Live Web Search Experiments for the Rest of Us

Timothy Jones
Computer Science Dept.
The Australian National University
Canberra, Australia
tim.jones@anu.edu.au

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

Ramesh Sankaranarayana
Computer Science Dept.
The Australian National University
Canberra, Australia
ramesh@cs.anu.edu.au

ABSTRACT

There are significant barriers to academic research into user Web search preferences. Academic researchers are unable to manipulate the results shown by a major search engine to users and would have no access to the interaction data collected by the engine. Our initial approach to overcoming this was to ask participants to submit queries to an experimental search engine rather than their usual search tool. Over several different experiments we found that initial user buy-in was high but that people quickly drifted back to their old habits and stopped contributing data. Here, we report our investigation of possible reasons why this occurs.

An alternative approach is exemplified by the Lemur browser toolbar, which allows local collection of user interaction data from search engine sessions, but does not allow result pages to be modified. We will demonstrate a new Firefox toolbar that we have developed to support experiments in which search results may be arbitrarily manipulated. Using our toolbar, academics can set up the experiments they want to conduct, while collecting (subject to human experimentation guidelines) queries, clicks and dwell times as well as optional explicit judgments.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*

General Terms

Measurement, Design

Keywords

Implicit measures, web search, browser extensions

1. INTRODUCTION

Understanding the effect of changing ranking on result quality is an important goal in Information Retrieval. In order to measure the impact of a change in result quality, it is important to measure ranking schemes *in context*, both by presenting entire ranked lists (as opposed to individual documents), and by testing using real users and their real information needs. One approach is to ask users to replace their normal search engine with an experimental one for a period of time [4]. However, for these experiments to be successful it is important that users use the experimental search engine for a significant period of time.

It would be unfortunate if academic researchers were unable to study whole-of-Web search, since many people's searching is predominantly across the whole Web. Furthermore there are areas of research, such as the effect of spam on search quality, which make little sense outside the Web context. When studying real Web search it is important that the search engine used by experimental subjects gives similar coverage, freshness and result quality to leading commercial search engines. In such studies the experimenter should be able to:

- manipulate the results pages,
- obtain explicit feedback (e.g. by adding rating buttons)
- record user behaviour data such as clicks and dwell times (implicit measures, [3]).

We demonstrate a browser toolbar which allows manipulation of result lists provided by a commercial search engine and records user interactions with the engine. With this toolbar, academic researchers are able to conduct experiments which would normally be restricted to employees of major search companies.

2. PREVIOUS WORK

Two methods previously used by academic researchers to study real web search are side-by-side panels and browser toolbars.

Thomas and Hawking proposed a two-panel methodology [4], in which users are presented with two side-by-side panels of search results for each query. Each panel is generated using a different ranking or search engine, and the left-right ordering of panels is randomised. Users are invited to indicate one panel as "better" than the other, or that there is "no difference". "Better" was intentionally not defined, allowing users to use their own preference judgement. In addition to the explicit feedback of the preference indication, implicit feedback by click log was tested. These were found to correlate with the preference judgement. Furthermore, there was also found to be no implicit bias towards either the left or the right result panel. In their first experiment, 23 users submitted a total of 306 queries, with 183 explicit preferences. They conducted four experiments in total, all with similar

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

numbers of users/queries/preferences. Those numbers are not repeated here for reasons of berevity.

Building on the two panel work, Bailey, Thomas and Hawking conducted an experiment looking at the influence of branding on perceived result quality [1]. They found that users continued to judge results from their preferred search engine as slightly better, even if brand labels were swapped or if the results were attributed to a fictitious search engine. From this experiment, we draw the conclusion that as long as a search service provides high quality results, users will be happy to use it.

The Lemur Toolbar[1] is a browser extension designed for collecting query and click logs from users of commercial search engines. It has versions for both the Firefox and the Internet Explorer web browsers, and supports most major search engines. Queries and clicks on result URLs are logged. These logs are periodically uploaded to a server, either automatically or manually by the user. Users have control over the private content of the logs, either by removing individual queries, or by setting up regular expressions to blacklist (phone numbers, credit card numbers, etc). However, the toolbar is only capable of collecting data about results provided by the search engine. It does not support additions to the results such as explicit feedback buttons or the provision of additional search results. Consequently, it can only be used to generate logs from search results provided by an existing engine.

