Assigned tasks are not the same as self-chosen Web search tasks

Daniel M. Russell *Carrie Grimes*
Google, Inc.
{drussell, cgrimes}@google.com

Abstract

Short assigned question-answering style tasks are often used as a probe to understand how users do search. While such assigned tasks are simple to test and are effective at eliciting the particulars of a given search capability, they are not the same as naturalistic searches. We studied the quantitative differences between assigned tasks and self-chosen "own" tasks finding that users behave differently when doing their own tasks, staying longer on the task, but making fewer queries and different kinds of queries overall. This finding implies that user's own tasks should be used when testing user behavior in addition to assigned tasks, which remain useful for feature testing in lab settings.

1. Introduction

In the course of testing and evaluating search engines and information retrieval systems in general, there is a real need to test the efficacy of the system. A common technique is to develop a battery of questions that can be answer by using the target system. These questions are then used by test subjects to evaluate the system in terms of the subject's behavior. Assigned tasks are a commonplace in usability lab settings where they are used to elicit system feature use, user interface features (and problematic features), and to provide observable instances of how people conduct their searches.

An obvious question, though, is to what extent such short assigned tasks are a good proxy for actual "in the wild behavior." In this paper we study the measurable differences in subject behavior when given short *assigned task* versus having the subject do a task that is self-chosen—what we call an *own task*.

We also realize that everyone, to a certain extent, has *assigned tasks* as a fraction of their day-to-day work. These may be simple tasks taken on in the course of the workday, or more complex assignments given for school or other reasons. In investigating *own tasks* versus *assigned tasks*, we hope to understand some of the behavioral differences that frame the way searchers ask their different kinds of questions.

2. Background

There are many ways to study search engine user behavior. The most common method is to do log file analysis. [5] [9] [8] Such transaction logs analysis is common practice throughout search engine engineering, Web log analysis is used to discover the patterns of queries, the relative frequencies of different kinds of queries, the variety of queries actually made and the overall characteristics of user behavior in the aggregate. Web logs analysis also informs search engines about what results are the most popular, information that is fed back into the ranking systems. From a user understanding perspective, Web log analysis is often ultralarge in scale, making sample population comparisons difficult when viewed with large numbers of web searches. [6] Smaller studies are possible with log analyses such as those done to understand smaller web sites and their more focused user tasks. [19]

Field studies of Web searching behavior is far more difficult than Web log analysis in terms of time and analysis depth, but can give a great deal of detailed information about both the context that surrounds the user's search behavior as well as a speak-aloud protocol to illuminate the otherwise invisible goals, intents and tasks that make logs difficult to decrypt. Going to watch search users *in situ* can reveal a host of unanticipated issues that would be otherwise impossible to detect.

Lab studies include traditional usability methods [7] and, increasingly, eye-tracking studies to monitor the user's eye movement over the search interface to indicate where their attention is focused. [1] [3] [4] These kinds of studies are often very revealing of user conceptual models and can be used to discover the fine

grain of user behavior while doing search in the laboratory.

Traditionally, of course, there has been a great wealth of user studies of IR systems. [5] [15] [16] While these studies are useful, it is often difficult to compare the results from these classic IR systems user-search behavior to web-search behaviors: there are large user interface differences, study technique differences and the variation in study tasks that were assigned. For instance, session length seems to have been much longer in the classical library studies (roughly 14 queries per search session, which is much longer than the average for a web search session, which hovers around 2.5). [5] A common finding across all of these older studies, however, was that there is a significant effect of the tasks used to study system effectiveness from the user's point of view.

Task-based studies. One common method in both IR studies and Web use investigations is a task-based study, i.e., when a subject is given a short question to answer, or a longer contextualized description of a situation with a task to resolve. [7] [10] [18] While the short trivia question approach is no longer in wide use (being deprecated as being relatively uninsightful and dependent on a person's individual knowledge), task situation assignments are in relatively common use in usability labs. [12] The key to successful task assignments is to make them understandable to the test subject, believable in the context of the task description, and engaging enough to elicit realistic behavior. [16] The objective of the task assignment is to generate an information need that the test subject will take on as if it were their own. TREC [13] uses many different kinds of questions, both short and contextualized to study search behavior from multiple angles.

