

Eye tracking in MSN Search: Investigating snippet length, target position and task types

Edward Cutrell

Microsoft Research

1 Microsoft Way, Redmond, WA 98052

cutrell@microsoft.com

Zhiwei Guan

University of Washington

Box 352195, Seattle, WA 98195-2195

zguan@u.washington.edu

ABSTRACT

Web search services are among the most heavily used applications on the World Wide Web. Perhaps because search is used in such a huge variety of tasks and contexts, the user interface must strike a careful balance to meet all user needs. We describe a study that used eye tracking methodologies to explore the effects of changes in the presentation of search results. We found that adding information to the contextual snippet significantly improved performance for informational tasks but degraded performance for navigational tasks. We discuss possible reasons for this difference and the design implications for the better presentation of search results. The studies reported here are to be published in CHI 2007[5, 8].

Author Keywords

Web search, eye tracking, contextual snippets, user studies.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

As an increasingly large fraction of human knowledge migrates to the World Wide Web and other information systems, finding useful information is simultaneously more important and much more difficult. In 2000, Jansen and Pooch estimated that 1 in 28 Web pages that users viewed were search results pages [11]. Today, search is among the most important activities that Web users engage in. Beyond the Web, search is a central activity for users of corporate intranets, specialized databases (from shopping to Medline), and increasingly for personal archives of documents and email [4].

Given the importance and ubiquity of search, it is remarkable how similar almost all search interfaces are. Users typically type a few words into a query box and receive a rank-ordered list of search results comprising document titles, brief descriptions of the objects and some metadata about the results. On the Web, such interfaces are extremely effective, considering the incredibly wide range of tasks they are used for and the very short queries provided by most users. However, even given their simplicity, it is not obvious how users utilize different information from lists of search results to complete their tasks. Do users read the descriptions? Are the URLs and other metadata used by anyone but expert searchers? Does

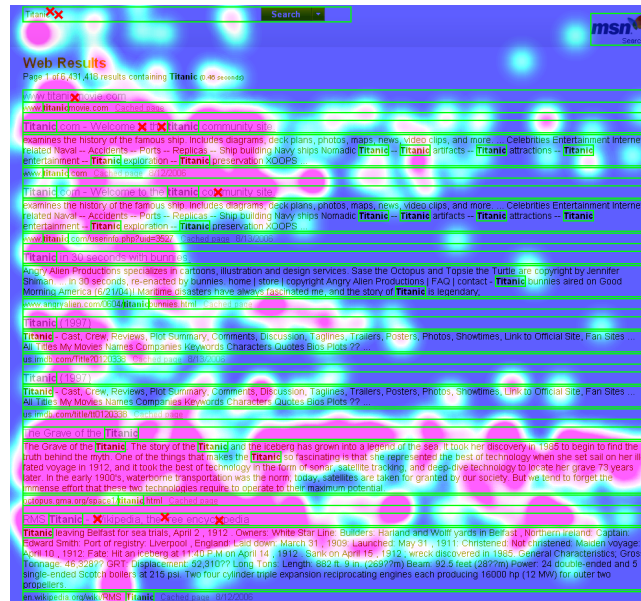


Figure 1. Heat map visualization of the number of fixations across 3 users on a page of search results for an informational task with long contextual snippets (see below). Boxes indicate defined areas of interest (AOIs).

the context of the search or the type of task being supported matter? Eye-tracking methodologies may help us to answer such questions by explicitly recording how users attend to different parts of Web search results. Figure 1 shows an example of users' fixation patterns for a page of Web search results in our study. For this task, users were clearly reading the contextual descriptions, especially on the seventh result.

Two or three broad classes of Web search tasks have been identified in the literature [1, 22] and used in various studies on Web search [12, 15, 17]. In *navigational* tasks, users are trying to find a specific Web site or homepage that they have in mind; the goal is simply to get to their destination. In *informational* tasks, the goal is to acquire some kind of information irrespective of where it may be located. For example, if a user is trying to find out the average June temperature in Caracas, he generally doesn't care where that information is found so long as it is reliable. Earlier research [17] reported that while informational tasks took longer to complete than navigational tasks, users also spent less time viewing search result pages for informational tasks (i.e., users spent more time looking for

information on destination sites). We thought that this might be because searchers do not have enough information on the search results page to make a good decision about where to go to find what they are looking for.

In this paper we describe an experiment that uses eye tracking techniques to help us understand how people use Web search to find information and whether strategies for scanning search results would be different for navigational and informational tasks. Do people view the same number of search results for different task types? Do they attend to different components of search results for navigational and informational tasks? Does the inclusion of more contextual information in search results help with informational tasks?

As noted above, most contemporary Web search engines return a few types of information with each search result: the document title, followed by a short bit of descriptive text, followed by the URL and perhaps some metadata or links (e.g., to cached pages or related links). The descriptive text is usually either a hand-authored description of the page or a *query-dependent contextual snippet*. Snippets are generated on the fly by the search engine from the text in the referenced page based on the query submitted by the user. Snippets are typically 1 to 2 lines of text and often contain sentence fragments from different parts of the documents.

Previous work in question-answering suggests that additional textual context to an answer provides a substantial boost to user preference and performance [16]. In the case of web searching, additional contextual information poses a tradeoff for searchers. On the one hand, even if the longer snippet didn't contain the information itself, we thought that increasing the size of the snippet might help users better decide whether a given result was likely to have what they wanted *before* they navigated to it. However, additional snippet length could also bear substantial costs. First, increasing the snippet length would reduce the number of search results that fit on the screen, forcing users to scroll to see the same number of results. Second, irrelevant search results also would include more information, and the "cognitive noise" associated with these snippets potentially could harm performance.

RELATED WORK

Understanding how users search for information on the Web has enormous practical implications for both commercial and academic endeavors. One of the most common techniques for studying Web search is examining search engine log files [18, 23]. Other researchers have used diary studies to explore the use (and limitations) of search engines in users' daily lives [24]. Researchers at PARC have done careful studies including the use of eye tracking and detailed activity-logging to develop user models for how people explore the Web [3].

As noted above, most Web-search interfaces are extremely simple. While there have been a number of attempts to enrich the UI for search results presentation, relatively few studies have attempted to evaluate how users interact with these new interfaces. Several careful studies have been done for interfaces that organize search results in different ways. "Faceted browsing" interfaces, in which search can be directed through dynamic filtering on orthogonal metadata properties (e.g., for an art database one might filter on date, artist, media, style, etc.) have been shown to be superior to traditional search interfaces for browsing tasks [25]. Other work has explored interfaces that dynamically categorize search results, grouping similar results together to aid directed search [7, 19]. In Relation Browser++, dynamic categorization and dynamic filtering and visualization of properties were brought together. This interface was significantly better for data exploration tasks than traditional form fill-in interfaces [26].

Eye-tracking methodologies seem particularly promising in the domain of Web search because gaze can be used as a proxy for a user's attention. While many techniques rely on the explicit actions of users (e.g., mouse clicks, query streams or diary reports), eye tracking can yield much more detailed moment-by-moment observations about how users interact with information. Because of this, eye tracking is particularly useful for developing user models (e.g., see [2, 8] for models involving search of computer lists and menus). Joachims, et al. [12] used eye-tracking techniques to characterize how users peruse search results. They then used these observations to inform measures of reliability for implicit feedback from clickthrough data in Web search.

Klöckner, Wirschum and Jameson [13] explored the order in which users examined search results before opening a document. They found that most people employed a linear strategy in which they evaluated each result in turn and decided whether to open the item before moving to the next result. A smaller number (15%) employed a different strategy in which most or all of the results were evaluated before a document was opened.

