Eye Tracking and Online Search: Lessons Learned and Challenges Ahead

Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, and Geri Gay
Cornell University Information Science, Ithaca, NY 14850, USA. E-mail: {lal2, hb47, lx33, gkg1}@cornell.edu; {maya ,tj}@cs.cornell.edu

Laura Granka
Google, Mountain View, CA 94043, USA. E-mail: granka@google.com

Fabio Pellacini
Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA. E-mail: fabio@cs.dartmouth.edu

Bing Pan
School of Business and Economics. College of Charleston, Charleston, SC, 29424, USA. E-mail: panb@cofc.edu

This article surveys the use of eye tracking in investigations of online search. Three eye tracking experiments that we undertook are discussed and compared to additional work in this area, revealing recurring behaviors and trends. The first two studies are described in greater detail in Granka, Joachims, & Gay (2004), Lorigo et al. (2006), and Pan et al. (2007), and the third study is described for the first time in this article. These studies reveal how users view the ranked results on a search engine results page (SERP), the relationship between the search result abstracts viewed and those clicked on, and whether gender, search task, or search engine influence these behaviors. In addition, we discuss a key challenge that arose in all three studies that applies to the use of eye tracking in studying online behaviors which is due to the limited support for analyzing *scanpaths*, or sequences of eye fixations. To meet this challenge, we present a preliminary approach that involves a graphical visualization to compare a path with a group of paths. We conclude by summarizing our findings and discussing future work in further understanding online search behavior with the help of eye tracking.

Introduction

Broad access to an abundance of information is one of the defining characteristics of today's environment. Internet search engines act as intermediaries between users' informational needs and the massive number of potentially relevant pages on the Web. To best design search engines, as well as understand the fundamental behaviors involved in the online search process, a closer look at what users are doing when they search online is needed.

One of the tools that helps us delve deeper into user behavior during online search is eye tracking. Eye tracking has been used to study human behavior for decades and has contributed to our understanding of activities such as reading, scanning, and overall processing of visual stimuli. Many of these contributions remain highly informative in the context of online search, specifically fundamental knowledge about reading, visual search, and cognitive load (Byrne, Anderson, Dougless, & Matessa, 1999; Hornof, 2004; Rayner, 1998). During interaction with a scene, our eyes make a series of *fixations*, or spatially stable gazes each lasting for approximately 200–300 milliseconds in which our eyes are focused on a particular area, and *saccades*, or rapid eye movements that occur between fixations lasting 40–50 milliseconds with velocities approaching 500 degrees per second. Because it is believed that little to no information is captured and processed during these rapid saccades, saccades are typically ignored in eye-tracking analyses, which is also the case for the studies discussed here.

Commonly used eye-tracking metrics, such as pupil diameter, fixation duration, and the number of fixations on an object or area, help us to determine user engagement and mental processing. In an online context, this translates to information about which components of a page are viewed and how often, and the cognitive load or engagement present upon viewing those components, which correlates with

Received April 19, 2007; revised October 14, 2007; accepted October 15, 2007

 • Published online 14 March 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20794

larger pupil size (Rayner 1998; Hess & Polt, 1960). A less commonly reported eye-tracking metric known as the *scanpath* reveals the order or sequence of the fixations, helping to depict the experience a user has while engaging with Web pages or, specifically, search engine results pages (SERPs), as well as inform design decisions for both layout and work flow.

In this article, we survey in depth how eye tracking has offered new insights to investigations of Internet search. In addition to discussing related work, we describe three studies that were undertaken by the authors. The first two studies are summarized and used for comparison and context because they are described in greater detail in Granka et al. (2004), Lorigo et al. (2006), and Pan et al. (2007). The third study is described for the first time in this article. Following an overview of these studies, we describe a key challenge that arose during our experiments. This challenge relates to the difficulty in analyzing *scanpaths*, or sequences of eye fixations that are captured by eye trackers. We describe current analysis approaches as well as our preliminary solution for overcoming this challenge.

Related Work

The application of eye tracking to online search has recently received a considerable amount of attention from research scientists, search engine companies, marketing firms, and usability professionals. While our discussion focuses on eye tracking for the purpose of understanding online search, a broad survey of eye-tracking usage across other application domains can be gathered from Duchowski (2002), Jacob and Karn (2003), and Rayner (1998).

Studies of online search behavior analyze who searches the Web, what tasks they perform, their perceptions of Web search tools, and how they perform searches (Hsieh-Yee, 2001). In an early study, an online survey, focus groups, and logs of telnet sessions were used to find out who used the Internet through an online catalog and information system and how they used it (Tillotson, Cherry, & Clinton, 1995). Another study looked into demographic data of the users of a particular search Web site, their search topics, search terms, and strategies (Spink, Bateman, & Jansen, 1999). Researchers have also investigated the users' cognitive, affective, and physical behaviors as the users sought to answer a fact-finding task (Bilal & Kirby, 2002). And, Teevan, Alvarado, Ackerman, and Karger (2004) conducted a version of diary studies, intermittently asking users what they had recently looked for to understand how users search for information on the Web, and they found that users vary between directly targeting their information need and reaching it in a number of small steps.

Eye-tracking methodologies, compared with the above-mentioned research methodologies, enjoy incomparable advantages in the field of Web search. Instead of relying on users' self-reported data, eye-tracking technology allows researchers to record users' real-time eye movements. This helps in understanding how users evaluate abstracts returned from a search engine, how choices are made, and how answers to queries are found on a page. In addition, eye-tracking methods enable informed interpretation of Web log files as feedback data, and thus they eventually influence and improve search engine performance and design. That is, if a log file shows that a particular advertisement was selected over another, eye tracking can then show if this was a conscious preference by revealing whether both advertisements were even seen.

Before the conception of Internet search, MacGregor, Lee, and Lam (1986) discovered specific search patterns on menus using eye tracking, which they named terminating, exhaustive, and redundant. In spite of this being a helpful distinction in search patterns, it is also important to note that search patterns may depend on the type of search being performed. To understand the different uses of search engines, Broder (2002) developed a Web search taxonomy that classifies the "need behind the query" into three classes: navigational, informational, and transactional. Navigational tasks are those in which the user's intent is to find a particular Web page. Informational tasks arise when the intent is to find information about a topic that may reside in one or more Web pages. Transactional search tasks reflect the desire of the user to perform an action, such as an online purchase.

Josephson and Holmes (2002) used eye movement measures to record participants' viewing behavior when presented with three different kinds of Internet pages: a portal page, an advertising page, and a news story page over the course of a week. They concluded that viewers' eye movements may follow a habitually preferred scanpath and that features of Web sites and memory may be important in determining those scanpaths. In the context of online search, we are also interested in understanding how habitual a users' scanpath is on a SERP.