However, in order for any kind of user study to shed light on authentic user behavior outside of a laboratory setting, we need to understand the differences between what happens when a user does the test task and when they do a more natural task of their own choice. This is the premise of this study: to begin to understand how closely short assigned tasks align to normal authentic user behaviors.

3. Experiment

In the fall of 2005, Google partnered with Keynote Systems to do a blind study of internet search behaviors. In the study reported here, 401 subjects were recruited and given a set of search tasks to do that would inform us about typical search behaviors. (Subjects did not know that Google was running the study.)

Tasks were given to the subject by a Web application that would present the question to be addressed and coordinate timing information. The task list was 45 items long and was done by the subject in their normal internet search use environment (typically their home).

Each subject would work through their task list by answering the questions at whatever time and place they would normally do their internet searches, avoiding the need to do their work in a lab that would be unfamiliar to them. The subjects were encouraged to make their work on the tasks as near to their normal search activity as possible.

The Web helper application presented each subject with a set of questions to answer. There was an initial set of 43 questions about demographics, internet use practices and background data that acted as practice trials to acclimate the subject to the data collecting application. After working through those questions, the task list was presented one item at a time in a separate window adjacent to the main browser window. All subjects ran Internet Explorer on Microsoft Windows XP. The subject could easily suspend working on the tasks at any time (between tasks only), and resume work at a later time. Thus, a subject could work on a single task for a bit, then break off for a bit. They could then return to complete the rest of the study tasks when they had another stretch of available time. While this is not a highly controlled lab study environment, this setup does nicely approximate the reality of internet search use behavior. As we know from other studies, internet search users often work in short bursts, interleaving multiple tasks with search tasks. [17]

This test was conducted on a panel of randomly selected subjects that was demographically balanced to reflect the population of US internet users. Our user population was split 50/50 male/female, with a range of connection speeds (from 1.5Mb or more--17%, 768Kb -- 20%, 384Kb--15%; and less--48%), a representative split of household income and education, age groups (18 – 29, 21%; 30 – 39, 38%, 40 – 49, 25%, remainder > 50 years old). Users were paid for their participation and were able to do all of the test tasks over a period of 2 weeks in their home environment.

The data was collected via remote upload to a server for later analysis. (The subjects knew about the data collection and upload, but this was fairly transparent part of the testing framework.)

3.1. Task list questions

The task list given to the subjects had 5 different tasks, each asking the subject to search for a different kind of content. The 5 types of search tasks were: (1) general Web search, (2) local information (such as the local of a pizza parlor in a nearby town), (3) product information, (4) image search within given guidelines and (5) news search for information on a recent topic.

A task could be an ***assigned task***--that is, one that posed specific question for the subject to solve, or a self-chosen problem, an ***own task***—where the subject was asked to search the answer to a question or problem that they genuinely wanted to do in the course of their normal search behavior.

In this experimental design, 20% of the questions were *own tasks*, interspersed in a balanced sequence among all the assigned tasks.

A typical assigned task would be:

> Suppose you recently moved and you would like to find the closest Stop N Shop grocery store to your new home. Use Google to search for the closest Stop N Shop grocery store. Please use 80012 as your zip code. What is the street address of the Stop N Shop nearest to your new home?

That is, the goal is pretty detailed and complete, with much of the background information provided by the task specification. Note that a great deal of contextual information is given by this description, but a method to do the search (other than using Google) was not prescribed.

By contrast, own tasks were given as very open-ended questions that were to arise from the subject's own experience and needs. An *own task*:

> For the following task, you will be asked to use Google to search for something on the Web. Please take a moment to think of a topic, Web site, or piece of information you would like to search for. It should be something you are genuinely interested in finding or learning about.

In this, the user's *own task*, no contextual information can be given. Thus, it's up to the user to understand what it is that they're actually trying to accomplish.

Before actually beginning the task, the user was asked to give a short description of their own task goal.