3. USER DROP OFF

In an as-yet-unpublished study, we conducted a two panel experiment looking at the preference difference when varying the number of spam results present in each result list. The experiment followed the methodology of Thomas and Hawking [4]. The experiment differed from previous two panel experiments in that we were introducing known low quality results. Result lists were constructed by presenting results from a well known commercial search engine salted with known spam pages heavily (3-5 spam results) in one list, and lightly (0-2 results) in the other. With one list containing 0-2 introduced spam results, at least one list remains high quality. Users were told that they were testing an experimental search engine we had built; they were not told that the majority of each panel came from commercial search. In order to encourage users to continue to use the search tool, an Open Search Description Document[2] was created. This enabled users to add our search tool to the search bar in their web browser. Over a period of 3 months, 23 users submitted 549 queries, and expressed 194 explicit preferences. These numbers fall within the ranges reported in previous experiments.

For the most convincing results, users need to submit many queries over the time of the experiment. In our experiments, we found many users submitted few queries. 19 of our users submitted a mean of 4 queries and preferences each. However, the other four users submitted more than 30 queries each, with one user submitting more than 300. This can be explained if we count the number of days between a users first and last access of the tool (duration of participation). Figure 1 shows the frequency of users with each duration of participation in our experiment. Nearly all users started and stopped using the tool on the same day. Under ideal conditions, all users would continue submitting queries for the entire duration of the experiment.

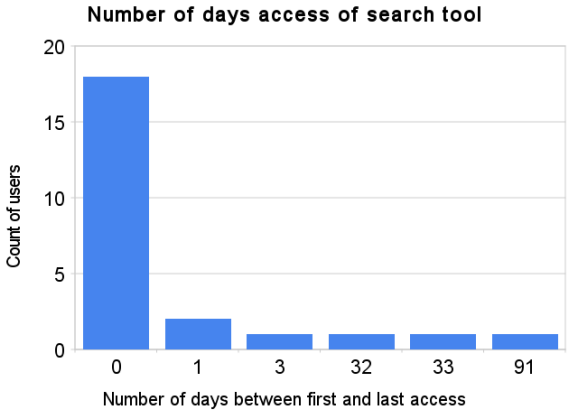

Figure 1: Frequency of users against total number of whole days access of the search tool. Note the X axis has no scale. Only four users used the tool for more than one day, and only one user stayed for the duration of the experiment.

At least one of the result panels was mostly high quality results. Since we know that branding does not affect perception of quality [1], we may infer that users did not stop using the tool because of a perceived lack of quality. Anecdotally questioning users of our tool gave reasons such as "I forgot" and "I didn't want to miss anything Google might tell me". While branding does not affect perception of quality, this last comment implies branding may affect initial *choice* of engine when presented with a search task.

4. CHOICE OF ENGINE AT SEARCH TIME

In this section, we report on a simple experiment to answer the question "when presented with a wide range of search engines and a search task, do users only pick their favourite search engine?".

In this experiment, users were given three search tasks, and asked to find a web page that best answers the task. Tasks were displayed one at a time, and users were presented with a list of search engines. This list included the three leaders in market share (Google, Yahoo, Microsoft's Bing), an engine with low market share (Dogpile), an engine which has received some bad press (Cuil), and a fictional engine with no reputation (TwoPotatoes, actually just a front end for Yahoo's Build Your Own Search Service). To get to the next task, users were asked to submit the URL of a web page that contained an answer to the search task. The experiment interface can be seen in Figure 2. After completing all three tasks, users answered a short questionnaire about their choice of engines. The questions can be seen in Table 1.

Here are the search tasks users were presented with:

1. "Funny People" is a recently released movie. Find session times for this evening at a cinema near you.
2. Every year, there is a conference about web technology called "World Wide Web" (WWW). Where is it going to be held in 2010?

[1] http://www.lemurproject.org/querylogtoolbar/
[2] http://www.opensearch.org/

3. What song is the following lyric from: "Yes, the tv's on, radio blastin' the news / Somebody down the hall, playin the low down dirty blues"

The first two tasks were chosen by asking two department members what their most recent search task was, and the third was created intending to suggest a different type of searching than the first two (phrase searching). Users for the experiment were found by posting to the popular social networking site Facebook[3]. In this way, we selected users with regular experience with the web.