The first use of eye tracking to investigate Internet search behaviors comes from Granka et al. (2004), to the best of our knowledge. This research analyzed users' basic eye movements and sequence patterns throughout 10 different search tasks performed on Google. We also considered how these behaviors were influenced by gender and also by task classification in (Lorigo et al., 2006). The task classification in that study was inspired by Broder (2002) and all assigned tasks were classified as either an informational or navigational task. The findings motivated us to undertake a second study, which manipulated the rank of various search results to explore further the impact of rank on behavior. The Studies 1 and 2: Internet Search Behaviors on Google section summarizes these two studies.

In 2005, search marketing firms Enquiro and Did-it, in collaboration with Eyetools, an eye-tracking company, performed a study that revealed a specific pattern of eye movements during the time participants were evaluating search results on Google (Sherman, 2005). This pattern mimics an "F" shape, with the eyes scanning the top of the page horizontally and then scanning downwards. This pattern has also been termed the "Golden Triangle," where the bottom of the viewing triangle extends only as far as the third or fourth result. Nielsen's Alertbox (Nielsen, 2006) also reports

DOI: 10.1002/asi

evidence of a dominant F-shaped pattern for eye movements exhibited while people read Web pages in general, based on a study of 232 users. The implications of these works tell us that the first two paragraphs or few results are the most viewed and thus must include the most important information of the Web page. Also, for information lower on the page, it is particularly critical that the first word or two catches the user's attention because it is likely to be all that is scanned initially.

In looking at different strategies that people employ for scanning search results depending on the type of task (navigational versus informational), Guan and Cutrell (2007) found that adding information to the contextual snippet of a search result significantly improved performance for informational tasks but degraded performance for navigational tasks.

Studies 1 and 2: Internet Search Behaviors on Google

Study 1 was the first to use eye tracking for investigation of online search, to the best of our knowledge. It was motivated by a need for a deeper understanding of online search behaviors and extended prior research in that domain by revealing which abstracts are viewed on a SERP, how they are viewed (i.e., sequentially or not), and how task and user characteristics may influence these behaviors. In addition to looking at general behavior for the entire data set, we also studied the results according to gender and task type to look for significant differences. After preliminary analysis, we became interested in whether internal or external forces are more influential in predicting online search behaviors and in what ways. Thus, we chose gender as one characteristic with which to model a user while also using task type as an alternative, nonpersonal factor. In summary, is user behavior more variable due to intrinsic factors, such as gender, or is it more task dependent, such as between informational and navigational tasks?

The importance of ranks one and two witnessed in that study led us to further investigate the role that search engines' rankings play in the information-seeking process. In Study 2, we were interested in determining how a user's choice of a particular abstract was influenced by (a) the rank of that abstract, (b) the user's perception of the relevancy of that abstract, or (c) a combination of the two. In that study, we actually manipulated the position of the abstracts (i.e., reversed the order returned by Google) to determine the extent to which position versus perceived relevance contributes to behavior, or if there is some level of "trust" in the result ranked in position one, for example.

After discussing methodology and design, we report important findings from the two studies. While summaries are provided, the reader is encouraged to refer to the referenced articles for more detail.

Research Methods and Design

For each of our earlier studies, we utilized the ASL Eye tracker (Applied Science Laboratories, 2005) together with the accompanying software, GazeTracker, for capturing the pixel coordinates of eye gazes. For Study 1, 36 undergraduate participants at Cornell University were recruited but difficulties with initial equipment set up and configuration resulted in complete eye-tracking data for 23 participants. Of these participants, 14 were male and 9 were female with ages ranging from 18 to 23 years. For Study 2, 22 undergraduate participants were recruited but equipment configuration errors resulted in complete data for 16 participants. Of these participants, 11 were male and 5 were female. The average age was 20 years, and in both cases the students represented a range of disciplines including communications, computer science, and humanities. In Study 1, nearly all participants reported a very high expertise with the Google search engine in a pretest survey, with an average expertise of 8.8 out of 10, and similarly, in Study 2, all reported a very high expertise of 10 out of 10. Because of the student sample used in both studies, our participants are likely to be, on average, somewhat more experienced with online search than a randomly selected sample of the population.

In both studies, each participant was read aloud 10 query tasks in random order, listed in Appendix A, and given 2 minutes to answer each task. These tasks were chosen to be representative of typical queries, inspired in part by Google Zeitgeist (Google Inc., 2005), a service from Google that highlights popular search terms, and also to vary in level of difficulty and task type, including both informational and navigational tasks as defined by Broder (2002). In Study 1, for example, this resulted in analysis of eye-tracking data for a total of 437 queries. All query pages went through a proxy server and Google ads were stripped from all pages before the participants viewed them to avoid adverse and varying effects on our study. For experimental purposes, it is highly desirable to have similarly structured query result pages that allow valid comparisons and analysis. Hence, in all three of the studies discussed here, advertisements were first removed.

Other than the removal of ads, no results were manipulated in Study 1. Instead, Study 2 consisted of three conditions to which participants were randomly assigned. The first, or normal, condition is equivalent to the setup of Study 1, with search results listed in their original order on the SERP. The second, or reversed, condition reversed results 1 through 10 so that the search result ranked 1 by Google moved to position 10 and vice versa. Third, we included a swapped condition that swapped only results of rank one and two. This condition was used to further explore the finding in Study 1 in which users nearly equally looked at results one and two but clicked on result one significantly more often. This was a less severe manipulation than in the reverse condition, and one that could lend specific insight into the rapid decision making involved in selection of the top ranked results. The participants were unaware that any manipulations were performed.

Furthermore, for Study 2, in order to better understand the influence of perceived relevancy as compared to ranking, we gathered relevance assessments of the abstracts for all SERPs evaluated in the study. Five nonparticipants were

DOI: 10.1002/asi

chosen as relevancy judges in the study. Judges were presented with the 10 query result abstracts in random order, and they were asked to order the results according to their believed likelihood that the result would lead to relevant information. Agreement between judges was also considered, and interjudge agreement on the perceived relevancy of the abstracts was 82.5%.

Study 1 Results

What is the relationship between rank and whether or not a result is viewed? Consistent with subsequent related work, the order in which search results are presented (result ranking) is extremely important when it comes to whether an abstract is viewed. In an impressive 96% of the queries, participants looked at only the first result page, containing the first 10 abstracts, and no participant looked beyond the third result page for a given query. Participants looked primarily at the first few results, with nearly equal attention given to results one and two. Despite nearly equal viewing times on those two result abstracts, participants clicked on results one and two 42% and 8% of the time, respectively.

Females, however, clicked on the second result twice as often as males. Specifically, males clicked on the second result only 7% of the time and females selected it 14.5% of the time. On the other hand, males were more likely to click on lower ranked results, from entries 7 through 10. As it happens, in our experimental environment, results one through six can be visualized without scrolling, while results 7 though 10 require scrolling. However, further investigation revealed that this characteristic was not solely due to scrolling. Furthermore, males looked beyond the first 10 results significantly more often than women. In our study, males were five times more likely than females to view additional Google result pages.