> Below, briefly describe what you will be searching for and what you're hoping to do with the information you find.
> {Text Response}

The *own tasks* were intended to be genuine, personally engaging, self-defined tasks in which the user could behave as normally as possible.

When each task was given, the subject would click on a button to indicate that they'd started working on the task. At any point during the task, they could suspend working on the task by clicking on a "Suspend Work" button, although this happened very rarely during the trials. (Less than 1% of the time.)

After completing a task, the subject would click a "Completed task" button, and would proceed to answer a short series of post-test questions about their performance on the task, their satisfaction with the search process and an open-ended opportunity to add comments about the task.

Subjects would typically work through the questions and tasks in a single sitting of about 1 hour, creating a session log for each task. After cleaning the data from failures and obvious corruptions, we were left with 399 own task sessions and 1587 assigned task sessions we could analyze. Each task varied in length from the rather short (1 query, 1 click to see the result) to the lengthy (30 queries with many results clicks as the user explored the search results looking for the answer to their question). Similarly, task time varied from a few seconds to many minutes.

4. Analysis

The Web browser helper application tracked each subject's searches and all Web sites they visited during the course of their search for the solution. Timing data, window open/close events and queries were all collected into sessions, where a task session is the sequence of all events from the moment the subject clicks the "Begin task" button until they click "Completed task."

Subjects were asked to do 4 assigned tasks, one each from the categories of "Local" search, "News" search, "Product," and "Image" search. The subjects were directed to begin their search at the home page of the search engine. However, in the case of news tasks task (e.g., "Find the latest information about Hurricane Katrina"), we quickly discovered that the behavior of users who chose to transition to a news specific search property was collectively very different than that in any of the other assigned or own tasks. Since subject performance was so different between subjects, we removed that set of data from the analysis. (We

removed the news task sessions from our pool, leaving 1188 assigned task sessions we could analyze.)

We began our analysis by asking a simple question: Do own and assigned tasks take the same amount of time?

A: Own tasks take longer than assigned tasks: Interestingly, the mean session length (in seconds) for our subject's own tasks was 287 seconds, median 182, while local search mean = 178, product search mean = 152, and image search mean = 132. The aggregate mean of all the assigned tasks was 154 seconds, with median of 97. Figure 2 shows the conventional histograms of each of the task categories.

In order to test our hypothesis more simply, we take the log of session time. The resulting log session length distributions are approximately normal. In the log-space, we tested the hypothesis that the time-per-query session was the same by checking the means of the two distributions (assigned tasks and own tasks) using a pooled variance t-test. This test returns a t-statistic 10.7081, with df: 1585, and a p-value of effectively zero for the two groups having the same value. Each individual task type (local, image and product) also tests with a mean of less than the mean of the own tasks.

Figure 1 shows the histogram of log-session length for the 4 task groups (own tasks, local search, product search and image search tasks) showing on the first group with the mean and the confidence interval for the mean, followed by the rest of the groups.

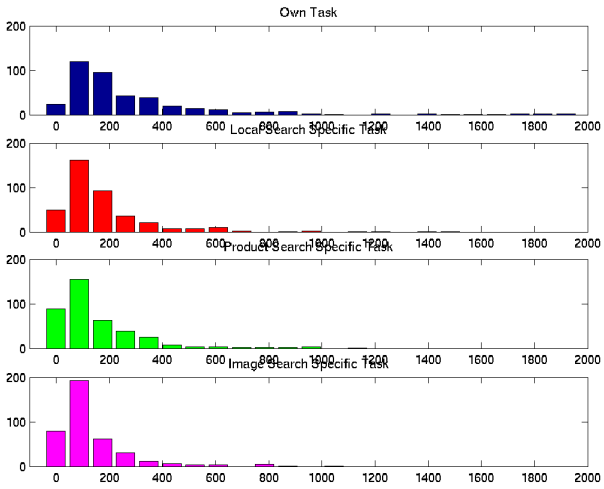

Figure 1. Own task simple time distributions for Web search, assigned local search, assigned product search and assigned image search tasks.

The comparison between log session time distributions is striking: all the sessions times for each of the task types is fairly normally distributed, much as you would hope given this sample size. (With the exception, as noted above, of the news search tasks, which were excluded.) This makes a comparative analysis straightforward and makes one confident in the outcome.

