In Google We Trust: Users' Decisions on Rank, Position, and Relevance

Bing Pan

School of Business and Economics
College of Charleston

Helene Hembrooke

Human-Computer Interaction Group
Information Science Program
Cornell University

Thorsten Joachims

Department of Computer Science
Cornell University

Lori Lorigo
Geri Gay
Laura Granka

Human-Computer Interaction Group
Information Science Program
Cornell University

An eye tracking experiment revealed that college student users have substantial trust in Google's ability to rank results by their true relevance to the query. When the participants selected a link to follow from Google's result pages, their decisions were strongly biased towards links higher in position even if the abstracts themselves were less relevant. While the participants reacted to artificially reduced retrieval quality by greater scrutiny, they failed to achieve the same success rate. This demonstrated trust in Google has implications for the search engine's tremendous potential influence on culture, society, and user traffic on the Web.

doi:10.1111/j.1083-6101.2007.00351.x

Introduction

Finding online information using search engines has become a part of our everyday lives (Gordon & Pathak, 1999). Currently the search engine serving the largest percentage of queries (at 47.3%) is Google, with an index of around 25 billion Web pages and 250 million queries a day (Brooks, 2004; Search Engine Watch,

2007). Google now provides search functions on handheld devices and smart phones (Google Inc., 2005a). With the ubiquitous presence of mobile devices, anytime and anywhere access to the information world has become a reality. Other popular search engines include Yahoo, MSN, AOL, and Ask.com, and all serve the pervasive need of finding pertinent information within the enormity of the Web. Despite the popularity of search engines, most users are not aware of how they work and know little about the implications of their algorithms (Gerhart, 2004).

All of the search engines noted above respond to a query with a ranked list of 10 abstracts in their default setting. The ranking reflects the search engines' estimated relevance of Web pages to the query. Individual search engines vary by both underlying ranking implementation, and also by various characteristics of how they display the ranked results, and any additional support they provide for finding related Web pages. Users can evaluate the abstracts, or other information displayed about a given result, before deciding whether to visit any of the suggested pages by clicking on a hyperlink. In this study, we chose to use Google because of the frequency of its use, the simplicity of its display of query results (which can serve as a common basis when compared to many other search engines), and our prior experience studying Web search on Google (Granka, Joachims, & Gay, 2004). We also confined the study to only one search engine to ensure a constant visual display on which we could analyze and interpret the subjects' eye movements.

The information search process is made possible through three parties: Web authors, the search engines themselves, and the users of search engines. The Web authors put their Web pages online with appropriate linking to other pages. The link structure has been used by popular search engine algorithms (Brin & Page, 1998) that can take advantage of this structure to rank relevant Web pages. Users of search engines enter various keywords (sometimes with Boolean commands) according to their understanding of the task and the functionality of the search engine, and they evaluate the results returned by the search engine, making a decision on whether or not to select one of the returned results or reformulate the query. Search engines act as an information intermediary that facilitates the information seeking process.

However, how well a Web page actually reflects the users' search intentions is hard to measure. For example, the ranking algorithm of Google uses a page's measure of in-links to help inform its quality and relevance (Pandey, Roy, Olston, Cho, & Chakrabarti, 2005). Some have argued that those algorithms, such as PageRank, simply set up a rich-get-richer loop, whereby a relatively few sites dominate the top ranks (Hindman, Tsiotsioliklis, & Johnson, 2003). Retrievability and visibility represent only part of the search process. We wondered what role the user plays in perpetuating this rich-get-richer dynamic. Particularly, we wondered how much of the correlation between Web traffic and site popularity (Hindman et al., 2003) is due to the alleged efficiency of these algorithms as opposed to users' tendency to simply trust the ranked output displayed by a search engine and forego any in-depth analysis or comparisons of the retrieved results. More importantly, Google's imperfect algorithm is open to abuses such as Google bombing (Tatum, 2005; see also Bar-Ilan, this

issue). This might deliver erroneous messages to a large population when the searchers trust Google without questioning its underlying ranking mechanism.