How well informed is the user before making a selection or refining a query? In roughly 40% of the cases, users chose to revise their query terms or view additional query results without clicking on a returned search result. This does not necessarily mean that in such cases Google did not return a useful result in its top 10 result entries. In fact, on average only three abstracts from each result page were viewed. The time spent on a given result page was often short, with the average number of fixations per result page being 16. On more than half of the result pages, users chose to revise their query terms without clicking on any abstract. However, it is important to note that eye tracking does not tell us how much users perceive in their peripheral field; to the best of our knowledge, nearly no literature studying peripheral vision exists from which we can effectively extrapolate to the context of online searching.

Still, from the patterns of eye fixations, it is clear that users are not reading each abstract in full but instead skimming them. Accordingly, bolding keywords and query terms in the search result page is likely to be advantageous.

Is the result list explored sequentially or otherwise? How users evaluate the search results and then decide upon their next action, whether to select a result or reformulate a query, is important to the search process and design. Only in one-fifth of the cases were users observed to read the results in the order in which they were presented.

In order to understand how the query result list is explored, we looked carefully at the scanpaths, or the sequences of eye fixations across the page. Rather than representing scanpaths as sequences of pixel coordinates, we treated scanpaths as sequences of ranked results, where pixels were mapped to results ranked 1 through 10. This approach is common, and the grouped regions of pixels are termed "areas of interest" or "lookzones" on the Web pages.

The average length of the scanpath, represented as a sequence of ranked results (including repetitions), was 16. If we remove repeated visits to search result abstracts, the resulting scanpath reveals that on average 3.2 distinct abstracts were viewed following each query. On those Google result pages in which a document was not selected, only 4% of the scanpaths contained all 10 abstracts. Hence, users may be using clues from the returned results that they do see to determine that it is in their best interest to reformulate the query rather than explore the results already returned.

Essentially, our scanpath analysis seems to indicate that most often, if none of the top three results are relevant, then the user does not explore further results.

What behaviors are more influenced by gender than task and vice versa? Our study found that time to complete the task varied by the type of search task (informational or navigational) but not by gender. The average time to complete informational tasks was 46 seconds, compared with 34 seconds for navigational tasks ($p < .05$). For informational tasks, participants also spent a greater proportion of their time away from the SERP, searching for their answers on the traversed Web pages whereas navigational queries can often be answered directly from the SERP. On informational tasks, subjects spend 60% of their time away from SERPs compared with 48% for navigational tasks ($p < .05$).

Pupil dilation, which shows a level of engagement or concentration, was found to be equivalent for informational tasks and navigational tasks on the SERPs but greater for informational tasks on the external Web documents (52 mm versus 46 mm, $p < .05$). No pupil dilation difference was found due to gender on either SERPs or external pages.

On the contrary, viewing patterns for evaluating the query result page were found to be strongly influenced by gender but not by task type. We observed, for example, that females made more regressions or repeat viewings of already visited abstracts while males were more linear in their scanning patterns. The process of searching within the search results and how users determine which result to click on is important in SERP design. Interactive interfaces for highlighting potential results or placing check boxes for elimination might improve efficiency, for example, and are based on this process.

DOI: 10.1002/asi

Success rates measured by whether the tasks were correctly completed in the time allowed were not significantly different amongst task types or gender.

Study 2 Results: Effects of Perceived Ranking

Did the modifications to the SERPs effect the success rate of participants? The success rates per task were significantly different among the three conditions. In the Reversed condition, participants completed search tasks successfully only 62% of the cases, compared with 85% and 80% in the Normal and Swapped conditions, respectively ($p < .05$). Forty-three percent of queries resulted in immediate reformulation of the query prior to selecting any result. This behavior did not vary significantly between the three conditions, so users did not immediately disregard the SERP in the reverse condition in favor or trying another query.

How did the Reversed condition effect results viewed? First, participants in the Reversed condition spent considerably more time viewing the SERPs. On average, participants in the Reversed condition spent 11 seconds on a SERP compared to roughly 6 seconds spent by participants in both the Normal and the Swapped conditions. It also follows that participants in the Reversed condition made more fixations (30 fixations on average compared with 18 in both the Normal and the Swapped conditions) and viewed a greater number of result abstracts (3.8 abstracts versus 2.5 in the Normal condition and 2.7 in the Swapped condition). These indices, combined with relevancy judgments suggest that placing less relevant abstracts in the higher ranked positions results in more scrutiny and comparisons before making a decision. This also shows that rank position alone is not a predictor of evaluative behaviors on the abstracts.

Is a user's choice of a particular abstract based solely on its rank? In Study 1, we saw that users viewed results one and two roughly equally, but they clicked on result 1 significantly more often. We can investigate this behavior further by comparing the Normal and Swapped conditions. In both the Normal and the Swapped conditions, the frequency with which participants viewed the top two ranked abstracts was roughly equal, similar also to the behavior observed in Study 1. However, in both of these conditions, participants clicked on what they perceived to be the first ranked abstract roughly three times more often than the abstract that was perceived to be ranked second. Thus, while the top two abstracts may both be scrutinized and plausibly relevant, the abstract in the rank one position is favored during click behavior. This suggests that rank alone can influence behavior. In contrast, rank does not solely influence viewing because participants in the Reversed condition were more likely to view the abstracts that were perceived to be of lower rank (i.e., abstracts positioned in places 5 through 10) than in the other conditions. However, similar to the Normal and Swapped conditions, participants in the Reversed condition still clicked on the top five presented abstracts significantly more often than the bottom five abstracts. Hence, participants in both the Swapped and Reversed conditions had their behavior affected by our manipulation of the results' order.

How does rank position compare to the intrinsic relevance of an abstract? Relevance was determined by the human judgments on the abstracts. A mixed model analysis was used to explore the significant influences on viewing and clicking. Position, relevance, and experimental condition were all significant determinants of whether an abstract was viewed and also whether an abstract is clicked on. In both models, values revealed that when all factors are considered, position is more influential than relevancy, though both had an impact on behavior.

Study 3: Comparison of Google and Yahoo! Viewing Behaviors

One important question that remained from the previous studies was whether the observed behavior by participants when using Google would extend to other search engines. In our most recent study, we performed extended experiments comparing behavior when using Google to that when using the also popular Yahoo! search engine. Our initial results are meant to both further validate and extend the results obtained in the previous studies, providing basic understanding as to whether users propagate the scanning behavior across varied search engines or whether behaviors differ in some respects. Similarly, we would like to understand whether familiarity or expertise with a particular engine influences users' viewing and click behaviors when performing a search task.