Lorigo, et al. [17] used measures of fixation, pupil dilation and sequence analysis of patterns of fixations (scanpaths) to look at differences in gender and task type for Web search. For task type, they found that informational tasks took longer to complete than navigational tasks. However, most of this time was spent on Web documents and not on search-results pages; users actually spent longer on the search-results pages for navigational tasks. They found no effect of task type on scanpaths (i.e., users evaluated search results in a similar way for both task types). In contrast, they did find a difference in scanpaths based on gender. Males tended to be more linear in the order in which they looked at results and looked at more results than females.

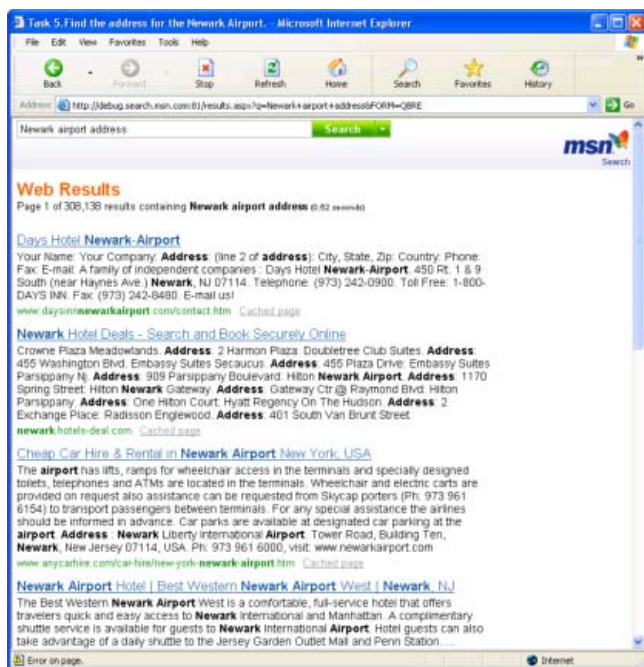


Figure 2. Screenshot of a search results page from the study. This example includes long query-dependent contextual snippets.

Understanding how users explore Web search results has large commercial implications as well. A number of companies have emerged that work with businesses in the area of search engine marketing. These companies help clients develop strategies for increasing traffic for their Web sites (e.g., “search engine optimization”). Detailed understanding of users’ behavior and expectations for Web search from eye tracking can be very valuable for these companies and their clients (e.g., [10]).

While most of the above work investigates Web searching with existing interfaces, none of these studies have examined how users respond to changes in the information provided to them. Joachims, et al. [12] is a possible exception to this because they did manipulate the order of search results for some of their users. The only work we know of that explicitly used eye tracking to explore differences in search interfaces compared a traditional list to a tabular interface for two informational and two navigational search tasks [20]. While no significant differences in performance were found, the eye-tracking measures did turn up a few interesting findings. In particular, they found that the mean number of fixations on the “summary element” (or snippet) for navigational tasks was higher for informational tasks across both interfaces. Unfortunately, this finding may have been driven by the fact that one of their navigational tasks was found to be especially difficult and may have required much more reading for selection confirmation.

EXPERIMENT

To investigate the effect of snippet length on how people use Web search, we designed our study to show results

Table 1. Example snippets of each length used in experiment for a single search result.

<p>Welcome to the Oklahoma City Zoo http://www.cpb.uhsc.edu/OKC/OKCZoo/</p>
<p>Short</p>
<p>The oldest zoo in the Southwest and one of the top in the nation, the Oklahoma ...</p>
<p>Medium</p>
<p>The oldest zoo in the Southwest and one of the top in the nation, the Oklahoma City Zoo's 110 acres are home to more than 2,800 of the world's most exotic animals.</p>
<p>Long</p>
<p>The oldest zoo in the Southwest and one of the top in the nation, the Oklahoma City Zoo's 110 acres are home to more than 2,800 of the world's most exotic animals." The Cat Forest/Lion Overlook was completed in 1997. New in 1993 was the Great EscApe , a simulated tropical forest with gorillas, orangutans and chimpanzees. Also found at the zoo are the Noble Aquatic Center: Aquaticus , a Children's Zoo and Discovery Area, Herpetarium, Island Life Exhibit, Dan Moran Aviary and the Safari Tram. Open 9-5 (Oct-March), 9-6 (April-Sept). Rides additional (weather permitting and seasonal). 2101 N.E. 50th Street Oklahoma City, OK (405) 424-3344 (OKC Zoo Phone Directory</p>

pages in various configurations. First, we presented results with three different snippet lengths (short, medium or long). In addition (for another set of questions), we simultaneously varied the position in the search results of the “best” search result for that task.

In our manipulations, short snippets usually contained a single line of words, medium snippets about two to three lines, and long snippets typically six to seven lines of words. For our browser and screen size, this meant that when we displayed results with short snippets, seven results were always at least partially visible on the first screen without scrolling. For medium snippets, there were an average of 5.7 (minimum of 5 and maximum of 7), and for long snippets, an average of 4.2 (minimum of 3 and maximum of 6) items were visible on the first screen. The screenshot in Figure 2 shows an example of a query with long snippets, and Table 1 shows all three snippets generated for a single search result entry. By default, MSN Search presents results with medium length query-dependent snippets. The short and long lengths were chosen to be realistic, but obviously different from the default lengths provided by MSN Search.

All manipulations were performed for two task types: *navigational* and *informational*. For our study, navigational tasks required the participant to find a specific Web page, and informational tasks required the user to find specific information that could be found in one or more places. In practice, tasks can vary widely in difficulty and, as seen in [20], this can have a major effect on performance. Therefore it was important to use several different tasks for each type and to counterbalance across tasks for all manipulations of interest.

Methods

Apparatus

All Web search queries were submitted to a special server for MSN Search (<http://search.msn.com>) that allowed us to control the length of the snippet information, but used the same methods for dynamic generation of snippets used in production servers (i.e., this used the same set of complex heuristics and algorithms used in MSN Search results). Search results were then dynamically intercepted by a proxy before being rendered to the browser. Advertising and editor-selected content at the top and side of the results page was removed, leaving only the standard UI associated with MSN Search and a list of 10 search results (see Figure 2). Note that because snippets were dynamically generated by the search engine based on indexed content, there were occasions when the snippet for a given item was considerably shorter than those of its neighbors.

Eye tracking was performed using the Tobii x50 eye-tracker (see, <http://www.tobii.se/>) paired with a 17" LCD monitor (96 dpi) set at a resolution of 1024x768. The eye-tracker sampled the position of users' eyes at the rate of 50Hz. An integrated log of eye movement data, user events and Web pages visited allowed us to map eye movements to various features on the screen during task performance. Areas of interest (AOIs) were generated by a javascript application that parsed the DOM for each page of search results that a user visited. This application provided us with the screen coordinates for each element that we were interested in for a given page (dynamically generated AOIs can be seen in Figure 1).

Participants

Twenty-two participants ranging in age from 18 to 50 years old with a diverse range of jobs, backgrounds and education levels were recruited for this study from a user-study pool. Of these, 4 participants were excluded from these analyses because of stability problems with the eye tracking and/or incomplete data, leaving us with 18 participants (11 male). All participants were moderately experienced at Web search, reporting that they searched the Web for information at least once a week, and all were familiar with several different search engines. None of them had experience using an eye-tracker.

