Evaluation by Comparing Result Sets in Context

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

David Hawking
CSIRO ICT Centre
Canberra, Australia
david.hawking@acm.org

ABSTRACT

Familiar evaluation methodologies for information retrieval (IR) are not well suited to the task of comparing systems in many real settings. These systems and evaluation methods must support contextual, interactive retrieval over changing, heterogeneous data collections, including private and confidential information.

We have implemented a comparison tool which can be inserted into the natural IR process. It provides a familiar search interface, presents a small number of result sets in side-by-side panels, elicits searcher judgments, and logs interaction events. The tool permits study of real information needs as they occur, uses the documents actually available at the time of the search, and records judgments taking into account the instantaneous needs of the searcher.

We have validated our proposed evaluation approach and explored potential biases by comparing different whole-of-Web search facilities using a Web-based version of the tool. In four experiments, one with supplied queries in the laboratory and three with real queries in the workplace, subjects showed no discernable left-right bias and were able to reliably distinguish between high- and low-quality result sets. We found that judgments were strongly predicted by simple implicit measures.

Following validation we undertook a case study comparing two leading whole-of-Web search engines. The approach is now being used in several ongoing investigations.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*performance evaluation*

General Terms

Experimentation, Measurement

Keywords

Evaluation, embedded comparisons

CIKM'06, November 5–11, 2006, Arlington, Virginia, USA.

Figure 1: An example of the range of types of information sources available to an individual. A tool which provides a unified search interface to all of them is desirable, but a challenge to evaluate.

1. INTRODUCTION

There are certain forms of information retrieval (IR) tasks and systems which are of both research and practical interest but for which current evaluation techniques are a poor fit. One example is a system which attempts to incorporate knowledge of its users' current context and past preferences; another is the system illustrated in Figure 1, which provides search of personal files, source code, calendars, public and private Web sites, enterprise databases, email, and so forth in a unified interface.

If we have such systems, how should we measure their effectiveness? How should we compare alternatives? How could we make them more effective?

We discuss established evaluation techniques in Section 2. In Sections 3–4 we propose a randomised comparison tool in which alternative result sets obtained in the actual context of a real search are presented in side-by-side panels and the searcher is asked for online comparative judgments. We describe its implementation in two different forms and the mechanisms available for recording information about

searcher behaviour. We also discuss limitations of the methodology.

Section 5 reports experiments we have undertaken to validate the use of the tool in Web searching and to test for inherent biases. We also investigate the value of implicit predictors of explicit judgments.

The method has application to other IR problems. Section 6 gives a case study, and Sections 7 and 8 discuss variants of the basic tool and outline a range of other possible uses.

2. ESTABLISHED APPROACHES

In this section we briefly describe established IR evaluation methods, and discuss their applicability to comparing contextual or personal IR services.

2.1 Test collection approaches

The style of experiment introduced by Cleverdon [6], and notably taken up by the TREC, CLEF, INEX, and NTCIR conferences [28, 22, 15, 21], relies on three elements: a standard 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.

In general there are strong advantages of the test collection approach:

1. The low cost of evaluating a system once the collection is in place;
2. reproducibility of experiments;
3. reusability of the collection; and
4. the possibility of creating test collections including a sufficient number of information needs to permit robust, reliable comparisons.

Unfortunately, applying test collection methodology to contextual IR systems, or those which include personal corpora, raises particular problems.

First, personal information corpora almost always contain private data, such as email folders, which may not be viewable by experimenters. Personal corpora may also contain information which is not strictly private but is still restricted in some way, such as proprietary encyclopædias and content made available by subscription services.

Second, some corpora will be rapidly evolving and may even change from use to use. A fixed test collection can at best be a snapshot of the documents available at one instant.

Third, most search will typically cover tens of billions of documents as most searches will include the Web. Distribution of the Web as a document collection is infeasible, and the scale of the Web makes full judgements impossible for practical reasons. (Measures such as Buckley *et al.*'s *bpref* [4] and techniques such as those used in the TREC Web Track [11] or Web search evaluations [12] offer alternatives to full judgements in certain cases.)