Even though the search interfaces provided by Google and Yahoo! present visual differences, their basic structure regarding how queries are submitted and how results are presented is very similar, leading us to believe that overall behaviors will not vary greatly. However, at the time of this research, no studies had revealed SERP viewing behaviors on engines other than Google, and for generalization, it was critical to examine this. Furthermore, recall that Study 2 exhibited that rank influences viewing behavior even when the result orders are compromised. In that study however it was also the case that all subjects reported an expertise of 10 out of 10 with the Google search engine. If trust or loyalty to a given search engine was influencing the observed bias toward perceived rank, then would the same be observed when subjects use an alternative search engine? Hence, by comparing results on Google and Yahoo! in a similar way to our previous studies, we strengthen our understandings gained from earlier work by both validating overall behavior across different search engines and looking more closely at the notion of user loyalty as it may relate to bias towards selecting abstracts that are perceived to be ranked highly by the search engine.

Study 3. Methodology

The same equipment described in the Research Methods and Design 35 section was used for this study. However,

DOI: 10.1002/asi

prior difficulties with the sensitivity of the eye-tracking equipment were overcome in the third study, and a robust sample of 40 undergraduate and graduate students of various concentrations at a large university in the Northeast of the United States served as participants in the third study. The average age of participants in this study was 20.7 years, with 22 males and 18 females. All participants but one reported using Google as their primary search engine. Meanwhile, the reported average of how often participants chose Yahoo! for searches was 2.9 (out of 7).

Participants were randomly assigned to use either Google or Yahoo! and asked to complete 10 predefined informational search tasks (listed in Appendix B). Each participant had 2 minutes for each search task and was instructed to go on to the next task if they believed to have found a satisfactory answer or if the time limit expired. The tasks were presented on a computer monitor next to the participant's computer to control for typing errors and language difficulties. Search pages resulting from mistyped queries were not included in our analysis. Similar to the earlier two studies, all query pages went through a proxy server and ads were stripped from pages from both search engines before the participants viewed them.

Normal and Reversed conditions were defined, as in Study 2. In the Normal condition, participants were presented with the search query results in the same order as returned by the search engines, after removal of advertisements. This condition provided us with an understanding and comparison of search behavior between the two different search engines (i.e., task success rate, rank influence, fixation patterns, and depth of ranked results viewed) in uncompromised conditions. In the Reversed condition, results were presented in the opposite order of what the search engine would normally present. The purpose of evaluating both Google and Yahoo! in this latter condition was to verify whether the importance of result ranking previously observed with Google also applies to Yahoo!, and to understand whether the determining factor responsible for user reliance on ranking is trust or familiarity with a search engine. To facilitate our analysis, we conducted both pre- and post-task surveys. The earlier survey asked about familiarity with the two search engines, and the post-task survey asked about ease and satisfaction with each task.

Study 3. Results

Is there a difference in eye movement patterns between Google and Yahoo! users? Our comparison of basic ocular indices between the two search engines included the number of fixations and fixation duration, as well as other typical eye-tracking measures (Table 1). The slight variations observed between participants using Google versus participants using Yahoo! are not significant. An interesting point of note is that the average number of abstracts seen in the Normal condition is fairly low (2.89 and 2.61 for Google and Yahoo!, respectively), an indication that users tend to reformulate queries quickly if they do not find the information they are looking for within a short span of time. In the Reversed condition, Google users presented a slightly higher average (3.60) compared to Yahoo! (2.68), but again the difference is not significant.

The number of abstracts clicked per page is less than one in all cases, which is a further indication of user reformulation. In the Google Normal condition, participants clicked on 75% of all the search result pages viewed, whereas in the Yahoo! Normal condition, users clicked on 61% of all the search result pages viewed. We found the average time spent on result pages before either the first click or query reformulation to be 6.29 seconds for Google Normal and 6.25 seconds for Yahoo! Normal. This timing is similar to that found in Study 2. There was also no significant difference between the average time spent on pages on which a ranked result was selected and those on which the query was reformulated instead. Hence, Yahoo! and Google users spent roughly the same amount of time inspecting search result abstracts before choosing to reformulate their query.

Is there a difference in perceived ease of use and satisfaction as well as success rates depending on search engine? In terms of behavioral measurements for comparing the search engines, we looked at the user experience and success rates. As part of the post task evaluation survey, participants were asked to rate each task, using a seven-point Likert scale, in terms of how easily they found the relevant information and how satisfied they were with the performance of the search engine they were using. Participants using Google reported averages of 5.23 (Normal) and 4.6 (Reverse) for task ease, whereas participants using Yahoo! reported averages of 5.04 (Normal) and 4.92 (Reverse). There was no statistically significant difference between the groups with respect to ease of the task. Considering satisfaction, participants using Yahoo! reported an averages 5.74 (Normal) and 5.01 (Reverse), compared to 5.48 (Normal) and 4.88 (Reverse) for Google, but again an independent t-test showed no significance. Overall, we can conclude that both groups found that their searching experience was equally easy and satisfactory.

TABLE 1. Average ocular indices for participants by search engine.

Ocular indices	Yahoo! Normal	Google Normal	Yahoo! Reversed	Google Reversed
Number of fixations per page	37.61	37.59	42.73	42.86
Fixation time per page in seconds	11.23	10.33	10.72	11.10
Duration of fixations in seconds	0.43	0.29	0.28	0.29
Number of abstracts seen per page	2.61	2.89	2.68	3.60
Number of abstracts clicked per page	0.77	0.99	0.71	0.96

DOI: 10.1002/asi

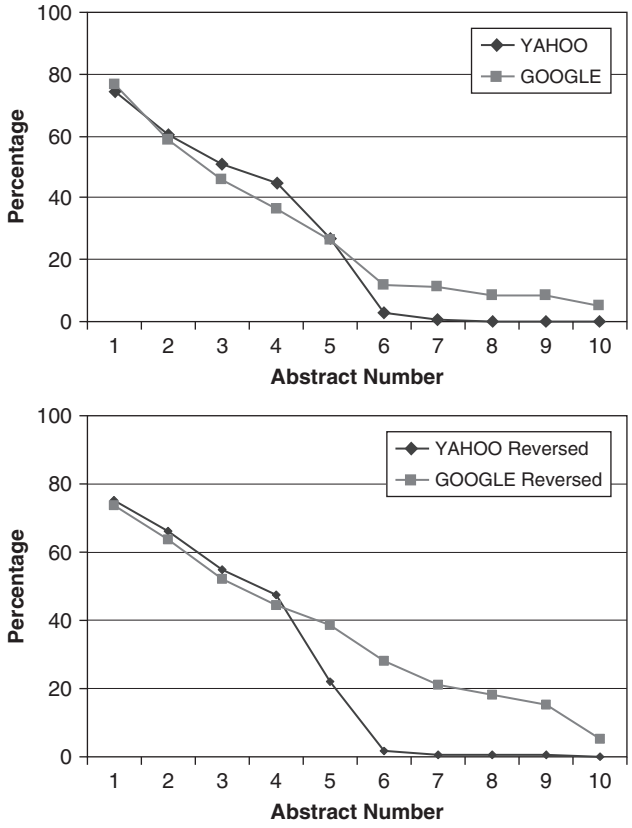

FIG. 1. Percentage of total pages in which abstract was viewed.