Why the difference between own and assigned tasks? It is clear from this analysis that own task sessions are significantly longer than assigned tasks. This led us to wonder why this would be so. Were the subjects reading each page longer? Were they posing more questions? What was going on?

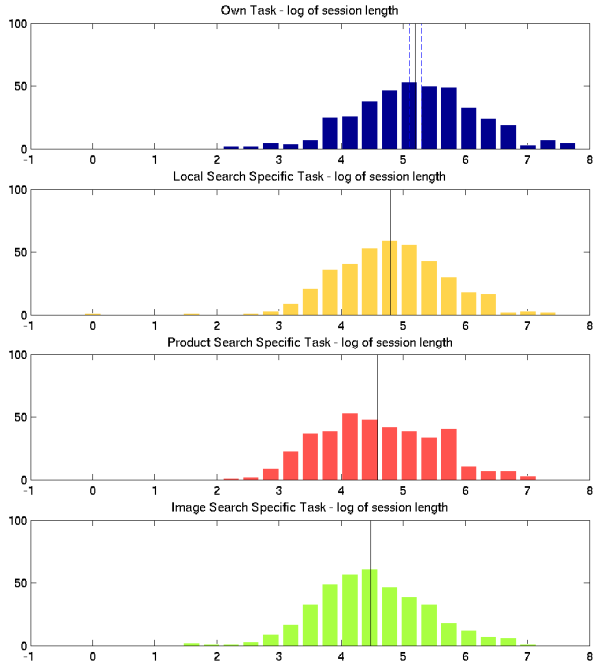

Figure 2. Tasks log time distributions for own Web search, assigned local search, assigned product search and assigned image search. This shows the differences between the mean search times very clearly, with own tasks (the top row) being longer than any other kind of assigned search.

B. Fewer unique queries in own tasks. We found that subjects ask relatively fewer unique queries in the own task condition than when given an assigned task. That is, subjects did fewer individual queries or refinements of the query in the own task condition. We defined a unique query as any "distinct" query – even if it was clearly related to a previous query, and also

including cases where the same query text was done on a different search property such as a geographic-specific search interface. As Figure 3 suggests, the distribution of unique queries is similar between the own and assigned tasks, but the tail of the assigned tasks carries more weight than is apparent from the figure. (The figure also shows some number of subjects who did 0 queries – all cases where there was a data collection failure due to subject interpretation of the question or potentially technical problems). The mean number of unique queries on own tasks is 1.36 (median = 1), while the mean is 1.83 on the aggregated assigned tasks (with median = 1). This difference in mean does have reasonable significance in aggregate, with a t-statistic of 4.75 for the mean of own tasks being different than the aggregate mean for assigned tasks. However, as we can see clearly from Figure 3, the assigned product tasks were the most similar to own tasks. The difference in means between the two is a result of difference in the tail of the distribution. The probability that a session observed in this data with more than 4 unique queries belongs to the assigned task group is over 95%.

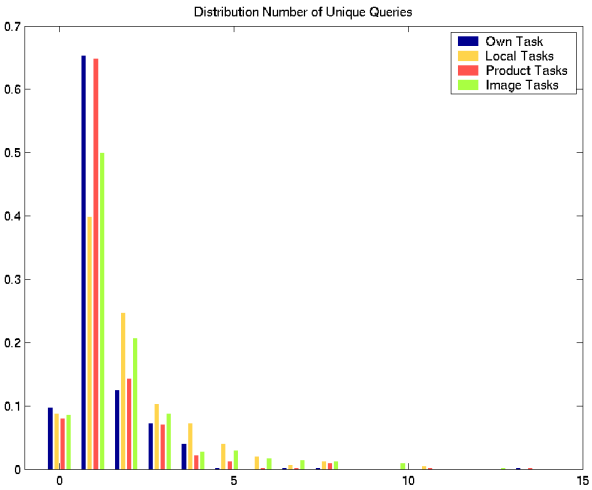

Figure 3. The fraction of queries that are distinct, separated by task type. Own tasks, like assigned product tasks, have a large number of only 1 and 2 unique queries in each session and far fewer in the tail (beyond 4 distinct queries per session).