Fourth, information needs are likely to be diverse and unarticulated. It is likely that future users of IR systems will use them for a range of purposes including question answering, known-item retrieval, service finding, and other tasks. Further, in many cases users may be unable or unwilling to articulate their information need, at least at early stages of the search process [27].

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 result 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[1].

2.2 Search log analysis

An alternative to explicit judgements is to consider user selection of a document as an indication of expected utility, and use clickthrough logging to evaluate systems. This technique is appealing in that it does not entail any extra burden on users, and it can capture judgements for a variety of information needs. Two forms of bias however need addressing.

"Trust bias" leads to more clicks on high-ranked documents, regardless of the documents' utility, as a result of users' faith in IR systems. "Quality bias" is the result of users being given a set of documents to choose amongst, not a single document at a time. A click on a document should be interpreted not as a vote for that document's relevance, but rather for its being more relevant than others in the set [17].

Further weaknesses in clickthrough data are relevant to our setting. First, although clickthrough data is known to correlate with utility in the Web [8, 17], it has not been established that this is the case for other types of data.

Second, many queries have no associated clicks. (From our observations of a commercial search engine log, this may be the case for a majority of queries.) The summaries provided may answer the user's information need without any further reading ("brilliant success"); the summaries provided may make it clear the IR system, or corpus, cannot answer the need ("abject failure"); or characteristics of the delivery mechanism may be responsible, such as Web page reloads or clicking more than once on the "search" button ("query bounce"). It seems important to at least distinguish brilliant success from abject failure, but this is not possible with clickthrough data alone.

To overcome trust and quality biases, Joachims suggests interleaving results from two systems, and using the number of results selected from each as an indication of that system's quality [16]. Tests with two presumed good systems (commercial Web searches) and a presumed poor-quality system (a commercial Web search but with reversed ranking) suggest this method can accurately predict user preference. We note however that this method is restricted to comparing ranked lists; it is not possible to consider sets of documents, for example for coverage or duplication, or lists arranged in some other manner (for example by source).

Finally, we note that search logs are generally maintained by individual search engines, which makes direct comparison difficult.

2.3 Human experimentation in the lab

A further method of evaluation involves observing test users in a laboratory setting, conducting searches in re-

[1]Indeed, inspection of logs covering ten million queries submitted to the search engine of a busy commercial website showed only around one query per 35,000 leading to clicks on all of the first ten search results.

sponse to a simulated information need.

The TREC Interactive Track [14] 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. This method is independent of any particular questions and corpus, but the same questions and corpus must be used for each participant in any given experiment.

Other techniques reported in the literature include post-search or post-experiment questionnaries [29] and manual judgements of results or result sets (*e.g.* [1, 26]). These can capture individual information needs and ideas of relevance, and are extensible to dynamic corpora, but in their common forms impose a significant burden on test subjects.

Borlund and Ingwersen [3, 2] suggest an alternative evaluation framework, the IIR (interactive IR) evaluation model. This model makes use of "simulated work task situations", which describe a work situation rather than a topic. The intention is to allow individual interpretation of utility, but this framework still relies on artificial information needs and may be confounded by inter-subject and order effects.

Evaluating IR by human experimentation in the lab is complicated by the problems of corpora and of artificial information needs, as with test collection methods. The complex experimental design, and associated overheads, also make this approach problematic.

2.4 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, in order to gain better understanding of user search behaviour.

Beaulieu observed library users as they used catalogue services and continued to browse the shelves [9]. Nordli [20] and Hansen and Järvelin [10] have carried out similar studies with, respectively, library users and with staff of the Swedish Patent Office.

There are serious problems with applying embedded observational techniques to our unified systems. First, it seems far too expensive for the benefit gained to employ experimenters to observe search behaviour of enough individual subjects over enough time to obtain an accurate general picture. We expect there to be large variability across the population. We also know that search activities may occur at any time of the day and usually comprise only a tiny proportion of a person's overall activities.

Furthermore, there are serious risks of altering the behaviour you are trying to observe. The mere presence of a search observer hovering in an office or around the home computer may seriously affect what searches are conducted and how. Asking a subject to vocalise their information needs or the process they are following is almost certain to affect search behaviour.