In terms of success rates, we looked at whether each participant could find the relevant information within the 2-minute time limit or not. In the Normal condition, participants using Yahoo! were slightly more successful on average, completing 95% of their tasks overall, whereas participants using Google were successful in 91% instances. This difference in averages was not significant. In the Reverse condition, the percentages were lower (87% for Yahoo! and 83% for Google), reflecting the increased difficulty levels.

What is the relationship between rank and whether a result is viewed? Similar to results observed in Study 1, participants looked past the first result page in less than 2% of the queries, and most of these cases contained only as far as the second result page. Only in one instance did a participant view search results past the second page, reaching the fifth result page.

Figure 1 presents the relationship between the rank order and how often a result was viewed by participants in both Google and Yahoo!. The graphs present the percentage of total pages in which a user viewed each particular abstract. Again, our findings were consistent with previous work. Independent of the search engine being used, rank is extremely important when it comes to whether an abstract is viewed. For example, Google and Yahoo! participants in both Normal and Reverse conditions mainly viewed the first few results. The first two ranks received most attention in both groups (76.7%, 56.6%, respectively, for Google Normal; 74.5%, 60.6%, respectively, for Yahoo! Normal). There is a significant drop in the number of results viewed for Yahoo! Normal at the fouth abstract appearing on the results page and for Google Normal at the sixth abstract. Considering this effect, we also observed that, on average, Google displayed a greater number of abstracts than Yahoo! given the same screen space; in our experimental setup, approximately 6 Google abstracts and 5 Yahoo! abstracts were visible prior to scrolling. Because users are typically not quick to scroll, the choice of placement and page real estate given to each abstract can affect how far down in the result list a user chooses to view. In our study, no participant from the Yahoo! Normal group viewed results positioned below the seventh result, whereas some participants from the Google group viewed all 10 ranked results.

Interestingly, the difference in later results viewed between Google and Yahoo! is even more prominent in the Reverse condition. While almost no abstracts past the sixth result are viewed in Yahoo! Reverse, Google Reverse suffers a steady decline all the way until the 10th result. This difference in behavior may indicate that lack of familiarity, under compromised conditions, leads users to rethink their queries earlier when using Yahoo! and to be more patient when using Google.

Similar patterns can be observed in Figure 2, which presents the relationship between the rank order and how often a result was clicked or not. In the Google Normal condition, participants clicked on results one and two 50.4% and 21.1% of the time, respectively. In the Yahoo! Normal condition, participants clicked on results one and two 40.3% and 20.2% of the

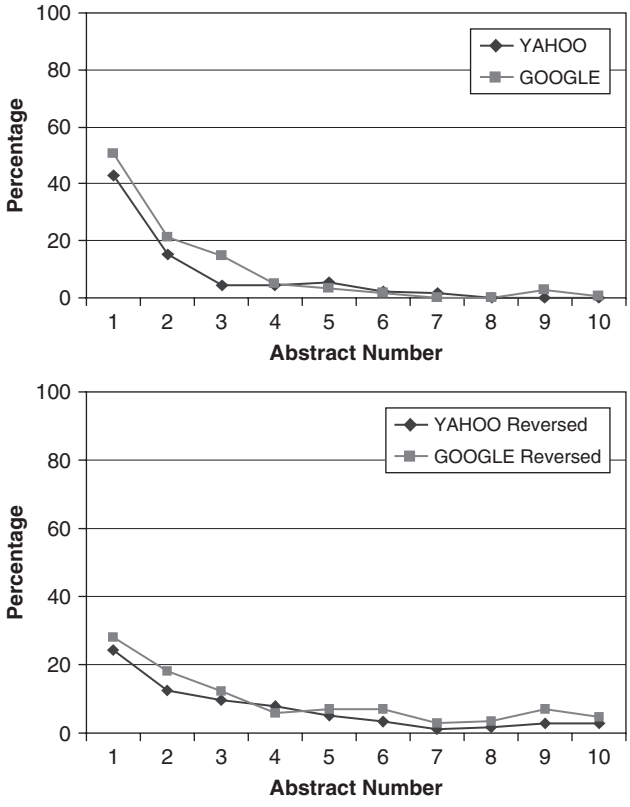

FIG. 2. Percentage of total pages in which abstract was clicked.

DOI: 10.1002/asi

time, respectively. There is a significant drop in results clicked for Yahoo! at the third returned abstract and for Google at the fourth abstract. Results after rank 7 were rarely clicked by participants in the Google group and never clicked in the Yahoo! group. In the Reverse condition, the number of clicks on result one is reduced for both Google and Yahoo! participants, further indicating that the behavior observed in Google can be extended to Yahoo!. The percentage of clicks on all other results (past the second) is similar to that observed in the Normal conditions.

Study 3. Discussion

One of the key results of this study is that little difference was found between eye-tracking results on Google and Yahoo!, indicating that the behavior observed in the earlier studies can be extended to Yahoo!. Behavioral trends followed similarly for both search engines, even though Google was rated as the primary search engine of all but one of the participants. These behavioral similarities combined with the similarly reported ease and satisfaction levels lead us to conclude that there are basic similarities in the fundamentals of how people experience and use Google and Yahoo!, and likely online search engines in general. This can be explained by the fact that the two search engines present a similar interface for returned results once advertisements and other peripheral information are removed from the result pages. Where the two search engines differed, namely in Google participants being more likely to look further in the list of returned results, can perhaps be explained by the fact that during the study, Google displayed more results than Yahoo! given the same screen space.

An additional key result of this study is that it served to further validate and extend findings from the earlier two studies. Although the earlier studies shared the same set of search tasks, this study intentionally included a selection of more challenging tasks. Despite the variation in search task difficulty, and the addition of the Yahoo! search engine, the importance of rank and the high tendency of participants to reformulate queries remained. For example, in all three studies, the average number of abstracts viewed (including reversed conditions) ranged from 2.5 to 3.8. These findings are consistent with related work (Nielson, 2006; Sherman 2005).

With respect to Study 2, questions remained as to whether the tendency to click on what is perceived to be top-ranked results even when the results are reversed could be explained by loyalty or trust in a favored search engine. Given that the behaviors were mimicked by participants using Yahoo! despite having reported Google as their primary search engine, we can infer that that behavior is not tied to a particular search engine, but instead a recurring pattern of viewing few results placed highly on the page and refining the query if those are not satisfactory. Note however that rank alone does not predict click behavior, and that there were increased likelihoods of selecting results farther in the search result list in the reversed conditions of Studies 2 and 3.