Question	Type
Were you able to complete the task?	Yes/No
If not, why not?	Free text input
Which search engines did you use?	Multiple checkboxes
During the task, did you use any search engine that you wouldn't use during your normal web use?	Yes/No
If you used any search engine you wouldn't normally use, why?	Free text input

Table 1: The questions asked once all tasks were completed. All questions were asked three times, once for each search task

5. RESULTS

28 users took part in the experiment, completing a total of 83 search tasks (Two users failed in a total of two tasks). Both failures were on the movie times task, because there were no local show times for that particular movie. 9 users in a total of 15 tasks (32% of users, 18% of all tasks) used an engine that wasn't their normal engine. When asked why they chose an engine other than their usual engine, answers fell roughly into 4 categories: *curiosity*, users trying a new engine to see what it was like; *practicality* users who were unable to use their usual engine for some technical reason; *poor results*, users who turned to a different engine because results from their favourite engine were unsatisfactory; and *no engine* users who were able to answer the question without the use of a search engine. Table 2) shows the frequency of users and tasks in each category. Of the users who used their usual engine, 24 users in a total of 66 tasks used Google, and 2 users in a total of 2 tasks used Yahoo (both of these users consider both Google and Yahoo to be their usual engine).

Category	Number of users	Number of tasks
Curiosity	4	6
Practicality	1	3
Poor results	3	5
No engine	1	1

Table 2: Breakdown of reasons for users not using their favourite search engine

6. DISCUSSION

Even though several engines choices were presented, users chose their usual engine in 82% of tasks. While users are not biased by branding when judging quality of results [1], we believe users *are* baised when reaching for an engine to complete a search task. This would explain the drop off in numbers from our earlier experiments. Users are prepared to use an experimental search tool at first, but when a real search need comes to them, they will reach for their usual engine. In the experiment described in Section 3, we had 23 users, 4 of whom used the experimental seach engine (instead of their usual search engine) for the duration of the experiment. When asked why they persisted to the end one user answered "I was curious to see if your results were any better than (their favourite search engine)". A comparable number of users took part in our search task investigation (28 users). 4 of those users used an engine other than their usual search engine for reasons of curiosity. The 4 users who continued to use our experimental engine, and the 4 curious users (both from groups of roughly 25 users) may well be the same demographic. As the size of the *curious* group is small, an experiment that the whole group participates in would be much better.

If users always reach for their favourite search tool, an effective strategy would be to replace or augment their favourite tool for the duration of the experiment. In the experiment in Section 3, we made it easy for users to add our search tool to their browser's the search bar, replacing one of the ways to access their favourite search engine. As we have already seen in Figure 1, the majority of users still did not use the search tool for more than a day. This may be because users would navigate directly to their usual search engine, or perhaps select their favourite search engine from the search bar once, and then forget to change it back to ours.

7. AUGMENTING WEB SEARCH ENGINES

We have seen that users will generally use their favourite search tool whenever they have a search task. Most of our two panel experiments have involved mixing results from commercial search with results of our choosing. Because of this, commercial search result pages already provide half of what we need. In this section we present a browser extension that allows us to manipulate result pages from commercial search engines to include arbitrary results. It also allows us to include arbitrary logging capabilities.

7.1 Tool architecture

We chose Mozilla Firefox[4] as our target browser, as the majority of users using our experimental interface were Firefox users. Similarly, we chose Google as a target search engine, though the tool can easily be modified to enable it to work with other search engines.

Firefox addons are written in a combination of JavaScript (for the logic) and XUL[5] (for the user interface). As our tool is mostly invisible to the user, most of it is written in JavaScript. The use of JavaScript allows us to include the comprehensive JQuery library[6], which simplifies interactions with the HTML search result page. The tool hooks in to the DOMContentLoad browser event, which fires when a page has finished loading. Then, the tool proceeds as follows:

```
if(SearchEngineResultPageDetected)
        logToServer(queryTerms, pageNumber)
```