Experimental design and procedure

The design of the experiment crossed *Task Type* x *Snippet Length* and *Task Type* x *Target Position* as independent within-subjects designs. For each participant, we randomly varied the order of the search tasks. Each of 12 search tasks (6 different tasks of each type) was counterbalanced across participants such that every task was seen with every snippet length (12x3=36 combinations) and every task was seen at every target position (12x6=72 combinations). We needed to counterbalance across tasks rather than just task types because of the large variability in individual tasks; since some tasks were easier than others, we needed to be able to average across all 6 tasks for each factor we are

Table 2. Search tasks (*queries*) used in study.

Navigational	
*	Find the homepage of the "Pinewood" software company. (<i>Pinewood</i>)
*	Find the homepage of the World Cup 2006 soccer games. (<i>World cup 2006 games</i>)
*	Find the homepage of Comfort Inn. (<i>Comfort Inn</i>)
*	Find the homepage of the National Weather Center. (<i>national weather center</i>)
*	Find the homepage of the St. John's law school. (<i>St Johns Law School</i>)
*	Find the homepage of the Yahoo! People Search. (<i>Yahoo People search</i>)
Informational	
*	Find when the Titanic set sail for its only voyage and what port it left from. (<i>Titanic</i>)
*	Find the address for the Newark Airport. (<i>Newark airport address</i>)
*	Find out how long the Las Vegas monorail is. (<i>Las Vegas monorail</i>)
*	Find out the name of the building that is Piano's most famous work. (<i>Renzo Piano</i>)
*	Find out the size (in area) of the Oklahoma City Zoo. (<i>Oklahoma City Zoo</i>)
*	Find the contact number for the Sylvan Learning Center. (<i>Sylvan Learning Center</i>)

interested in. For a detailed discussion of the experimental design, please see the Appendix.

We designed the search tasks for this study to be representative of common search tasks on the Web, varying in difficulty and topic. On a control page, we gave participants a brief query description and motivation for each task (e.g., *Renzo Piano is a famous architect. Find out the name of the building that is Piano's most famous work*) paired with a link of one or more query words that would launch a search when clicked (e.g., *Renzo Piano*).¹ See Table 2 for the list of all tasks and initial queries. After launching the initial query, participants were free to use the search engine however they chose to complete the task. Although participants generally agreed that the initial queries were reasonable for each task, they frequently submitted new queries if they felt they could not find what they were looking for with the query we provided.

All the results pages for the initial queries (those generated by the links) were cached locally so that we could be assured that all participants in a given condition would see exactly the same information at the beginning. All search-results pages to subsequent queries were generated on the fly as describe above.

¹ We provided the link and query terms for the initial query for every task because we wanted to control the first set of search results that every participant saw. While this does potentially threaten the ecological validity of our experiment, there is considerable benefit in making sure that all users see the same set of initial results.

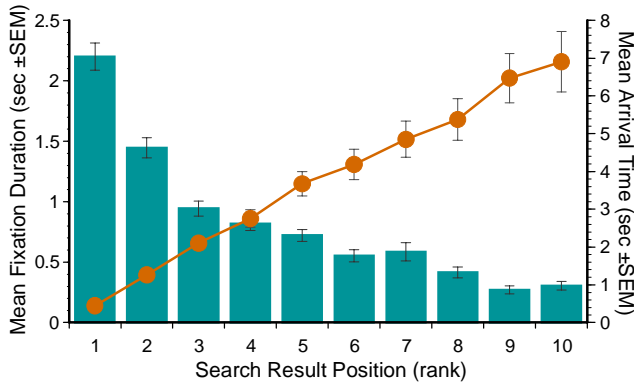


Figure 3. Mean fixation duration (bars) and mean time for gaze to arrive at each result (circles). As search results move downward in rank, it takes longer for searchers to look at them (upward trend of circles) and they spend less time looking at lower-ranked results (decreasing trend of bars). This figure is across all search results pages visited by participants. All error bars are \pm standard error of the mean.

For each query we generated, we made sure that the task could be completed with a site presented in the initial set of 10 results. For navigational queries, only one result was associated with the target, while for informational queries there was always one “best” result where a user could quickly find the searched information (e.g. the searched information was directly shown in the snippet, or the information was shown at a very obvious place on the Web page). However, as is common for informational queries, the task could often be completed by navigating to several different locations if the participant was willing to spend some time “orienting” around target sites (see [24]).

At the beginning of each session, participants were calibrated for the eye-tracker and given a practice query to familiarize them with the procedure. At the beginning of each task, participants read the task description and motivation in their Web browser and clicked the underlined query when they were ready. Each task was considered completed when the participant clicked on the target page, confirmed it was the desired site and vocally announced that they had found the Web site or information requested. Following the end of the study, participants answered a

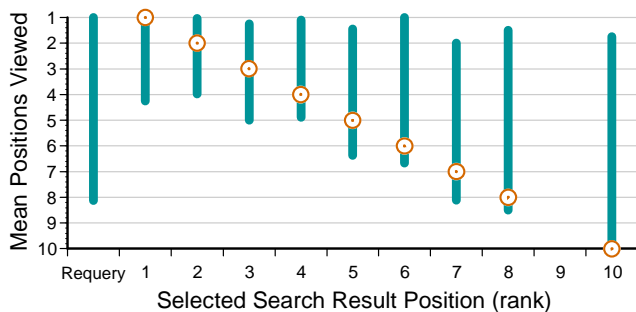


Figure 4. Mean number of search results looked at before users clicked on a result (above and below that result). E.g., if a user clicked on result 5, on average they looked at almost all items above it and about 1.4 results below it.

short questionnaire about their experiences in the study, some demographic information, and their past search history.

Results

Common eye-tracking measures include pupil dilation, fixation information and sequence information such as scan paths. For our analyses we relied on measures related to gaze fixations with a minimum threshold of 100 ms in areas of interest. Here we consider AOIs including each individual search result and each sub-element therein (e.g., title, contextual snippet, and URL).

In addition we looked at two non-gaze-related behavioral measures: *total time on task* (measured from when the first set of search results appeared on the screen until the participant announced they had finished), and *click accuracy* (whether a participant clicked on the “best” result from the first set of results).

General gaze characteristics for search results

Before describing the results of the various manipulations, we present some aggregate characteristics for how people view Web search results across all our search tasks and conditions. First, confirming previous findings [12], we found that people viewed search results in a roughly linear order. Most gaze activity was directed at the first few items; items ranked lower got users’ attention last and least (Figure 3).

We were also interested in the number of items viewed before and after a selected item. That is, if a user clicked on a result, on average how many other items above and below that item did they look at? This is interesting because it relates to how completely users search a set of results. Figure 4 shows that no matter what result they eventually clicked on, our participants usually looked at the first 3 or 4 search results. When they clicked on the first or second result, they still looked at the first 4 results. When they clicked on lower ranked results, they usually had looked at most of the items ranked above them. Finally, people go through about 8 results on a page before changing their queries without clicking on anything (indicated by “Requery”). With the exception of position 1, these results are very similar to findings reported by Joachims, et al. [12]. In their study, participants rarely looked at more than 1 or 2 items following the one that they clicked on, even when they clicked on the first item.