The extended results described above help us to understand what users expect from search engines. The generalization of the strong query reformulation behavior from Google to Yahoo! implies that, independent of search engine, users have a strong expectation that search engines should provide the most relevant result entries in the first page. If this expectation is not fulfilled, they try to adapt their input to the search engines by reformulating queries until they are immediately provided with satisfactory results matching their search, rather than spending much time on further pages returned by their original query. It is also notable that in Study 1, a large percent (40–50% task dependent) of the total search task time was spent on the search results pages, and furthermore in Study 2, a significant 43% of queries resulted in query reformulation rather than clicks. Such an effort suggests that adding query formulation suggestions to results pages, as some search engines have done as well as clearly highlighted key words on the results pages, could be advantageous.

Challenges Ahead

In our experiences conducting these studies, the need for better eye-tracking support to understand the process participants follow when interacting with search results pages became apparent. Two major challenges that demand considerable attention are the difficulty in analyzing and interpreting eye-tracking data and the difficulty in integrating eye-tracking methods with other usability testing techniques, such as think aloud and bio feedback. While these two areas are rich with open problems, we present a preliminary solution towards meeting the first of these challenges.

Visualization of Scanpaths in Online Search

Eye movement scanpaths are typically represented as sequences of areas of interests (AOIs), or *lookzones*. In our studies, each ranked result 1 through 10 represented lookzones 1 through 10 on the page. The ability to map pixel-based gaze coordinates to areas of interest comes standard with state of the art eye-tracking software.

Despite the capabilities to automatically capture and record fixations by lookzone, the analysis of the sequences themselves proves to be time consuming and challenging. For example, can one find a representative path or compare paths? Once even two or three paths are plotted on the same image, the paths become unreadable. What about outliers? How does the collection of paths sum to reveal characteristic path behavior in the same way that static fixations can?

Heat maps (Reeder, Pirolli, & Card, 2001), which are similar to shadow maps and utilize a color spectrum to reveal intensity, are one visualization technique for analyzing eye-tracking data and are currently part of Eyetools' commercial eye-tracking analysis software. A sample heatmap visualization downloaded from their Web site (Eyetools, Inc., 2007) is included in Figure 3. While heatmaps make eye-tracking analysis less cumbersome and provide a great high-level visual overview, they lack the means to reveal the scanpaths or sequences that are important to the process of search. For instance, recall that analysis of these paths in the above-

DOI: 10.1002/asi

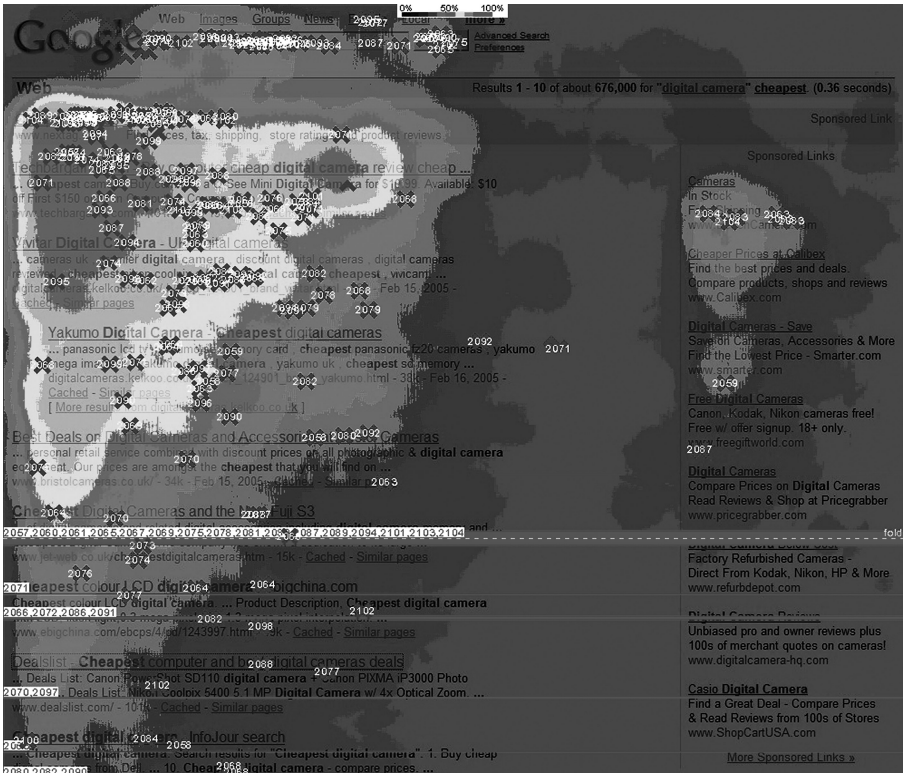

FIG. 3. Heat map illustration of a Google results page from Eyetools, Inc.

mentioned studies revealed search strategy patterns as well as gender differences which fixation data alone did not.

Work by Pellacini, Lorigo and Gay (2006), shown in part in Figure 4, sets out to help overcome this shortcoming. The goal of this work was not simply to depict scanpaths but to depict scanpaths in the context of the others paths in the data set. Hence, Figure 4 shows a scanpath together with other visual clues that relate that scanpath to the group behavior. In that figure, the inner 10 circles of varying sizes reflect the 10 ranked search results, with circle "A" representing rank 1, and moving counterclockwise in order of the next 9 ranks so that circle "J" represents rank 10. Circle size is proportional to total number of fixations on that rank, or lookzone, for all subjects. The scanpath starts with two sequential fixations (noted by two grey dots) at lookzone B, or abstract of rank two in our case. From the size of circle B, we can see that this was a popularly viewed rank, and by looking at the series of grey dots that emanate out from rank B, we can see that this participant also made a substantial number of fixations on this abstract. The path proceeds through abstract of rank three, then again through two, followed by 10 sequential fixations on abstract four (labeled "E"). The path then goes through other lookzones, including unpopular lookzones 9 and 10 (labels I and J, and concludes with a fluctuation

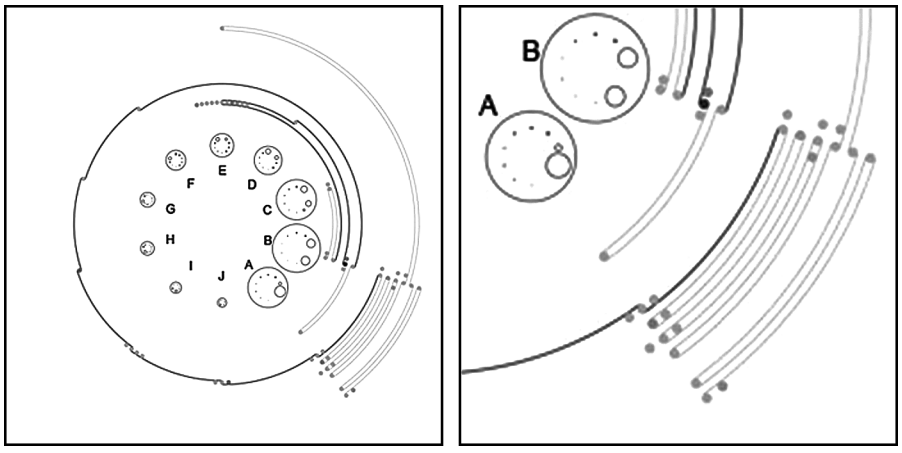

FIG. 4. Visualizing a search Path in the context of a group of paths: entire path (left), close-up of zig-zag pattern found between results ranked 1 and 2 (right). Label A represents the area of the Google results Web page containing result of rank 1. Labels B through J (counterclockwise) represent ranks 2 through 10, respectively.