C. Repeated query results page views are infrequent in assigned tasks: We also noted that on their own tasks, subject frequently return to the same search-engine results page (SERP), a condition we term "return-to-SERP." The apparent fact that the number of unique queries performed during the session on own tasks was relatively low appears somewhat at odds with the result in section *A,* but it turns out that users do fewer unique queries in own tasks, but spend more time viewing and revisiting the initial search page for a query.

When subjects do a return-to-SERP, they're going back to a search page they've already created, generally getting back to that page by clicking the back button a number of times. On average, as shown below in Section *E,* number of times the "Back" button is used in own tasks is also significantly larger.

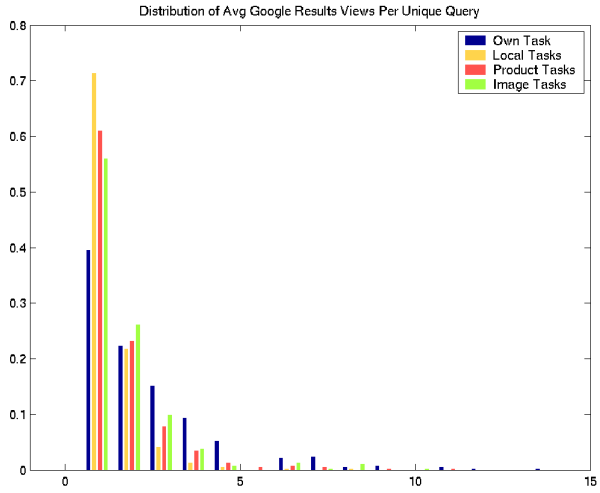

Figure 4. The number of return-to-SERP views is significantly higher on own tasks than for assigned tasks.

D. Fewer related queries in own tasks: Subjects doing own tasks do far fewer "related queries" (that is, queries that are not identical repeats of a first query, but overlap by at least 1 term). Related queries also include the case where a user repeated a query manually using a different type of interface such as a specific local or geographic search interface. As shown in Figure 5, 78% of all own task sessions have 0 related queries, while the assigned tasks have only 0 related queries 62% of the time.

The average number of subsequent related queries per session is 0.41 for the own tasks, and 0.79 for assigned tasks. However, the difference in mean comes from the fact that the two are fairly similar until the upper quantiles (where the number of subsequent related queries is greater than 1). The proportion of assigned tasks (that is, all 3 task types in aggregate) with more than 2 additional related queries is 8.8% (out of 1188 assigned task sessions). By contrast, the

proportion of own tasks with more than 2 unique queries is only 4.26% (out of 399 own tasks). We tested significance using pooled and worst-case variance. Since we know the distributions of the fraction of users with more than 2 additional related queries will have a moderately valid normal approximation, testing whether the proportion for the own tasks is the same and the proportion for the aggregated others. In this test, the t-stat is 4.51 or 4.52, which is significant at the 99% level.

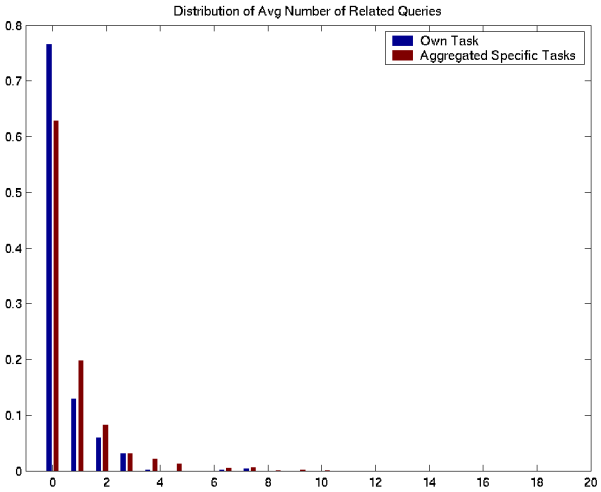

Figure 5. The number of related queries is far higher for assigned tasks than for own tasks. Note that the far left column shows that 78% of own tasks have very few (i.e., 0) related queries.