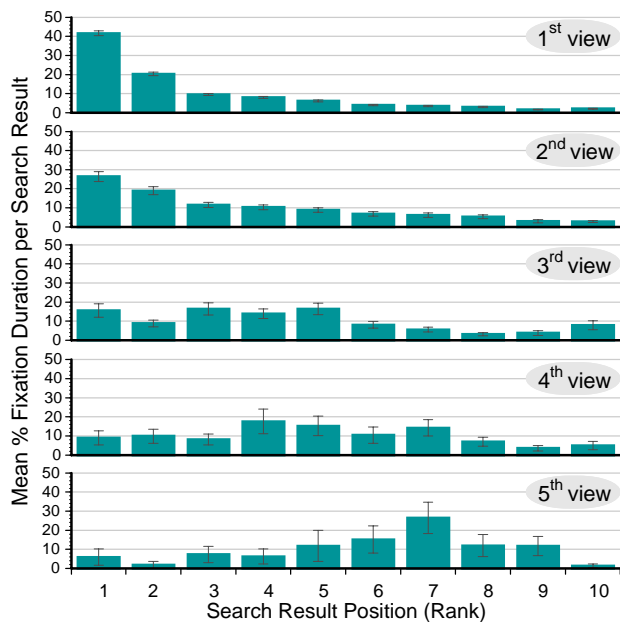


Figure 5. Mean percent of gaze fixation on each search result position broken down by visits to a given page of search results. Note that with subsequent viewing, searchers spend more time looking at lower-ranked search results.

A common observation in Web search is a “hub and spoke” pattern of exploration in which users go back and forth between a search results page and different target sites using the “back” button. We found that the distribution of users’ fixations also changed with subsequent visits to the search page (see Figure 5). In the first visit, higher-ranked items got the most attention, as described in Figure 3. When a user returned to a results page for a second examination, higher ranked items still received most attention, but the slope of the fixation distribution decreased with proportionately more time spent on lower ranked items. As the user returned to the page again, results 3, 4 and 5 became the main focus, and the focus steadily moved down the page with subsequent viewings.

Task Type & Target Position

To investigate the effect of task type and target position on how people search, total time on task and fixation measures were analyzed using 2 (*Task Type*) x 6 (*Target Position*) repeated measures analysis of variance (ANOVA). The click accuracy was analyzed using a chi-square analysis.

General Effects on Task Performance

We found a significant main effect of target position on the total time on task ($F(5, 85)=3.544, p=.006$); people spent more time on a task when the target is displayed at a lower position. We also found a main effect with query type, $F(1, 17)=54.718, p<0.001$, confirming findings in [13] that informational tasks took longer than navigational. There was no significant interaction between target position and task type.

While participants took more time to finish tasks when the target position moved down, it didn’t help them make

accurate selections. A chi-square analysis on the number of accurate clicks showed a significant effect for target position ($\chi^2(5)=58.5, p<0.001$). The click accuracy rate dropped from 84% (average of 78% and 89%) to about 11% when the target was displayed at position 8. Figure 6 shows that for navigational searches, people had the highest click accuracy rate when the target was on position 1 or 2 (78%, 83%). With the target on position 4, 5, 7, and 8, their click accuracy dropped to 33% or less. For informational search, the effect of target position on click accuracy was much more dramatic. When the target was displayed at lower positions (4, 5, 7, and 8), participants correctly selected the target for less than 20% of time. None of our participants correctly selected the target when it was at position 8. A closer look at Figure 6 shows a more interesting phenomenon: for navigational search, when people couldn’t find the target when it was placed at a lower position (4, 5, 7, or 8), they either clicked the first result (40% of the time, an average of 44, 28, 39, 50), or to switch to a new query (15% of the time, average of 11, 11, 22, and 17). For informational search, people rarely changed their query without a click (4% on average). Instead, they chose the first result over half the time, or randomly click on other results.

It is not surprising to see that when the target position was moved to the lower part of the results list, participants spent more time on the tasks yet achieved poorer accuracy.

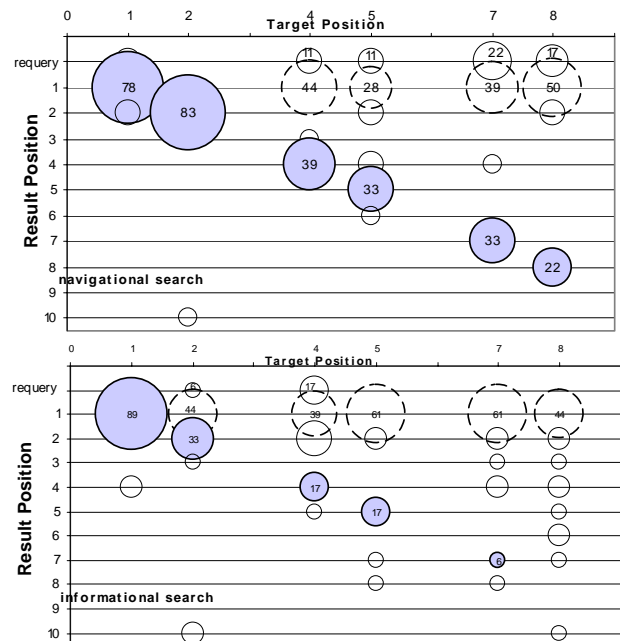


Figure 6. Chance of clicking on search results broken down with target position. The numbers inside the bubble indicate the chance (%) that the result was clicked (e.g. when the target position was 2 for navigational search, 83% of participants clicked on result 2, which was the target result.) The shadowed bubbles indicate the target results. The bubble with a dashed border indicates the first result. Bigger bubbles indicate a larger probability of clicking at the result at its particular position, which is also shown with a number inside the bubble.

However, we did not expect such a dramatic effect from placing the target low in the results list, particularly for informational search where participants achieved less than 10% accuracy. We hypothesized there might be two reasons for the general decrease of click accuracy across different tasks and the dramatic effect on informational search:

1) Since participants rarely went through the whole result list, they never saw the target result when it was placed at a low position, especially for informational search. Therefore, they couldn't find the correct target. This could be tested by looking at the number of results people fixated upon.

2) Alternatively, participants may have seen the target result for both navigational search and informational search, but they did not feel the result at lower position was as compelling as others. This could be tested by looking at the effect of task type and target position on fixation duration (an indicator of participants' attention.)

Examination of the gaze distribution helps us understand the dramatic difference in selecting target results depending on their ranks.

Did Users Look at Target Results?

A 2 x 6 ANOVA (see above for model) on the number of results participants fixated upon within the first page shows that there was a main effect of target position, $F(5, 85)=4.958, p=.011$. Participants went through more results (for position 1, mean=3.47, SE=.409; for position 8, mean=6.06, SE=.572) in order to complete the task when the target was placed lower. This indicates that participants sensed the fact that the top results were not correct and felt difficulties in finding the target when it was placed lower. No significant effect was found for *Task Type* or for the *Task Type x Target Position* interaction.

We looked further at the accumulated times that people fixated upon the target results (Figure 8). For navigational search, everyone looked at the target result when it was the first result (100%). When the target was position 2, this dropped to 89%, then to 72% for 4th, 56% for 5th, 7th and 8th. For informational search, the chance of looking at the target result drops a little faster from over 90% for position 1 and 2, to 22% for position 8 (see Table 3).

Table 3. Percent of people who looked at the target result (percent of people clicked on target) for navigational and informational search.