DOI: 10.1002/asi

between ranks one and two. In this example, the participant viewed all the results on the page and, during that scan, spent time reading the result ranked five, as witnessed by the large number of consecutive dots, and spent time alternating between results of ranks one and two, which is shown also in the close-up image on the right. The close-up reveals a zig-zag pattern that went unnoticed until these paths were visualized.

Contextual comparisons in this tool make use of Markov models where probability likelihoods are witnessed by inner circle sizes (note, the positions of the two large inner circles at label B show that it is most likely that from B a user will proceed to either A or C). We also use Levenshtein-distance (Levenshtein, 1966), also known as *string-edit*, and depict string alignments using bold lines.

Other recent work, West, Haake, Rozanski, and Karn (2006), shares a similar goal by introducing a tool named eyePatterns, designed specifically for scanpath analysis. Both this work and our work on similar techniques to compare paths, namely, Levenshtein distance (Levenshtein, 1966) and Needleman-Wunsch (1970), a modification of Levenshtein distance that allows for additional control of the distancing parameters. For example, to compute the difference between two paths, or sequences of lookzones, we compute the number of insertions, deletions, or substitutions needed to transform one path to the next. For instance, path ABCDEF is paired with path BCDEFA in its original and aligned form below.

ORIGINAL	ALIGNED
A B C D E F	A B C D E F –
B C D E F A	– B C D E F A

These paths are in fact similar and have a Levenshtein distance of two, shown by the two differences in the aligned pairing found at the first and last character. This distance metric has also been used in other eye-tracking experiments, such as Josephson and Holmes (2002).

One difference between our approach and that by West et al. (2006) is that in our work the scanpaths themselves are portrayed visually, rather than solely textually via lookzone sequence strings. The value of this visual approach has already been revealed when looking at our data from Study 1 to test out this new tool. Even though the scanpaths had been previously analyzed with traditional and time consuming means, the actual visualization of the paths themselves quickly revealed both reading and zig-zag patterns. The zig-zag pattern, for example, was depicted in Figure 4 and was quickly found to be a trend when flipping through the paths in Study 1 with this tool. This was likely overlooked in our prior analysis because (a) line crossings that occur when a scanpath is plotted on a Web page can make it difficult to decipher path structure and (b) the scripts we wrote to analyze the scanpath sequence strings looked for specific patterns such as linearity. Previously, we did not have the intuition or foresight to look for trends that only the image later made quickly apparent. While our visual tool is limited to eye-tracking data that is studied according to lookzones, and for which there are a relatively small number of lookzones, it added value when looking at Web search behaviors.

Future Work

This article has highlighted the results, lessons, and implications of controlled experimental studies using eye tracking as a methodology tool. However, one of the challenges in better interpreting ocular indices is effectively integrating eye tracking with other methods, especially methods traditionally deployed by usability practitioners in industry work.

Thus far, eye tracking has really only been successfully combined with click through data (Granka, et al., 2004, Joachims, Granka, Pan, Hembrooke, and Gay 2005), in which a strong correlation has generally been found between the results that users look at and what they click on. It is more difficult however to integrate eye tracking with other known usability methods, particularly the think aloud protocol. Think aloud is likely to distort eye-tracking measures due to an increase in cognitive load by having users both verbally speak and silently read the information that they are presented with. Rayner (1998) reported ocular differences during silent and verbal reading, namely significant differences in fixation duration and time, and this is just one of the basic differences that can be expected when trying to interpret eye-tracking data tainted by think aloud. For this reason, usability studies relying on the accurate interpretation of eye-tracking data generally do not use think aloud in conjunction with eye tracking. However, given the value of the two approaches for evaluating user interaction and behavior, it would be interesting in future work to determine how these two may be combined.

Another area for future work would be to integrate physiological measures. Fields of psychology and physiology have standards for interpreting biometric data, such as heart rate, skin response, and others. It would be useful to see if ocular data can be collected in parallel with these measures to determine if increases in arousal or frustration can also be detected through differences in eye movements. Right now, eye tracking only offers insight into what a user might be processing cognitively and thinking; combining eye tracking with physiological measures could offer insight into how a user is feeling when processing stimuli.

As we look forward, eye tracking is not only useful for analyzing user behavior, as in the work we describe ahead, but also used as an input mechanism and a means of interacting with a program, game, or other technology. Work by Oyekoya and Stentiford (2006) has shown that eye tracking can be valuable input in the online search of images.

Furthermore, work including that of Joachims et al. (2005) and Agichtein, Brill, and Dumais (2006) has shown that implicit user behaviors can positively impact search ranking. Another area for future research would be to investigate how some of these implicit behaviors might be used to improve search ranking functions, including search along a variety of media and from a variety of devices.

DOI: 10.1002/asi

Conclusion

In this article, we described three studies we undertook that used eye tracking to study online search, a ubiquitous behavior on the Web. These studies revealed characteristic behaviors of participants using Google, and also Yahoo!. Our studies, as well as related work, confirm the strong influence that rank has on viewing behavior, with only three to five abstracts viewed on average. Task type, difficulty level, gender, search engine, and search engine familiarity were all considered in our experiments, and conditions were also manipulated to reverse the perceived ranking of the abstracts. Despite these variations, rank continued to significantly influence behaviors. Our newest study suggests strong similarities between behaviors on Google and Yahoo!. In fact, no significant differences were found between success rates for the query tasks, time spent, number of fixations, viewing and click behavior, or user-reported satisfaction and ease.

We have adopted eye tracking to study online search because of the evaluative and cognitive nature of this process and its ability to explain and enrich click-through data. The search within the search results is illuminated by the path of fixations that participants make on the page of ranked abstracts. Analyzing these paths remains cumbersome in practice, and we also presented related work on a visualization designed to help decipher and understand paths from large data sets.