Our curiosity regarding this question was piqued by an earlier study conducted by Granka et al. (2004). Their results indicated that most student subjects only view and click the top two results returned by Google. The design of this earlier study did not tease apart whether those choices were the result of the top positions of the two abstracts as influenced by Google's ranking algorithm, or if those were truly the most relevant results as evaluated by the subjects. We were interested in finding out whether a user's choice of a particular abstract was based on the position of that abstract, the user's evaluation of the relevance of that abstract, or a combination of the two.

In order to explore the relative contribution of relevance and position, we employed eye tracking as the methodology in the current study. Eye tracking devices are able to record eye movements and reveal subjects' attention and cognitive process. In areas of cognitive psychology, human-computer interaction, and marketing, eye tracking methods have been used for decades (Rayner, 1998). We use eye tracking to investigate how users make decisions when confronted with returned Google results following a query. Eye tracking adds meaning to the more traditional log file or click behavior analysis. It allows for a more complete assessment of the information-seeking process by revealing which query result abstracts users looked at, or were aware of, before selecting a query result or refining their query. This article provides behavioral evidence that sheds light on the influential factors in the evaluation process of search engine uses.

Related Research

The hypertext nature of the Web has changed how people search and access information (Bilal & Kirby, 2002). It is imperative to understand user behavior on the Web in order to design better search engines. This section describes past research on user behavior and search engines. The relevant literature is organized into three parts: past research on user behavior and search engines, past eye tracking research related to Web viewing behavior, and a review of the major results of the first Google eye tracking study the authors conducted, which provided the basis for the current eye tracking study.

User Behavior and Search Engines

Most of what is known about user behavior and search engines comes from Web log analysis and clickthrough analysis. These studies are descriptive in nature and report general user behavior (Bilal & Kirby, 2002). For example, most users input two or three terms in search engines, they rarely go to the second page of results, and most of the time they only click on one document in the result set (Jansen, Spink, Bateman, & Saracevic, 2000). Furthermore, different user groups use search engines differently. For example, children and adults have different information search

strategies and different success rates (Bilal & Kirby, 2002), and search engine users in the U.S. and Europe engage in different search behaviors (Jansen & Spink, 2004). Spink and Ozmultu (2002) discuss the characteristics of searching using question-type queries and identify four distinct types. Other researchers show the differences of search behavior over the course of time (Jansen, Spink, & Pedersen, 2005; Ozmutlu, Spink, & Ozmutlu, 2004). Hsieh-Yee's (2001) review of the literature summarizes research on user behavior focusing on search patterns and many studies that have investigated the effects of selected factors on search behavior, including information organization and presentation, type of search task (Lorigo et al., 2005), Web experience, cognitive abilities, and affective states.

Eye Tracking and the Web

Studies of eye movements date back to work by Javal in 1879 (Huey, 1908), and over time they have informed fundamental facts about eye movements, behavioral and experimental psychology, and human-computer interaction, an application that is greatly benefiting from advances in the ease and accuracy of eye trackers. In a typical user study, the subject is calibrated to the eye tracking device by the researcher asking the subject to look at specific targets while the software configures the respective target locations. This is done by sending a weak infrared light to the subject's eye and measuring the light reflection on the screen. The result is that eye movements on a computer screen can be recorded, with a high degree of accuracy, for the majority of people.

Ocular behavior on Web pages has been the focus of several recent studies, including eye movement research on news websites (Stanford-Poynter Project, 2000), analysis of scanpaths (sequences of eye gazes) on Web pages (Josephson & Holmes, 2002), ocular behavior of Web users completing tasks on a Web portal page (Goldberg, Stimson, Lewenstein, Scott, & Wichansky, 2002), and eye movement behavior on several types of websites (Pan et al., 2004). The Stanford-Poynter Project (1998) investigated reading behavior on news websites and concluded that text was frequently the first entry point for a majority of online news readers. Rayner, Rottello, Stewart, Keir, and Duffy (2001) reported similar findings in which viewers of print advertisements spent more time on text than pictures. Josephson and Holmes (2002) studied the eye viewing behavior on three different Web pages. They concluded that users develop habitually preferred scanpaths and that features of websites and memory might be important in determining those scanpaths.