An alternative approach which avoids these objections is to use instrumented search software, which records aspects of interactions for later analysis. Kelly and Belkin [19] used monitoring software on specially-configured laptops to gather very extensive information on user's interactions with the Web. The "Curious Browser" of Claypool *et al.* [5] also recorded intractions with the Web, and explicit judgements of web pages, while Dumais *et al.* [7] used pre- and post-search questionnaires and recordings of interface actions to evaluate their "Stuff I've Seen" retrieval system with natural information needs in a large organisation. This technique is similar to our proposal below, although users in these experiments were not given the chance to directly compare two systems.

3. EMBEDDED COMPARISONS

In the future, it may well become possible to accurately predict from test collection evaluations the actual effectiveness of information retrieval tools in the field. Experiments with human subjects will also be able to assist in designing and comparing better retrieval tools.

In the meantime, we argue that collecting observational data about real search is a prerequisite both for choosing or building appropriate test collections and for designing useful human experiments. Following Hawking *et al.* [13], we propose a means by which some types of such data can be cost-effectively collected while at the same time allowing experimental conditions to be compared in full context.

Our approach combines aspects of embedded observation and search log analysis. We support a large range of real-world information needs and judgements by providing a real, working, IR system which takes the place of the searcher's usual system and logs interactions. Using metasearch techniques [25] we can provide a front-end to many different primary search interfaces and generate our own logs.

We also need however to compare two or more systems; and if we are to have users participate in our tests we need to minimise the additional overhead they face.

To this end we propose an interface that provides two or more panels, each of which presents a different IR system (or results from the same system presented in a different way, or generated with different parameters). A single field for user queries, which are passed to each system, provides an interface not much more complex than users are accustomed to. Figure 2 shows an example, using our pilot software searching corpora including email, calendars, addressbooks, local files, and the Web.

By logging interactions with each panel, we can infer which presentation is preferred. (At the start of each search, the systems are randomly assigned to the panels to control for any bias towards, for example, the left-hand side.) We also ask for explicit judgments of preference.

Using a live search system has several advantages. The corpora being searched are those really 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 similarly 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 judging entire result sets — for example for coverage — if this is important to users.

Presenting two conditions side by side and eliciting either explicit or implicit comparisons controls for differences between subjects (because the same person compares both conditions) and also controls for presentational order effects (because the two conditions are presented simultaneously). Furthermore, subjects are required to judge differences rather than to make absolute ratings which will later be compared.

The interface has two further advantages. Unlike the interleaved results list of [16], we can offer two different result presentations: for example, we can compare two clustering algorithms or compare graphical presentations such as