Acknowledgements

We wish to thank and acknowledge Matt Feusner for his assistance in data preparation and insights for the earlier two studies and discussions. We are also thankful for Google, Inc. who provided partial funding for parts of this research.

References

Agichtein, E., Brill, E., & Dumais, S.T. (2006). Improving Web search ranking by incorporating user behavior. In Proceedings of SIGIR 2006 (pp. 19–26). New York: ACM Press.

Applied Science Laboratories (2005). ASL 504 eye tracker manual. Bedford, MA: Applied Science Group, Inc.

Bilal, D., & Kirby, J. (2002). Differences and similarities in information seeking: Children and adults as Web users. Information Processing and Management, 38(5), 694–670.

Broder, A. (2002). A taxonomy of Web search. SIGIR Forum, 36(2), 3–10.

Byrne, M. D., Anderson, J. A., Douglass, S., & Matessa, M. (1999). Eye tracking the visual search of click-down menus. In Human Factors in Computing Systems: CHI 99 Conference Proceedings (pp 402–409). New York: ACM Press.

Duchowski, A.T. (2002). Abreadth-first survey of eye tracking applications, behavior research methods, instruments, & computers, 34(4), 455–470.

Eyetools, Inc. (2007). Retrieved July 9, 2007 from http://www.eyetools.com/inpage/research_google_eyetracking_heatmap.htm

Google Inc. (2005). Google Zeitgeist: Search patterns, trends, and surprises. http://www.google.com/press/zeitgeist.html

Granka, L., & Rodden, K. (2006) Incorporating eyetracking into user studies at Google. Workshop paper presented at CHI 2006, ACM.

Granka, L., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in www search. Poster Session presented at the Conference on Research and Development in Information Retrieval (SIGIR).

Guan, Z. & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In Conference on Human Factors in Computing Systems (pp.). New York: ACM Press.

Hembrooke, H.A., Granka, L.A., Gay, G.K., & Liddy, E.D. (2005). The effects of expertise and feedback on search term selection and subsequent learning. Journal of the American Society for Information Science and Technology, 56(8), 861–871.

Hess, E., & Polt, J. 1960. Pupil size as related to interest value of visual stimuli. Science, 132, 3423, 349–350.

Hornof, A. (2004). Cognitive strategies for the visual search of hierarchical computer displays. Human-Computer Interaction, 19(3), 183–223.

Hsieh-Yee, I. (2001). Research on Web search behavior. Library and Information Science Research, 23(2) 167–185.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science and Technology, 44(3), 161–174.

Jacob, R.J., & Karn, S.K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. H. Radach & H. Deubel (Eds.), In the mind's eye: Cognitive and applied aspects of eye movement research (pp. 573–605). Amsterdam: Elsevier Science.

Jansen, B.J., & Pooch, U. (2000). Web user studies: A review and framework for future work. Journal of the American Society of Information Science and Technology, 52(3), 235–246.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: Study and analysis of user queries on the Web. Information Processing and Management, 36(2), 207–227.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting click-through data as implicit feedback. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 154–161). New York: ACM Press.

Josephson, S., & Holmes, M. (2002). Visual attention to repeated internet images: Testing the scanpath theory on the World Wide Web. Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (pp. 43–49). New York: ACM Press.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. Doklady Physics, 10, 707–710.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google. Information Processing and Management, 42(4), 1123–1131.

Macgregor, J.N., Lee, E.S., & Lam, N. (1986). Optimizing the structure of menu indexes: A decision model of menu search. Human Factors, 28, 387–400.

Needleman, S., & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. Journal of Molecular Biology, 48(3), 443–453.

Nielsen, J. (2006). F-Shaped pattern for reading Web content. Jacob Nielsen's Alertbox for April 17. Retrieved April 17, 2006 from http://www.useit.com/alertbox/reading_pattern.html

Oyekoya, O.K., & Stentiford, F.W.M. (2006). An eye tracking interface for image search. Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, p. 40. New York: ACM Press.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position and relevancy. Journal of Computer-Mediated Communication, 12(3), article 3.

Pellacini, F., Lorigo, L., & Gay, G. (2006). Visualizing Paths in Context, Technical Report #TR2006-580, Dept. of Computer Science, Dartmouth College.

Rayner, K. (1998). Eye movements and information processing: 20 years of research Psychological Bulletin, 124(3), 372–422.

Reeder, R.W., Pirolli, P., & Card, S. (2001). Webeyemapper and Weblogger: Tools for analyzing eye tracking data collected in Web-use studies. Conference on Human Factors in Computing Systems (pp. 19–20). New York: ACM Press.

Sherman, C. (2005). A new F-word for Google search results. Search Engine Watch. Retrieved March 8, 2005 from http://searchenginewatch.com/showPage.html?page=3488076

DOI: 10.1002/asi

Spink, A., Bateman, J., & Jansen, B.J. (1999). Searching the Web: A survey of excite users. Internet research, electronic networking applications and policy, 9(2), 117–128.

Teevan, J., Alvarado, C., Ackerman, M., & Karger, D. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems (pp. 415–422). New York: ACM Press

Tillotson, J., Cherry, J.M., & Clinton, M. (1995). Internet use through the University of Toronto library: Demographics, destination and users' reaction. Information Technology and Libraries, 14(3), 190–198.

West, J.M., Haake, A.R., Rozanski, E.P., & Karn, K.S. (2006). eyePatterns: Software for identifying patterns and similarities across fixation sequences. Proceedings of the 2006 Symposium on Eye Tracking Research & Applications (pp. 149–154). New York: ACM Press

Appendix A

Query tasks used in Studies 1 and 2.

1. Find the homepage of Michael Jordan, the statistician.
2. Find the page displaying the route map for Greyhound buses.
3. Find the homepage of the 1000 Acres Dude Ranch.
4. Find the homepage for graduate housing at Carnegie Mellon University
5. Find the homepage of Emeril - the chef who has a television cooking program.
6. Where is the tallest mountain in New York located?
7. With the heavy coverage of the democratic presidential primaries, you are excited to cast your vote for a candidate. When is/was democratic presidential primaries in New York?
8. Which actor starred as the main character in the original *Time Machine* movie?
9. A friend told you that Mr. Cornell used to live close to campus - near University and Stewart Ave. Does anybody live in his house now? If so, who?
10. What is the name of the researcher who discovered the first modern antibiotic?

Appendix B

Query tasks used in Study 3.

1. What are the rules of Australian rugby?
2. Find information on the Icelandic personal identification number system.
3. Find student housing for transgendered students in Ithaca.
4. Find 3 common ergonomic risk factors in computing.
5. Find a substitute for cornstarch in recipes.
6. How do hurricanes get their name?
7. What services are offered by the Calidris company?
8. Find information on herbal treatment of hypothyroidism.
9. Who directed *Powers of Ten?*
10. What are the main muscle groups worked in a spinning class?

DOI: 10.1002/asi