Goldberg et al. (2002) used eye tracking methods to test the performance of subjects in completing several tasks on a Web portal page. Their research demonstrated the characteristics of subjects' eye movements on that portal page. This research gave rise to implications for improving the design of the Web portal. Pan et al. (2004) showed that gender, website types, and the interaction between search sequence and website type all affect Web viewing behavior. For example, female subjects had shorter mean fixation durations than males; the subjects had longer mean fixation durations on the first Web pages viewed than the second ones; and the subjects spent more time gazing on the first pages than on the second ones. In

general, as an indicator of information processing, eye movements on Web pages are influenced by both individual variables, such as gender, and the characteristics of the stimuli, such as the layout and content of Web pages. The current study combines eye tracking with clickstream data in order to make inferences regarding the impact of positions versus judged relevance on decision-making processes involved in information search. By doing so, we gain an in-depth understanding of how users evaluate search results and the factors that influence their choices.

Previous Eye Tracking Search Research

In general, knowing how users evaluate result pages through eye tracking methods can help researchers to understand users' motivations, tasks, and cognitive processes. Understanding this evaluation and decision-making will enable correct interpretation of Web log files as feedback data and thereby improve search engine performance (Joachims, 2002). As a result, it may be possible to design Web-based information retrieval systems to better satisfy user needs.

In the study conducted by Granka et al. (2004), 10 tasks were devised, ranging from informational tasks such as "Who discovered the first antibiotics?" to navigational tasks such as "Find the homepage of Emeril - the chef who has a TV cooking program." Informational tasks require finding a particular fact, while navigational tasks involve searching for a particular Web page (Broder, 2002). Among the 10 tasks, some were inspired by popular topics from Google Zeitgeist, while others covered local or specialized topics. Google Zeitgeist (Google Inc., 2005b) is a report provided by Google that reveals the most popular queries or other related trends about the queries Google receives. Our resulting mix of popular and specialized topics was an effort to simulate a likely query task situation.

In Granka et al. (2004), the subjects were given the 10 tasks in random order and asked to start with Google and search for answers with a limit of three minutes for each task. The time constraint allowed for the collection of a substantial amount of eye tracking data for each task and also minimized the total time required of each subject. Most subjects voluntarily stopped the search tasks in the given amount of time. In this earlier study, complete eye tracking data were obtained for 23 subjects. The results showed that the subjects viewed the top two abstracts almost equally often and much more often than any other abstract. However, they clicked the number one ranked abstracts significantly more often than the number two ranked abstract (Granka et al., 2004). This behavior prompted us to ask whether the subjects were simply defaulting to Google's ranked results, or if their decisions were based on some critical evaluation of the results.

Research Methods and Design

In the present study, rank refers to the original sequence of abstracts returned by Google. Lower ranks indicate that the Web pages are less relevant as judged by Google's algorithm and thus placed later in the sequence; position represents the

actual physical locations of the abstracts on the Google results page, for example, from top to bottom (1 to 10) on the first Google result page; relevance represents the subjective judgments of the likelihood that the information piece is related to the answer of the question or the goal of a search task. In this study, we obtain relevance through human judgments of the abstracts returned by Google (abstract relevance), as well as the pages associated with those abstracts (Web page relevance). Judgment data were important to ensure that all of the 10 results were not equally relevant.

Based on the findings from our previous study (Granka et al., 2004), the current work was designed to exploit Google's ranking function in order to investigate how much the subjects rely on Google's ranking to make their decisions about relevance. Unbeknown to the subjects, we manipulated the order of Google's returned results in some cases, such that abstracts of actual lower ranked Web pages appeared higher in position and vice versa. Thus, choosing a lower ranked abstract that is in a higher position in the Google results page but is evaluated to be less relevant by human judges would be evidence that the subjects have assigned priority to Google's "expertise" over the actual relevance of the abstract.