E. Own tasks use the back button more often: The back button is used to return to a previous page in the current search session. So it comes as no surprise that own tasks—which also have significantly more repeat-SERP views also have a higher level of back button use. Own task subjects used the back button on average 3.08 times per session, as opposed to the subjects doing assigned tasks which went backward only an average of 1.5 times per session. Given the number of page visits and the number of sessions we studied, this is a significant difference.

5. Discussion

The differences between own tasks and assigned tasks are consistent, somewhat subtle, but important. From this analysis it's clear than own tasks have longer sessions than assigned tasks while generating fewer unique queries (*and* fewer related queries).

From our hand inspections of the session logs it seems apparent that our subjects are simply more engaged in when working on their own tasks than tasks that have been assigned to them, no matter how benevolent the force that does the assigning.

So when doing their own tasks, subjects simply seem more focused on the task and dive deeper into the searches they perform. This leads to longer sessions and more clicks as the subject explores a topic of real interest to themselves.

On the other hand, fewer unique queries (that is, wholly new queries not seen before in that session) during the course of the user's own sessions suggests that own tasks are either clearer in the minds of the user, leading to a reduced need to generate unique queries, or that they simply can't think of any alternatives.

The fact that there are also fewer *related queries* suggests the same thing: own task session are characterized by a smaller range of possible alternate queries.

It also might be that users often browse after creating a search with the intent of recognizing their desired information when they see it, and not looking for a specifically pre-determined ideal document. [14] This is the kind of behavior we observe in field studies of searchers (i.e., the use of a search engine to locate another site, then browsing with recognition as an important part of the process), but have not yet adequately measured. (We leave this for future work.)

In addition, specific assigned task questions cue subjects with far more query terms to use and a very specific search termination criterion. From the user's self-reports on own task goals, their own tasks were much more loosely specified. A representative own task description was expressed as, "find a leather jacket," rather than "find a man's watch for a gift that costs less than $200" as was given in the assigned tasks.

Subjects doing own tasks spent more time overall, and tended to spend more time browsing through the results set rather than immediately digging for specific information. One possible explanation is that the subjects were defining their own task satisfaction criteria for the search while in the process of searching. (That is, as the searcher finds out more about a given topic, they continually refine their model of what would constitute a successful search.) Another possibility is that searchers "in the wild" simply don't define tasks as precisely as a typical assigned task.

Task engagement might also play a significant role in the differences we find. As [12] and [19] point out, a searcher with no real, engaged interest in an assigned topic might very well accept any plausible answer as

an acceptable one in the context of a test, even a test that pays well. A sense of personal engagement (or situational relevance [19]) with an own task almost certainly causes at least some of the differential behavior we see. This factor alone might account for the relative tendency of own task subjects to keep returning to the same SERP as they continue to drill ever more deeply into the topic area because they really do want to know the best answer to their own task, and not just one that will let them complete the test expeditiously.

There are ways to create test questions that require non-trivial amounts of search and information integration. [11] describes a test instrument to evaluate how well search tool users could pull together information from a number of sources in order to answer test questions. Such *integrative* questions might also cause user behaviors that emulate certain kinds of own task behaviors in terms of query patterns, time and return-to-SERPs.

6. Future

Clearly, understanding the nuances of search behavior will continue. We expect in the near future to add field studies results to this initial analysis to help us understand how the level of engagement and other markers of own tasks might be observable. In those field studies we anticipate being able to do both think-aloud protocols for own tasks in situ (to augment this study's own goal statement collection), and to do retrospective studies of time-lapse Web behavior captures where the subject relives the experience of a week's worth of searches, explaining what happened with that search, and why.

One of the big questions that emerged is why own tasks, which consume more session time, result in fewer related and fewer unique queries. Is this caused by the search-to-browse behavior we've seen elsewhere?