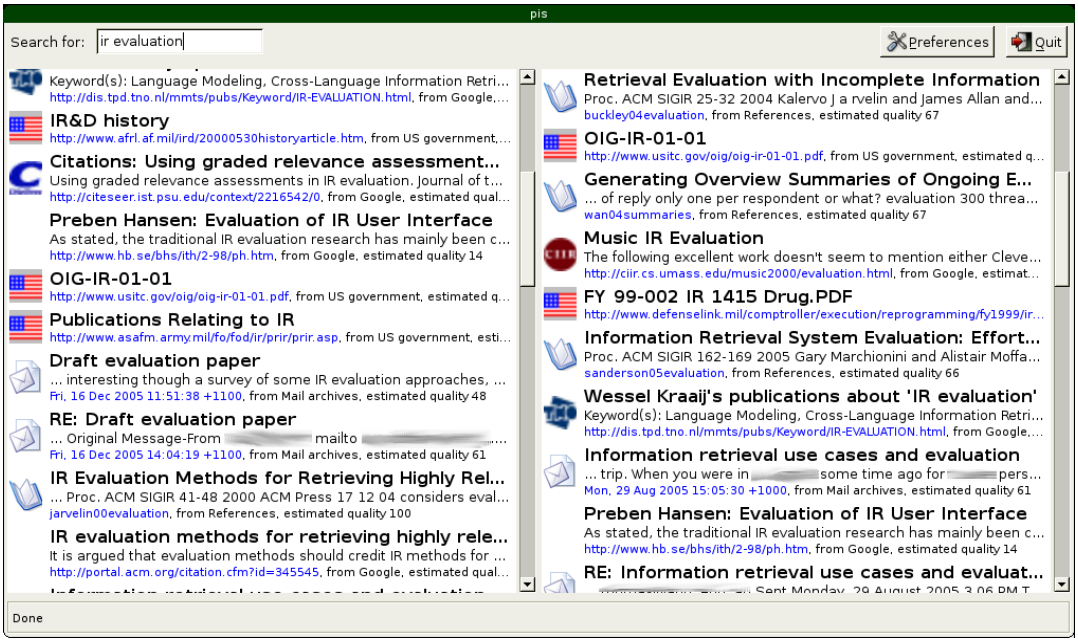

Figure 2: Sample two-panel interface

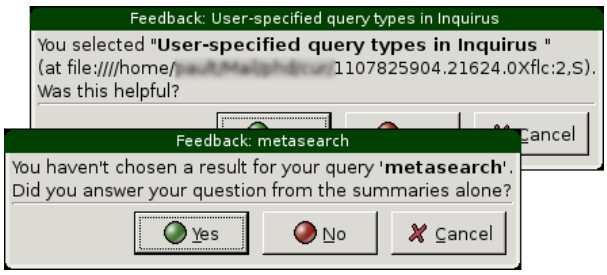

Figure 3: Extra feedback from the two-panel interface

Kartoo [18] with lists. At the expense of greater intrusion and effort for our subjects, the interface also allows us to prompt for extra information. We currently do this in two cases: periodically after a result is selected, to ask whether it was useful, and when no result is selected, to distinguish brilliant success from abject failure[2]. Figure 3 has examples.

We finally note that a variant of this tool could be used to compare systems in a similar manner to TREC ad-hoc, by combining the results from two systems into one list and asking users to judge each for relevance. This bears similarity to Shen's evaluation [26].

3.1 Limitations of the 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 and likely takes longer to present them. Careful design of the interface is required to minimise these effects. Feedback elicitation must be unobtrusive and make minimal demands on users.

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. We believe, nonetheless, that the ability to directly compare two IR systems in a natural setting is invaluable.

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, or exactly why. It is also difficult to make multi-way comparisons. Dividing a screen into more than two panels is feasible but it magnifies the experimenter effect by forcing larger presentational changes, requiring more time for judging and making the judging interface more complex. The alternative approach of making multiple pairwise comparisons inevitably makes it harder to control for between subject variation and order effects.

Finally, we note that if one panel takes significantly longer than the other to retrieve results, this could be the source of significant bias. We control for this in our experiments by retrieving all results for both panels before displaying either.

4. IMPLEMENTATION

We have implemented a version of the search system described above, which can access any corpus available to the user. Search is handled by external programs or libraries and results are merged and re-sorted using configurable methods including a random sort, a sort by each engine's reported quality score, and a sort based on term frequency in titles and summaries [23]. Modules exist for searching the Web, local files, email, BiBTEX bibliographies, LDAP

[2]From other logging, we can detect query bounce, and there is no "reload" command, which completes the set of possibilites described in Section 2.2.

directories, addressbooks, calendars, and any service with a Web interface (including for example `firstgov.gov` and `wikipedia.org`).

There are a number of interfaces to the tool, including a stand-alone application with its own graphical interface, illustrated in Figure 2 and which implements the pop-ups of Figure 3, and a Web-based interface which we have used in the experiments reported below[3]. Technical problems associated with the stateless nature of web interaction prevented us implementing the extra pop-ups.

5. VALIDATING THE DESIGN

We have carried out four experiments to verify that the approach described above can provide a useful comparison between two search systems. The first three of these requested explicit feedback on two search systems, one of which was known to be better than the other, and the fourth looked only at implicit feedback given the same two systems.

In the experiments reported below, one system is considered preferred to another if the number of users reporting it as better overall is greater than the number reporting the other or with no overall preference. (Users who submitted no judgements were not counted.) Significance is measured with a binomial sign test with criterion $p < 0.05$.