Target position	1	2	4	5	7	8
Navigational	100(78)	89(83)	72(39)	56(33)	56(33)	56(22)
Informational	94(89)	94(33)	89(17)	44(17)	39(6)	22(0)

This result supports the first hypothesis above, that the decrease probability of clicking on the target is related to the probability of looking at the target: if a user doesn't see a result, he won't click on it. However, this still doesn't explain the dramatic decrease in click accuracy for informational search: participants were fairly likely to look at the targets at positions 2 and 4, but were extremely reluctant to click on them (see Table 3). Is this because for

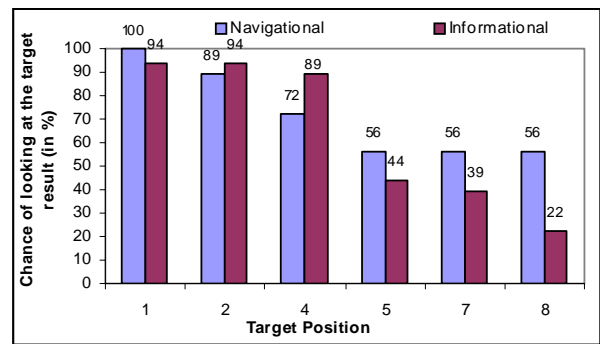


Figure 8. Chance of looking at target results. E.g., when the target position is 1, everyone looked at the target result for navigational search, and 94% of them looked at the target for informational search.

informational search participants put less attention on lower results even though they attended to them (hypothesis 2)? Further analysis of fixation duration, as we will now discuss, rejects this possibility.

How Much Attention Did Users Invest on the Target Results?

Our analysis of how long people looked at search results when the target results were at different positions leads us to suspect other reasons (e.g. high confidence in search engine) to explain people's reluctance to select the target results during informational search.

A repeated measures ANOVA found a main effect of the target position that the average fixation time on the target result decreases as the target position gets lower ($F(5,85)=7.06, p<0.001$). However, we found no main effect of *Task Type* and no *Task Type x Target Position* interaction effect. This means that people looked at the

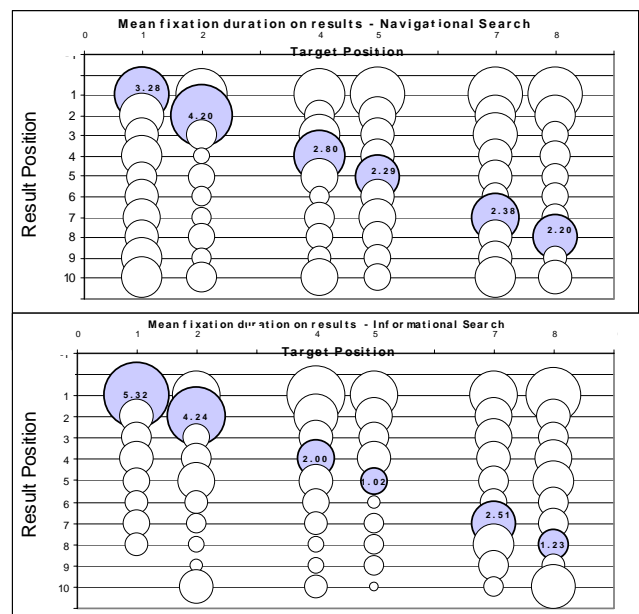


Figure 7. Fixation duration on results when people looked at them, break down by target position. The shadowed bubbles indicate the target results.

target in a same way for navigational and informational searches. Furthermore, the fixation duration on targets at lower positions decreased at the same rate for navigational and informational search.

This result shows that for informational tasks, people looked at the same number of lower-ranked results as they did for navigational tasks, but they clicked much less frequently on them. Figure 7 also indicates that people often lingered on the target results for informational search even though they didn't click on them. This suggests that users trust the search engine more for informational search or invest less scrutiny in judging the results with higher rankings. Eventually they are more likely to choose the top few results to try them out in spite of their lower objective relevance to the task. In the post-questionnaire, several responses from participants on their expectation on the search results also speak to this effect: they highly agreed on the statement, "I expect the information I'm looking for to be in the top five results" (mean=5.78, SE=.94, on a 7 point Likert-scale). Participants showed no preference on the statement "I often scroll to the bottom of the first page of search results looking for what I want" (mean=4.06, SE=1.63.)

Task Type & Snippet Length

To investigate the effect of task type and snippet length on how people search, we analyzed the following measures:

- Total time on task
- Total number of search results fixated for the task²
- Total summed duration of fixations on titles
- Total summed duration of fixations on snippets
- Total summed duration of fixations on URLs

For these measures, we performed a 2 (*Task Type*) x 3 (*Snippet Length*) x 2 (*Repetition*) repeated measures multivariate analysis of variance (RM MANOVA).

For *Task Type*, we found a significant main effect only for total time on task, $F(1,17)=54.7$, $p<0.001$. As in prior work [17], informational tasks took longer to finish than navigational tasks (78.1 s vs. 36.9 s).

As expected, there was no main effect for *Repetition*, suggesting that search strategies did not change as a result of time and experience with the tasks. Nor was there a main effect on any measure for *Snippet Length*. This indicates that averaging over both task types, changing the length of the query-dependent snippet had no effect on any measure of people's search behavior.

² When we calculate the total number of results fixated for a task, we sum all the search results a participant looked at across all the result pages visited for that task. For example, if a user looked at 5 results on the initial page, revised their query, and then looked at 6 on the new page, the total number of search results fixated for that task would be 11.

Table 4. F-values for Task Type x Snippet Length

Measure	F(2,34)	Sig. (p)
Time on task	4.4	0.02
# results fixated	5.6	0.01
Fix duration on titles	5.2	0.01
Fix duration on snippets	3.2	0.05
Fix duration on URLs	5.2	0.01

However, when we looked at the interaction between *Task Type* x *Snippet Length*, we found significant effects for all 5 measures analyzed (see Table 4).

Figure 9 illustrates the mean time on task for each task type as we varied the snippet length. For navigational tasks, the time on task remained the same for short- and medium-length snippets but increased by 10 seconds for long snippets. In contrast, informational tasks showed an *improvement* in task time of 24 seconds with long snippets.

If we focus only on informational tasks, this result supports the notion that more information in the snippet may help searchers determine whether a given site is likely to have the information they are interested in. To investigate this further, we also looked at the accuracy of our searchers' selections on the first query page where we know what the "best" result is. Figure 10 shows that as snippet length increased, the accuracy of clicks for informational queries increased from 28% to 39%. Because of the small number

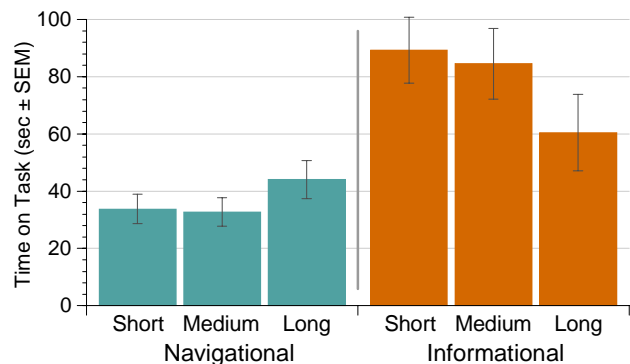


Figure 9. Mean time to complete search task for each task type, broken down by snippet length.

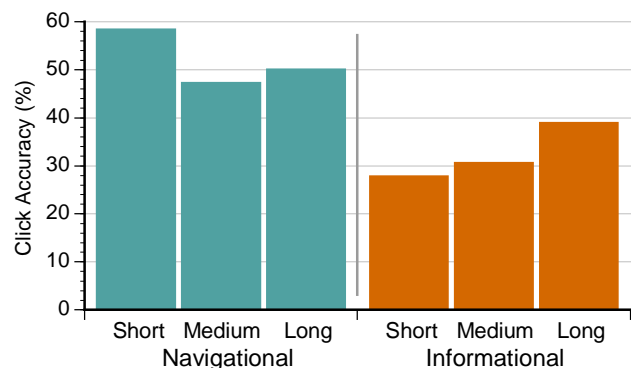


Figure 10. Accuracy in clicking the "best" result for each task type, broken down by snippet length.