This section introduces eye tracking as a methodology and introduces the details of the research methods and procedures used in the current study. A laboratory setting was necessary to capture all aspects of the search sessions and related eye movements, in order to compare the variables across all subjects systematically. Although some scholars have argued that external validity is compromised in a laboratory setting, previous studies have shown that in laboratory settings and Web settings there are few or no differences in the subjects' behavior on information search, especially on those tasks using keywords (Epstein, Klinkenberg, Wiley, & McKinley, 2001; Schulte-Mecklenbeck & Huber, 2003).

The Subjects

In this study, participants were undergraduate students with various majors (including communication, engineering, and arts and sciences) at Cornell University (U.S.A.). All students were given extra class credit for their participation in the experiment. Twenty-two subjects were recruited, and 16 complete data sets, including 11 males and 5 females, were obtained. Attrition was due to random recording difficulties and the inability of some subjects to be calibrated precisely.[2] The average age of participating subjects was 20 years and 4 months. All subjects reported that they used Google as their primary search engine and had a high familiarity with the Google interface (all scored 10 out of 10); when asked about the levels of trust in Google, they reported an average of 7.9 (out of 10). Thus, our subjects, in general, are savvy users of Google and tend to trust Google to a high degree.

Search Tasks

Ten search tasks were included in this study, each of which addressed a unique aspect of the information retrieval experience. Half of the searches were navigational in nature, asking subjects to find a specific Web page or homepage. These were

definitive searches, meaning that only one correct Web page would provide an acceptable answer. The other five tasks were informational, asking subjects to find a specific bit of information (Broder, 2002). Much of the content for the tasks was generated according to the content of top searches listed on Google Zeitgeist (Google Inc., 2005b). Our purpose was to ensure that the tasks in this experiment represented the various genres of searches that the general population uses on a regular basis, including travel, movies, current events, celebrities, and local issues. These tasks were also pre-tested to ensure that the most intuitive queries would not always result in top-ranked results; therefore, the findings should be interpreted in light of the fact that these queries are on average more difficult than a subject's typical query. The following table is a brief description of the 10 search tasks included in the experiment and the correct answers to these tasks (Table 1).

Experimental Procedure

All participants were required to give informed written consent prior to the start of the experiment. Before the actual experiment, the eye tracker was calibrated using a nine-point standard calibration procedure for each subject (Duchowski, 2003). Participants were instructed to search for the 10 different tasks through the Google interface. Subjects were told to view the Web pages and search as they typically would under normal conditions, with the opportunity to scroll up and down the page at their leisure.[3] The experimenter sat to the right of and behind the subject, where she was able to watch the subject, the subject's eye, and also the corresponding eye movements on the two control monitors. If the experimenter recognized that the eye tracking system temporarily lost a subject's eye path due to extreme movements, she could re-center and if appropriate, perform a quick recalibration fix. This happened rarely and randomly; it did not interrupt the experimental session since the experimenter could perform the quick fix in a few milliseconds.

The 10 search tasks were read aloud to the subject by the experimenter to eliminate unnecessary eye movements away from the computer monitor; such eye movements could potentially hinder the accuracy of the ocular calibration. Typically, due to the monitor size, scrolling was required to view abstracts ranked seven and higher on the Google results pages. To eliminate the potential bias from question order effects, all search questions were completely randomized for all subjects. The maximum time for completing each task was restricted to three minutes. As before, the time constraint allowed for a sufficient amount of eye tracking data to be collected for each task and also minimized the total time required of each subject. The most important data for this study come from how each subject responds and interacts with the 10 results following each query and not from the full completion of the task itself.