5.1 First experiment: popular queries

Our first experiment addressed two questions:

1. Given two result sets with a difference in quality, do users' judgements reflect this difference?

2. If so, can the two-panel design tell which is better?

Question 1 acts as a validity check — if the answer is "no", and users' judgements do not seem to distinguish a supposedly high-quality from a supposedly low-quality result set, other results will be very doubtful. Assuming however that our users agree there is a real difference, Question 2 is key. In answering this, we consider two further questions:

3. To what extent do users tend to prefer the left-hand panel, as the one they read first?

4. To what extent does clickthrough, timing, or other implicit feedback correlate with user judgements?

If the left-hand bias is small, and one or more types of implicit feedback correlates well with stated preference, then we will be able to use this implicit feedback and a two-panel design to compare systems. If, on the other hand, no implicit data seems to correlate with user judgements, we can only use a two-panel design with explicit judgements.

A convenience sample of 23 users were given the Web-based software described above, which simply acted as a proxy to the Google search engine. One panel, chosen at random, displayed Google's first ten results; this was assumed to be a (relatively) high-quality set. The other panel, assumed to be (relatively) low-quality, displayed Google's 21st through 30th results. Users were given queries from Google's list of popular searches[4] and after each search were prompted onscreen to indicate which set of results were "better", if either. No further definition of "better" was given.

[3]Please contact the authors for more information on this tool and its availability.

Age	22–54 (mean 35, std. dev. 10.5)
Sex	Male: 16, female: 6
Education	Postgraduate degree: 16, first degree: 6
Computer use	Daily: 21, occasional: 1
Web use	Daily: 21, occasional: 1
Search engine use	Daily: 17, occasional: 5
Computer experience	7–37 years (mean 19, sd 8.4)
Web experience	5–12 years (mean 9, sd 2.1)
Search engine exp.	5–11 years (mean 8, sd 2.4)

Table 1: User demographics for first experiment (23 users total). Not all users answered all questions.

Users were asked a number of optional demographic questions prior to the experiment, the responses to which are summarised in Table 1.

306 queries were recorded and 239 judgements. We observed a significant preference for the higher-quality set of results: 19 users preferred the high-quality set overall, 1 the lower-quality set, and 1 had no overall preference (sign test, $p \ll 0.01$). (The remaining two users made no judgements.) There was no significant difference in the preference for result sets in the left-hand or right-hand panels (sign test, $p = 0.09$).

Of the 183 queries where a preference was recorded, 34 had associated clickthrough data with between one and ten clicks per query. We considered four attributes of this clickthrough data: the panel which received the first click, the panel which received the last click, the panel which received the most clicks, and the panel which received the highest-ranked click. All agreed with the user's final judgement in 67% to 79% of cases, significantly better than chance alone (sign test, $p = 0.04$ to 0.001). There was no significant difference in the number of clicks in each panel ($p = 0.43$).

In this experiment, since users were assigned tasks and did not have a real need for information, the clickthrough rate is low and click patterns may be different to those in a more natural setting. We consider this possibility with our other experiments below.

These first results are very encouraging. In answer to Question 1, users' judgements certainly do reflect differences in quality; this suggests that they are able to judge two result sets when they are presented side-by-side in this manner. Further, it seems our design can tell us which set of results, and hence which search service, users prefer (Question 2). There is no significant bias to either side (Question 3), and our results suggest that clickthrough data may be able to predict user preference with this design (Question 4). We note, however, that our participants cover a limited demographic range and in particular that they are all well-educated and experienced with search engines and the Web.

[4]`http://www.google.com/intl/en/press/zeitgeist/archive2005.html`, downloaded 20 December 2005. Top queries for English-speaking countries from June to September 2005 were selected, duplicates removed, and a small number of apparently pornographic requests removed. 100 queries remained.

5.2 Second experiment: natural queries

Our second experiment considered the same questions and used the same technique, but participants were not assigned search tasks: users were instead encouraged to use our software in place of their regular Web search engine. Since users were issuing their own queries, for their own needs, and in their own time, we believe this gives a good indication of how well a two-panel evaluation method would work for IR systems in the field. For privacy, we suggested that users may choose not to use our software in cases where they would prefer us not to record data, but we expect this to have only a small impact on the number and range of tasks represented.