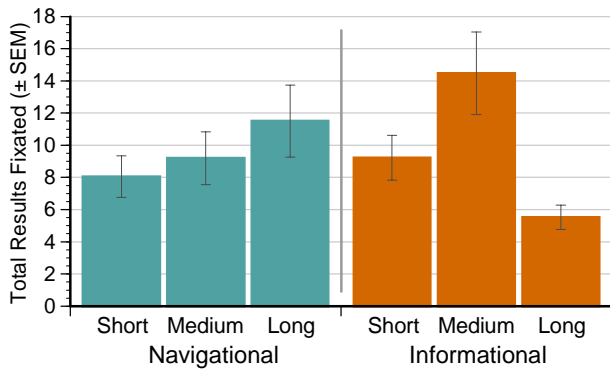


Figure 11. Mean number of search results fixated for each task type, broken down by snippet length.

of observations, statistical tests for this difference were not significant, but we believe that the trends provide converging evidence further supporting our hypothesis.

In stark contrast, Figure 9Figure 10 indicate that increasing the snippet length had exactly the opposite effect for navigational tasks. Long snippets increased the total time on task, and snippets of even medium length were associated with a *drop* in accuracy from 58% to 47%. In sum, users performed best on navigational tasks with short snippets and best on informational tasks with long snippets.

To better understand what might be driving this overall effect on task performance, we looked at the gaze measures provided by eye tracking. Figure 11 shows that when searchers were given short snippets, they looked at about the same average number of search results independent of task type (about 8 or 9 search results). However, as the snippet length increased, they began looking at more results for navigational tasks. In contrast, for informational tasks with the longest snippet length our searchers looked at a third fewer results than at the short length.

This finding presents a puzzle: it seems plausible that increasing the amount of information on the search-results page would result in looking at fewer results simply because there is more information to read: more lines of text in each result means that fewer results will be visible without scrolling. However, why would adding more information cause one to look at *more* results? And why

Table 5. Mean fixation duration (with SEM) for each component of search results, broken down by task type and snippet length.

	Navigational		
	Short	Medium	Long
Title	3.36 (0.44)	4.31 (0.95)	5.49 (1.27)
Snippet	2.72 (0.52)	5.10 (1.29)	7.85 (1.90)
URL	2.93 (0.54)	3.25 (0.77)	3.32 (0.71)
	Informational		
	Short	Medium	Long
Title	5.56 (1.09)	6.68 (1.68)	3.61 (0.81)
Snippet	4.38 (0.82)	7.51 (1.54)	6.54 (1.34)
URL	3.79 (0.76)	4.16 (0.9)	1.47 (0.39)

would this effect be task-dependent? One explanation could be that adding more information would lead a searcher to be more thorough, spending more time with search results because the information density is higher. Therefore they just spend more time reading the results and less time reading Web documents. But if this were true, we should find that users look at more results in both informational and navigational tasks. What’s going on here?

One possibility is that because the goal of navigational tasks is locating a specific site, the information provided by contextual snippets is much less relevant for navigational than for informational tasks where details related to site content, authority, etc., are more important. In contrast, URLs may be proportionately *more* relevant for navigation because they are directly related to the location of target sites. If this were true, we would expect that searchers would spend proportionately more time looking at the URL in navigational than informational tasks. In our study, this was true but the difference was small: across all snippet lengths, people spent 25% of their time looking at the URL in navigational tasks vs. 22% in informational tasks. However, if we break this down by each snippet length, a pattern begins to emerge. Figure 12 shows the relative proportion of total fixation duration for each search result component (title, snippet and URL) broken down by snippet length and task type (for reference, the mean fixation duration for each condition is shown in Table 5).

Examination of Figure 12 shows that as we increased the snippet length, the relative time looking at the snippet also increased for both task types. However, while the proportion of time looking at the title stayed roughly constant, the increase in time looking at the snippet came primarily at the cost of looking at the URL. This decrease was particularly dramatic for the informational tasks, but it was true for navigational tasks as well. Figure 12 suggests that when our participants looked at search results with long snippets, they read them, whether or not the snippets were relevant to their task.

Post-experimental questionnaire

After the experiment, participants answered a short questionnaire with questions about demographic information and their Web search experience. In general, our participants appeared to be quite savvy at Web search;

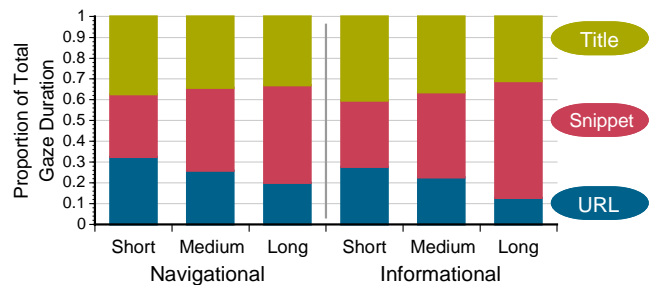


Figure 12. Proportion of total fixation duration for each component of search results, broken down by task type and snippet length. As snippet length increases, the relative proportion of gaze devoted to the URL decreases.

most reported that they typically search the Web at least once a day, and all were familiar with and had used a variety of different search engines. Google was by far the most popular search engine, but several participants also reported using Yahoo! as their primary search engine. One reported using the AOL default engine as his primary search engine.

As part of the questionnaire, participants were given several 7-point Likert-scale questions of the nature, “Click 1 if you completely disagree, 7 if you completely agree and 4 if you neither agree nor disagree.” Of particular interest were answers to the following questions:

“The search terms automatically selected for each task were usually close to what I would have entered myself for that task.” For this question, the median score was 6 and the mean was 5.8. This was important, because a possible concern was that the query terms that we created would be so different from what users would generate on their own that their behavior would be unnatural. These scores suggest that for most of our participants, the query terms were reasonably close to what they would spontaneously generate.

The next two questions were also very interesting: *“When I’m searching the Web, I often look at the URL of each search result to help me decide if the page will be useful.”* And: *“When I’m searching the Web, I usually read the snippet (text under the title) to help me decide if the page will be useful.”* For these questions, the median scores were 7 and 6 respectively, and the means were 6.4 and 6.2. These answers suggest that our participants deliberately use various elements in the search results to help them find what they are looking for. We were particularly surprised to see the overwhelming endorsement of the URL because this is often characterized as a “power-user” feature that is used by only a small percentage of users.

DISCUSSION AND DESIGN IMPLICATIONS

Task Type & Target Position

Web search engines have become commoditized tools for finding information in our daily lives. Most search engines display the search results in a rank ordered list, with the closest matched results placed on top and others ordered below that. However, this display has the potential side effect that users may not utilize more relevant results displayed at lower positions on the list.

This study showed that people spent more time on tasks and were less successful in finding target results when targets were displayed at lower positions in the list. When people could not find the target results for navigational search, they usually selected the first result or switched to a new query; for informational search, people rarely issued a new query and were more likely to try out the top-ranked results despite their lower objective relevance.

Further eye movement analysis showed that people mainly look at the results on top of the list and this explains the

uniform decreases in click accuracy for both navigational and informational search. The analysis also suggests that the large decrease in performance for informational search was likely due to the strong confidence on search engine performance even though people know that target results at lower positions *can* be relevant. People are more likely to degrade their own judgment of target results and trust the ranking determined by the search engine.