7. Summary

Assigned tasks are a common method of doing user testing of search interfaces and a common aspect of ordinary life. While they are a useful and straightforward method that are especially useful for eliciting feature use and highlighting problems in the interface, the behavioral consequences of assigned tasks are different from self-selected "own" tasks, with longer search sessions and qualitatively different query use. Consequently we recommend that own tasks be blended into the testing task mix whenever the goal of a study is to understand Web search behavior over the course of an entire session. Own tasks offer a look at the more open-ended behaviors of searchers, and can provide a view into users behavior outside the lab study environment.

8. Acknowledgements

We would like to thank our colleagues at Google, Ellen Konar and our partners at Keynote Systems for making this study possible.

9. References

[1] D. Beymer, D. M. Russell, P. Z. Orton
"Wide vs. Narrow Paragraphs: An Eye Tracking Analysis" Proc. INTERACT Conference, Rome, Italy. p 741-752. Sept. 2005.

[2] Goldberg, J.H., M. J. Stimson, M. Lewenstein, N. Scott, A. M. Wichansky. Proc. of the 2002 Eye Tracking Research Conference (ETRA), New Orleans, Louisiana. 51 – 58, 2002.

[3] Granka, L. A., Joachims, T., Gay, G. "Eye-tracking analysis of user behavior in WWW search" Proceedings of SIGIR'04. ACM Press, 478—479 (2004)

[4] Hotchkiss, G., S. Alston, G. Edwards. *Eye Tracking Study: An in depth look at interactions with Google using eye tracking methodology.* Enquiro technical report. http://www.enquiro.com/eyetrackingreport.asp June, 2005.

[5] Jansen, B. J., Spink, A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing & Management, 42(1), 248-263. 2005.

[6] Jansen, B. J., Pooch, U. "Web User Studies: A Review and Framework for Future Work." Journal of the American Society of Information Science and Technology, 52(3), 235-246. 2001.

[7] Kuniavsky, M. *Observing the User Experience: A Practioner's Guide for User Research.* Morgan Kaufman, San Francisco, CA. 2003.

[8] M. Mat-Hassan, M. Levene. "Associating search and navigation behavior through log analysis" Journal of the American Society for Information Science and Technology, v 56, 9 , 913 – 934. 2005

[9] Moukdad, H., & Large, A. "Users' perceptions of the Web as revealed by transaction log analysis" Online Information Review, 25(6), 349-358. 2001.

[10] Nielsen. J. "Authentic behavior in user testing" http://www.useit.com/alertbox/20050214.html Posted: Feb 14, 2005. (Last checked: June 15, 2006)

[11] Russell, D. M., Slaney, M., Yan, Q., Houston, M. "Being Literate with Large Document Collections: Observational Studies and Cost Structure Tradeoffs" HICSS 2006

[12] Spool, J. "Interview-based tasks: Learning from Leonardo DiCaprio." http://www.uie.com/articles/interview_based_tasks/ March, 2006. (Last checked: 6/15/06)

[13] Voorhees, E. M., D. K. Harman. (eds.) *TREC: Experiment and evaluation in information retrieval*. MIT Press, Cambridge, MA. 2005. Chapter 10: "Question answering in TREC."

[14] Draper, S. W., Dunlop, M. D., "New IR—New Evaluation: The impact of interaction and multimedia on information retrieval and its evaluation." The New Review of Hypermedia and Multimedia, 3:107--121, 1997

[15] Belkin, N. Oddy, R., Brooks, H. "ASK for information retrieval: Part 1. Background and theory." Journal of Documentation, 38(2), 61-71.

[16] Hsieh-Yee, I. "Effects of search experience and subject knowledge on the search tactics of novice and experience users." Journal of the American Society for Information Science, 44(3), 161-174. 1993.

[17] Spink, A., Park, M., Jansen, F., Pedersen, J. "Multitasking Web Search on Alta Vista" p 309, International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 1, 2004.

[18] Borlund, P. "The development of a method for the evaluation of interactive information retrieval systems" Journal of Documentation, 53(2), 225-250. June, 1997.

[19] Jansen, B. J. "The Wrapper: An open source application for logging user-system interactions during search studies" Workshop on Logging Traces of Web Activity: The Mechanics of Data Collection. 15th International World Wide Web Conference (WWW 2006), 22- 26. May, 2006.