Demographics for participants in the second experiment, some of whom had already participated in the first experiment, were similar to those in Table 1 and are not repeated here.

179 queries were recorded and 147 judgements, 119 of which were for one panel or the other. The data proved similar to that from the first experiment: 17 of 20 users preferred the high-quality result set overall (a further 2 preferred the lower-quality set and 1 had no overall preference; $p \ll 0.01$), and there was no significant difference between the judgements in favour of the left- or right-hand panels.

85 queries had clickthrough data as well as an explicit judgement. All four predictors discussed above agreed with the final judgement 81–85% of the time, significantly better than chance alone ($p \ll 0.01$). Again there was no significant difference in the number of clicks in each panel.

We note that the second experiment has a much higher conversion rate from queries to clicks, and a marginally lower rate of "no difference" judgements. This may be explained by observing that in this experiment users are carrying out their own searches to fulfil genuine information needs; the higher click rate may be a reflection of the inadequacy of document summaries for some needs, and the lower "no difference" rate to our users' stronger sense of what constitutes a useful set of documents.

5.3 Third experiment: overlapping result sets

Result sets in our first two experiments were disjoint by construction. A third experiment considered the case where result sets overlap, and we ask: in these cases, where there may be less difference in quality, can our method still predict which set is preferred?

The same method was used as for the second experiment above, but with Google's results 1–10 in one panel and results 6–15 in the other (so there was an overlap between the bottom five results of one set and the top five results of the other). Demographics of the 37 users in this experiment, which included some users from the first two experiments, were again similar and are not repeated.

348 queries were recorded and 121 judgements, 79 of which were for one panel or the other. Despite the overlap and the presumed smaller difference in quality, results proved similar to the second experiment. Of the 25 users who offered one or more judgements, 18 preferred Google's results 1–10 overall and 6 Google's results 6–15; the remaining user had no overall preference. Although still significant ($p = 0.02$), this is a smaller difference than we observed when the two sets did not overlap. There was no significant difference in the number of judgements in favour of either panel.

The four attributes of clickthrough data considered earlier remained good predictors of the final judgement, with accuracy between 80% and 86% over the 66 queries with both click data and explicit judgements; again this was significantly better than chance ($p \ll 0.01$).

5.4 Fourth experiment: implicit feedback

The first three experiments suggest that in the context of our two-panel design, clickthrough data is a good predictor of final preference; they also suggest that users have a strong preference for a higher-quality result set. These observations suggested a final experiment, using the same system as the second experiment 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? It is possible, for example, that the presence of the explicit voting prompts users to read both result sets more carefully. In the absence of these their reading, and hence click patterns, may be different.

Demographics of our 18 users were again similar and are not repeated here. The four attributes of clickthrough data remained good predictors of the high-quality result set, with agreement between 68% and 87% over 129 queries with clickthrough data recorded ($p \ll 0.01$). As in the earlier experiments there was no significant difference in the number of clicks in each panel, suggesting no bias towards one panel or the other on the basis of their position alone.

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; we note however that this result is yet to be confirmed for non-web data or documents without appropriate summaries.

5.5 Observations

Although privacy concerns prevented us contacting test users directly, a number offered informal feedback on their use of the two-panel design. Comments were positive: none reported finding the two panels distracting (one even found this layout more useful, and requested that we maintain the software long-term). Several users commented that they found the process of scanning the result sets easy, but that especially in experiment three there was sometimes no reason to choose one over the other; this is consistent with our reducing the difference in quality and allowing overlap.

We have not carried out any formal investigation, but these comments and the number of queries collected suggest there is minimal extra burden for test users in comparing results side-by-side.

6. CASE STUDY

We have carried out a study to demonstrate the method in action. This compared two major whole-of-Web search engines, as exposed by their public APIs. (We refer to these as "engine A" and "engine B" below.) Again, users were asked to use our comparison tool as they would a regular Web search engine, and to indicate which set of results, if any, they preferred. The result sets were not re-ranked or modified except to ensure consistent display, and overlap between the two was allowed.