Task Type & Snippet Length

This experiment presents designers of Web search engines with something of a dilemma. Our results showed that changing the user interface of Web search results by varying the length of the query-dependent contextual snippet had opposite effects on task performance depending on what a user was trying to do. For navigational tasks, optimal performance occurred with short snippet lengths, while for informational tasks, long snippets helped the most.

The different task types in our study involved very different informational needs from the search results. All of our navigational tasks required participants to find a specific destination Web site. For these queries, the URL was likely to be a very useful source of information. This is not to say that the title and snippet were irrelevant—indeed in the condition that proved best for these tasks, searchers looked at the title, snippet and URL almost equally (Figure 12). In contrast, the URLs of search results were probably much less relevant for our informational tasks, because these tasks could be answered by any of a number of Web sites. So long as the destination site referenced by the search result looked authoritative and contained sufficient information scent, the searcher could be satisfied going to the result and looking for an answer from there.

If we take gaze fixation as a proxy for users’ attention, we can start to explain what is going on. For both task types, as the snippet length gets longer, the relative attention to the snippet also increases. However, this increase comes at the cost of relative attention to the URL. The proportion of attention on the title also decreases somewhat, but the decrease is quite small. For informational tasks, where the URL is less relevant, the cost in task performance and information scent is minimal; the attention paid to the longer snippets more than makes up for any cost from missing the URL. However for navigational tasks, the information in the longer snippets is not as relevant, and the cost of the lack of relative attention to the URL is more acute. Even though the total amount of time spent looking at URLs did not vary much for navigational tasks (see Table 5), we believe that the increased amount of attention to the longer snippets interfered with the information located in the URL, decreasing the URL’s relevance for our users.

If users applied attention equally to any information in the search result, more information in the snippet could actually decrease their certainty that a result was the target, and they

would continue looking at other results. That is, as more information is included in the results, users may unconsciously down-weight the relevance of URLs for their decisions. When multiple results have rich snippets, it would be more difficult to decide which result is the target for navigational tasks, but this wouldn't be an issue for informational tasks, where the goal is any Web site likely to have the answer. This idea suggests that users may not consciously realize the benefit they receive from URLs and do not strategically devote attention to different parts of results depending on their task, but rather simply use what they are given.

This hypothesis is testable in a few ways. First, we could perform a similar study, this time removing the URL entirely from the results. This would address the hypothesized importance of the URL for different task types. A more subtle variation would be to carefully choose a subset of navigational targets to be Web sites that have URLs with little information (e.g., hosted by a generic provider, or with very long GUIDs in them). In this case, we would expect increased snippet lengths to have little effect (or perhaps a *positive* effect) on the performance because the URL simply isn't useful for these tasks.

Another way to explore this hypothesis is to vary the attentional salience of different elements in the results list. If the above explanation is correct, we should be able to improve performance in navigational tasks by increasing the attention to navigational information such as URLs. Conversely, we could harm performance by emphasizing other, pseudo-relevant information. In either case, performance should directly correlate with the amount of gaze devoted to (ir)relevant information.

Our results suggest that for a substantial fraction of queries in Web search (informational tasks), extended snippets are useful. Despite having to scroll more, accuracy and task times were improved, and users actually looked at fewer total items. For another large class of Web search tasks (navigational), long snippets are problematic. However, our results suggest several possible solutions to this problem. In search results provided by MSN Search (and all major search engines), the URL is always placed at the bottom of each search result, immediately following the snippet. As seen in Figure 12, when the snippet is only a single line, all three elements receive almost equal attention as searchers linearly scan the results from top to bottom. However, as the snippet length grows, searchers begin to lose the URL in the mass of text. It would be very interesting to place the URL below the title, immediately above the snippet. This would guarantee that as users scan the results they would always see the URL before the snippet. When the URL is an important navigational aid, it would easily be seen; likewise, a single line of stylized text would be fairly easy to ignore if the useful information is in the snippet below it.

Another solution would be to radically alter the design of the results presentation to interrupt the linear top-to-bottom

scanning. One might divide the display, placing the snippet in a dedicated pane to the right of the title, URL and other metadata. This would de-emphasize the snippet, but it would still be available for detailed examination.

Both of the above solutions assume an "all things for everyone" design where search engine providers present results in a single style for all queries. Another solution would be to provide different information for different kinds of queries. This could be an explicit gesture of intention by the user (e.g., in the trivial case, a button), or using automatic classification [14, 15]. If the search provider is able to determine reliably that a user is engaged in an informational task, it could provide results with richer content. Likewise, for clearly navigational tasks, it could minimize such content while emphasizing navigational information.

Our results also might have implications for search domains outside of Web search. There are many domains in which informational search is the primary activity (e.g., medical and academic databases). For these domains, our results suggest that long contextual snippets can greatly improve a user's search experience. Similarly, there are other domains such as directory searches, where navigational tasks clearly dominate. For these domains, users may be better served by brief snippets and search results that emphasize navigational information such as the URL or other location context.

CONCLUSIONS AND NEXT STEPS

We presented a study using eye tracking techniques to investigate user strategies for Web search. We looked at how people respond to search results when the target is systematically manipulated to be displayed at different positions on two kinds of search tasks and found that users seem to exhibit an implicit trust for the rank generated by the search engine, particularly for informational tasks. In addition, we looked at how varying the amount of information in Web search results affected user performance on the same tasks. We found that as we increased the length of the query-dependent contextual snippet in search results, performance improved for informational queries, while it degraded for navigational queries. Our eye tracking results suggest this difference in performance was due to the fact that as the snippet length increased, users paid more attention to the snippet and less attention to the URL located at the bottom of the search result.

Web search is a very attractive domain for the use of eye tracking techniques, and we believe this study is only a prelude to a wide range of interesting studies in UI for information retrieval. For example, the experiments outlined above would provide excellent information about how users deploy their attention when they view search results. Similarly, it would be interesting to verify whether or not moving the URL above the snippet would improve users' experience in navigational search. There are many

kinds of metadata that are potentially useful for Web search. How can this information be presented to users in such a way that is complementary to existing information in search results?

In addition, this study raises interesting theoretical questions about how our results might be situated with respect to information foraging theory [12]. It would be an interesting exercise to fit our data to the concept of information scent.

The future of Web search interfaces will probably be very different from we see today [20]. Studies like those outlined here can help to inform what they will look like. Finally, we would like to perform similar studies in other search domains to see whether our findings apply outside of Web search (e.g., search in corporate intranets, medical databases, personal desktop indices, etc.).

APPENDIX

Details of Experimental Design

Eye tracking experiments are often associated with a fairly high degree of attrition of participants. Because of this, we designed our experiment to require a minimum of 6 participants, to be repeated as frequently as we had time and participants. In the end, we collected usable data from 18 participants, yielding 3 repetitions of the complete design. Table 6 shows the full design for the experiment. Along the left margin are listed each of the 12 tasks (6 informational and 6 navigational), and each column (A-F) represents each of the 6 participant groups. For each cell, we show the snippet length (S, M or L) and target position (1, 2, 4, 5, 7 or 8). For example, when each participant in group C performed navigational task 3, they saw short snippet lengths and the target result was placed in position 7 (S, 7). Please note that for each participant we randomly varied the order of the search tasks.