[3] http://www.facebook.com

[4] http://www.mozilla.com/firefox

[5] http://www.mozilla.org/projects/xul

[6] http://jquery.com

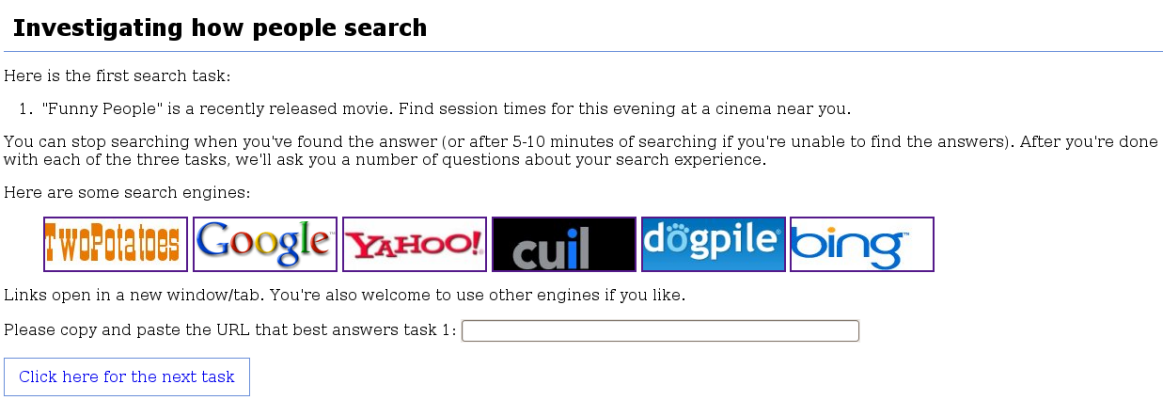

Figure 2: The experiment interface. The order of the list of search engines was randomised for each user.

```
// The following are optional
// depending on the experiment
augmentResultsWith(ClickListeners)
augmentResultsWith(FeedbackButtons)
insertExtraResults(sourceUrl, pos)
```

All logging is done with XMLHttpRequests to a log server, specified at the creation time of the tool. The intention is that each experiment would have a custom instance of the tool built for it. We will briefly describe each of the optional features currently implemented in the tool.

- *Click Listeners* obtain the user's click log. Each search result gets given an additional JavaScript onClick event, which logs the target URL, query terms, result page number, and position in the result list.

- *Feedback Buttons* add extra buttons to label results. These can be binary good/bad buttons, as in the case of Google's Search Wiki (Figure 3), or even a classification task (Figure 4).

- *Extra Results* can be placed among the real results. These can replace or be inserted in existing results. Because extra results usually need to come from another server via an asynchronous request, all results are hidden until the extra results appear. This extra time delay may introduce a bias to perception of quality, so it is logged.

8. OUR DEMONSTRATION

Our demonstration will allow WWW attendees to participate in an experiment using our toolbar, which will comply with a protocol approved by the Australian National University Ethics Committee. Subjects will interact with modified results from a high quality web search engine and may be asked to rate the result set as a whole or the value of individual results. Their interactions with the results will be recorded. It is not appropriate to disclose the experimental question in advance. We will invite attendees to download the toolbar and to continue to participate in our ongoing experiments. We will also provide a URL to download the source code for an extendible skeleton of our tool.

Figure 3: Google's SearchWiki allows users to promote, demote, or comment on search results

Figure 4: Classification buttons added to search results by our tool

9. CONCLUSIONS

We have presented and will demonstrate a new advanced browser extension, capable of supporting realistic Web search experiments outside the confines of Google, Bing or Yahoo! labs. The approach we describe supports any type of experiment in which the content of the search results is unimportant (other than a requirement that they be high quality), such as investigations into query biased summary generation and/or result presentation. It supports the collection of rich information about real web results (eg. spam content, usefulness, and many implicit measures). Furthermore, it supports the inclusion of new results to facilitate experiments about result content and ranking, such as the comparison of multiple rankings (possibly provided by multiple engines), either by interleaving the results [2], or by extending the tool to allow the two panel approach described in [4].

10. REFERENCES

[1] P. Bailey, P. Thomas, and D. Hawking. Does brandname influence perceived search result quality? yahoo!, google, and webkumara. In *Proceedings of ADCS 2007*, December 2007.

[2] T. Joachims. Evaluating retrieval performance using clickthrough data. In J. Franke, G. Nakhaeizadeh, and I. Renz, editors, *Text Mining*, pages 79–96. Physica/Springer Verlag, 2003.

[3] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[4] P. Thomas and D. Hawking. Evaluation by comparing result sets in context. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 94–101, Arlington, Virginia, USA, Nov. 2006.