Age	20–54 (mean 33, std. dev. 9.4)
Sex	Male: 33, female: 11
Education	Postgraduate degree: 23, first degree: 18, other post-school qualification: 3
Computer use	Daily: 45
Web use	Daily: 43
Search engine use	Daily: 42, occasional: 2
Computer experience	10–38 years (mean 19, sd 6.9)
Web experience	5–15 years (mean 10, sd 1.9)
Search engine exp.	3–12 years (mean 9, sd 2.1)

Table 2: User demographics for whole-of-Web case study (49 users total). Not all users answered all questions.

49 users participated in this experiment; demographic details are summarised in Table 2.

We recorded 444 queries and 250 judgements, 158 of which were for one search engine or the other. Of the 40 users who recorded one or more judgements, 25 users preferred engine A overall and 13 engine B; two users had no overall preference. This is not a significant preference (binomial sign test, $p < 0.08$). There was no apparent difference in the number of judgements or clicks for either the left- or right-hand panel.

As in the validation experiments, we also considered clickthrough data to determine whether it reflected overall judgements. There were 117 queries recorded with both explicit judgements and clickthrough data; the four predictors considered were again accurate to 77–84%, significantly better than chance ($p \ll 0.01$).

Since we are recording user preference directly, this method may indicate how users rate search engine usefulness in real situations. By comparison, methods based on precision or similar scores indicate performance only in an abstract way and do not necessarily reflect user satisfaction.

7. APPLICATIONS

Having shown that the method we proposed can capture useful judgements in realistic settings, we feel confident that we can use this method to answer some of our real questions.

Our primary purpose in creating the evaluation tool is to enable us to develop better information retrieval systems of the type described in Section 1, which are hard to evaluate with current methods. We envisage that the tool could be used to shed light on many design questions, including:

1. Is it better to merge multi-source results into a single ranked list, to segment by source, or to cluster?
2. Can intelligent selection of sources lead to better results?
3. Do searchers prefer result sets which take into account aspects of their context such as role or location?
4. Do query-biased summaries help users?
5. Is the exact order of high-ranked results important?
6. How much does it matter which results are "below the fold" and not on the first screen of results?

There are many potential applications outside the immediate domain of contextual or unified information retrieval. Questions 3–6 above are also relevant to many other types of search, such as general web search, enterprise search, encyclopædia search, medical abstract search, etc. The technique could also be used in guiding purchasing decisions for website and enterprise search software.

8. FURTHER WORK

We are currently using the method in a few lines of enquiry, covering different IR domains.

Our major interest is in studying IR tools such as that described earlier and illustrated in Figure 1. A prototype tool has been developed which combines a working metasearch engine with a two-panel display and functions for recording preference, including the pop-up windows of Figure 3. The corpora and queries used with this tool can be assumed to be personal and will not be shared with researchers, so the technique is valuable for evaluating algorithms and design decisions.

Our second set of experiments is investigating the benefit of including a small number of external results in intranet search. We are interested to learn whether, for example, including a client's homepage or results from the public face of an organisation would prove useful even in intranet searches. Employees of a local research organisation have been asked to use our tool for their intranet searches; one panel contained the top 20 results from the installed search engine, and the other contained both these results and the top two results from the same query at a whole-of-Web search engine. Is this instance the side-by-side method allows us to capture real information needs and judgements, as well as to work with a constantly changing corpus.

A third ongoing experiment is investigating the influence of brand (such as search engine name, layout, or colour scheme) on the perceived quality of result sets. In this case it is also valuable to work with real information needs and a changing corpus; further, we are able to record user preferences which are due only to branding and not to the quality of result sets or individual results.

9. CONCLUSIONS

After due consideration of familiar evaluation methodologies we have proposed and implemented an IR evaluation method based on a tool which takes the place of a user's normal search interface and offers results from two systems side-by-side. This tool can collect queries, interactions, and explicit judgements as they occur, and can be used with private or dynamic corpora. Use of the tool avoids many of the costs and biases of familiar evaluation methods.

Experiments to validate our method, using Web search systems, confirm that we are able to detect a difference in user preferences between a high-quality set of results and a lower-quality set. We did not detect any significant preference for either the left-hand or the right-hand panel of results.