As seen in Table 6, the design of the experiment crossed *Task Type* x *Snippet Length* and *Task Type* x *Target Position* as independent within-subjects designs. Each of 12 search tasks (6 different tasks of each type) was counterbalanced across participant groups such that every task was seen with every snippet length (12x3=36 combinations) and every task was seen at every target position (12x6=72 combinations). We needed to counterbalance across tasks rather than just task types because of the large variability in individual tasks; since some tasks were easier than others, we needed to be able to average across all 6 tasks for each factor we are interested in. Although all task types, snippet lengths and target positions were counterbalanced (every snippet length and target position co-occur in at least 2 navigational and informational tasks), we could not fully counterbalance all tasks, snippet lengths and target positions (12x3x6=216 combinations). This means that each of the *Task Type* x *Snippet Length* and *Task Type* x *Target Position* designs are complete, but we cannot test for interactions between *Snippet Length* and *Target Position* because the variance

Table 6. Full design of experiment, showing assignment of experimental conditions for each group and task. Each participant group has 3 participants for the purpose of repetition. Each cell contains the snippet length (S, M, or L) and target position (1, 2, 4, 5, 7, or 8).

		Participant Group					
		A	B	C	D	E	F
Tasks	Nav-1	(S,1)	(L,2)	(M,4)	(S,5)	(L,7)	(M,8)
	Nav-2	(M,2)	(S,4)	(L,5)	(M,7)	(S,8)	(L,1)
	Nav-3	(L,4)	(M,5)	(S,7)	(L,8)	(M,1)	(S,2)
	Nav-4	(S,5)	(L,7)	(M,8)	(S,1)	(L,2)	(M,4)
	Nav-5	(M,7)	(S,8)	(L,1)	(M,2)	(S,4)	(L,5)
	Nav-6	(L,8)	(M,1)	(S,2)	(L,4)	(M,5)	(S,7)
	Info-1	(S,1)	(L,2)	(M,4)	(S,5)	(L,7)	(M,8)
	Info-2	(M,2)	(S,4)	(L,5)	(M,7)	(S,8)	(L,1)
	Info-3	(L,4)	(M,5)	(S,7)	(L,8)	(M,1)	(S,2)
	Info-4	(S,5)	(L,7)	(M,8)	(S,1)	(L,2)	(M,4)
	Info-5	(M,7)	(S,8)	(L,1)	(M,2)	(S,4)	(L,5)
	Info-6	(L,8)	(M,1)	(S,2)	(L,4)	(M,5)	(S,7)

for individual tasks is confounded with this interaction. So, for example, the combination of short snippets and target position 1 only occurs in tasks 1 and 4. The large variation in task difficulty could potentially skew the effect of interactions between these two factors, so we cannot reliably test this. While it is remotely possible that a complicated 3-way interaction of *Task*, *Snippet Length* and *Target Position* could skew the results of an individual factor, we feel that this is quite unlikely considering the robustness of the effects we report.

ACKNOWLEDGMENTS

We thank Susan Dumais, Dan Liebling and Muru Subramani for their extensive assistance and intellectual horsepower. In addition we would like to thank all the participants who spent an hour searching the internet on our made-up tasks.

REFERENCES

1. Broder, A. A taxonomy of web search. *SIGIR Forum*, 36, 2(2002), 3-10.
2. Brumby, D.P. and Howes, A. Good enough but I'll just check: Web-page search as attentional refocusing. *Proc. 6th Int'l Conference on Cognitive Modeling*, Lawrence Erlbaum (2004), 46-50.
3. Card, S.K., Pirolli, P., Van Der Wege, M., Morrisson, J.B., Reeder, R.W., Schraedley, P.K. and Boshart, J. Information scent as a driver of web behavior graphs: Results of a protocol analysis method for web usability. In *Proc. CHI 2001*, ACM Press (2001), 498-505.
4. Chi, E., Pirolli, P., Chen, K., & Pitkow, J. Using information scent to model user information needs and

- actions and the Web. In *Proc. CHI 2001*, ACM Press (2001), 490–497.
5. Cutrell, E. & Guan, Z. What are you looking for? An eye-tracking study of information usage in Web search. Accepted for publication in *CHI 2007*.
 6. Cutrell, E., Dumais, S.T., & Teevan, J. Searching to eliminate personal information management. In *Communications of the ACM (Special Issue: Personal information management)*, 49 1(2006), 58-64.
 7. Dumais, S.T., Cutrell, E. and Chen, H. Optimizing search by showing results in context. In *Proc. CHI 2001*, ACM Press (2001), 277-284.
 8. Guan, Z., and Cutrell, E. An eye tracking study of the effect of target rank on Web search. Accepted for publication in *CHI 2007*.
 9. Hornof, A.J., and Halverson, T. Cognitive strategies and eye movements for searching hierarchical computer displays. In *Proc. CHI 2003*, ACM Press (2003), 249-256.
 10. Hotchkiss, G., Alston, S. and Edwards, G. Eye Tracking Study. <http://www.enquiro.com/eyetrackingreport.asp>.
 11. Jansen, B.J. and Pooch, U. A Review of Web Searching Studies and a Framework for Future Research. *Journal of the American Society of Information Science and Technology*, 52 (2000), 235–246.
 12. Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. Accurately interpreting clickthrough data as implicit feedback. In *Proc. SIGIR 2005*, ACM Press (2005), 154-161.
 13. Klöckner, K., Wirschum, N. and Jameson, A. Depth- and breadth-first processing of search result lists. In *Ext. Abstracts CHI 2004*, ACM Press (2004), 1539-1539.
 14. Lau, T. and Horvitz, E. Patterns of search: analyzing and modeling web query refinement. *Proc. Seventh Int'l Conference on User Modeling*, Springer-Verlag (1999), 119–128.
 15. Lee, U., Liu, Z., and Cho, Junghoo. Automatic identification of user goals in web search. In *Proc. WWW 2005*, (2005), 391-400.
 16. Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D.R. The role of context in question answering systems. In *Ext. Abstracts CHI 2003*, ACM Press (2003), 1006-1007.
 17. Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., and Gay, G. The influence of task and gender on search and evaluation behavior using Google. *Info. Processing and Management: an Int'l Journal*. 42, 4 (2006), 1123-1131.
 18. Mat-Hassan, M. and Levene, M. Associating search and navigation behavior through log analysis. *J American Society for Information Science and Technology*, 56 (2005), 913-934.
 19. Pratt, W. and Fagan, L. The usefulness of dynamically categorizing search results. *J American Medical Informatics Association*, 7 6(2000), 605-617.
 20. Rele, R.S. and Duchowski, A.T. Using eye tracking to evaluate alternate search results interfaces. In *Proc. HFES, 49th Annual Meetin.*
 21. Rose, D.E. Reconciling information-seeking behavior with search user interfaces for the web. *J American Society for Information Science and Technology*, 57 (2006), 797-799.
 22. Rose, D.E. and Levinson, D. Understanding user goals in Web search. In *Proc. WWW 2004*, (2004), 13-19.
 23. Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. Analysis of a very large AltaVista query log. *SRC Technical note #1998-14*. <http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html>. (1998).
 24. Teevan, J., Alvarado, C., Ackerman, M.S., and Karger, D.R. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proc. CHI 2004*, ACM Press (2004), 415-422.
 25. Yee, K.P., Swearingen, K., Li, K., and Hearst, M. Faceted metadata for image search and browsing. *Proc. SIGCHI 2003*, ACM Press (2003), 401-408.
 26. Zhang, J. and Marchionini, G. Evaluation and evolution of a browse and search interface: relation browser++. In *Proc. CHI 2005*, ACM Press (2005), 179-188.