Our experiments and case study also demonstrated that clickthrough data is useful in inferring preference in side-by-side presentations of web results, even in the absence of explicit user judgements. We have gained enough confidence from our validation experiments to use this technique to compare IR systems, although we can ask for explicit

feedback in cases where clickthrough data may be unreliable.

Demonstration of the method in a case study indicated that it is a viable means of gathering preference data with real users and real information needs. A comparison of whole-of-Web search engines, as demonstrated, could be of particular value to Web search providers as it captures user preference directly.

We are particularly interested in studying IR over a diverse range of information sources. This method is also however applicable to a number of questions in other areas of IR, and future work will apply it to questions outlined in Sections 7 and 8 above.

The major concern in interpreting our results is the narrow demographic range of the subjects involved in experiments thus far. A natural direction for future work is to determine whether these results, obtained on a highly educated and computer-literate group of users, generalise to a broader demographic; however we note that the early adopters and power users of future IR tools such as the personal system of Figure 1 are very likely to have similar demographic profiles.

10. ACKNOWLEDGEMENTS

We would like to thank the Statistical Consulting Unit at the Australian National University for their advice, and Tom Rowlands of the CSIRO ICT Centre for the analysis of search engine logs described in Section 2.1. We are grateful for the help of the approximately 100 people who participated as test users.

11. REFERENCES

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. ObjectRank: Authority-based keyword search in databases. In *Proc. VLDB*, 2004.

[2] P. Borlund. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 2003.

[3] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life. In *Proc. ACM SIGIR*, 1998.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. ACM SIGIR*, 2004.

[5] M. Claypool, P. Le, M. Waseda, and D. Brown. Implicit interest indicators. In *Proc. Intelligent User Interfaces*, 2001.

[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] S. Fox. Evaluating implicit measures to improve the search experience. Talk presented at *SIGIR Workshop on Implicit Measures of User Interests and Preferences*, 2003.

[9] 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.

[10] 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.

[11] D. Hawking and N. Craswell. Very large scale retrieval and web search. In Voorhees and Harman [28].

[12] D. Hawking, N. Craswell, P. Bailey, and K. Griffiths. Measuring search engine quality. *Information Retrieval*, 4(1), 2001.

[13] D. Hawking, C. Paris, R. Wilkinson, and M. Wu. Context in enterprise search and delivery. In *Proc. IRiX Workshop, ACM SIGIR*, 2005.

[14] W. Hersh and P. Over. TREC-9 interactive track report. In *Proc. TREC*, 2001.

[15] INitiative for the Evaluation of XML Retrieval. `http://inex.is.informatik.uni-duisburg.de/`.

[16] T. Joachims. Evaluating retrieval performance using clickthrough data. In *Proc. SIGIR Workshop on Mathematical/Formal Methods in IR*, 2002.

[17] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proc. ACM SIGIR*, 2005.

[18] Kartoo, S. A. `http://www.kartoo.com/`.

[19] D. Kelly and N. J. Belkin. Display time as implicit feedback: Understanding task effects. In *Proc. ACM SIGIR*, 2004.

[20] R. Nordli. "User revealment" — a comparison of initial queries and ensuing question development in online searching and in human reference interactions. In *Proc. ACM SIGIR*, 1999.

[21] NTCIR (NII-NACSIS Test Collection for IR Systems) Project. `http://research.nii.ac.jp/ntcir/`.

[22] C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors. *Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001*, volume 2406 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.

[23] Y. Rasolofo, F. Abbaci, and J. Savoy. Approaches to collection selection and results merging for distributed information retrieval. In *Proc. CIKM*, 2001.

[24] Y. Rasolofo, D. Hawking, and J. Savoy. Result merging strategies for a current news metasearcher. *Information Processing and Management*, 39(4), 2003.

[25] E. Selberg and O. Etzioni. Multi-service search and comparison using the MetaCrawler. In *Proc. WWW4*, 1995.

[26] X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In *Proc. CIKM*, 2005.

[27] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: A study of orienteering behaviour in directed search. In *Proc. Conf. Human Factors in Computing Systems*, 2004.

[28] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[29] 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.