Design

Because Google is continuously updating its search algorithms, one specific query will not produce the same exact results on two separate occasions. Because much of the data analyses were to occur after the experimental sessions, it was necessary to

Table 1 The 10 information search tasks

Task Type	Task	Correct Answer
Navigational	Find the homepage of Michael Jordan, the statistician.	http://www.cs.berkeley.edu/~jordan/
	Find the page displaying the route map for Greyhound buses.	http://www.greyhound.com/maps/
	Find the homepage of the 1000 Acres Dude Ranch.	http://www.1000acres.com/
	Find the homepage for graduate housing at Carnegie Mellon University	http://www.housing.cmu.edu/graduatehousing/
	Find the homepage of Emeril—the chef who has a television cooking program.	http://www.emerils.com/emerilshome.html
Informational	Where is the tallest mountain in New York located?	The Adirondacks OR High Peaks Region
	With the heavy coverage of the democratic presidential primaries, you are excited to cast your vote for a candidate.	March 2, 2004
	When are/were democratic presidential primaries in New York?	
	Which actor starred as the main character in the original Time Machine movie?	Rod Taylor
	A friend told you that Mr. Cornell used to live close to campus—near University and Steward Ave. Does anybody live in his house now? If so, who?	Members of Llenroc, the Cornell chapter of the Delta Phi Fraternity live in the mansion.
	What is the name of the researcher who discovered the first modern antibiotic?	Alexander Fleming

cache the Web pages with which the subjects actually interacted. A proxy server was set up to mediate the interaction between the subjects and the Google Web server. The proxy script was run on the subject's computer and stored every search query typed by the subjects, as well as all links and Web pages that were viewed, along with the corresponding times that they were accessed and viewed. When the subject typed in a query, the query was sent to the proxy server, and the proxy server relayed it to Google. After receiving the results from Google, the proxy server manipulated the results and passed on the modified results to the subject's Web browser.

Results were modified in two ways. First, the proxy server removed the advertisements on the Google results page to avoid distraction and ensure consistent stimulus exposure across all subjects. This also saved the authors from having to filter out eye movements on ads, since the goal of the study was concerned with which query results were viewed and selected. Second, in order to explore the relative contribution of relevance versus position to the decision making process, the results were further manipulated for each subject in one of three ways. In the "Normal" condition, the proxy server returned the results in their original ranked order; in the "Swapped" condition, the proxy server swapped the positions of the first ranked abstract with the second ranked abstract, keeping the rest of the ranking intact; and in the "Reversed" condition, the proxy server reversed the positions of the abstracts on the first result page as follows: The first ranked abstract was swapped with rank 10 abstract, the rank 2 abstract was swapped with rank 9 abstract, and so on.

Eye Tracking Indices

During an eye tracking experiment, several measurements are typically recorded that are relevant for studying college students' interactions with search engines. 'Fixation' refers to a relatively stable eye-in-head position within some threshold of dispersion (typically ~2°) over some minimum duration and with a velocity below some threshold (typically 15–100 degrees per second). In this study, we set the minimum duration as 50 milliseconds, as suggested in the ASL504 eye tracker manual (Applied Science Laboratories, 2005). Eye fixations are the most relevant metric for evaluating information processing in online search. Fixations represent the instances in which most information acquisition and processing occurs (Rayner, 1998). The total number of fixations is often used as an indicator of processing difficulty, with fixation density related to the complexity and informativeness of the visual stimulus (DeGraef, De Troy, & d'Ydewalle, 1992; Friedman, 1979; Henderson, Weeks, & Hollingsworth, 1999), such that as informativeness increases, so too does the number of fixations in that area. In the current study, we also used measures of the average number of fixations. A higher number of fixations on an abstract will represent intensified information processing.

'Pupil Dilation' refers to widening of the pupil. It has long been known that pupils dilate in response to emotion-evoking stimuli (Beatty, 1982). While it is also the case that pupil size is affected by light, the lighting remained constant in our experiment. As Rayner and others have pointed out (Rayner, 1998), using only a single indicator of processing difficulty may result in an oversimplification of the relationship between the indicator and processing difficulty. Hence, both fixation and pupil dilation measures are frequently used as corroborating measures of cognitive workload (Hess, 1965; Just & Carpenter, 1980; Kahneman, 1973). Last, a 'scanpath' is the spatial arrangement of a sequence of fixations, or simply the sequence of LookZones that a subject views, as in the present study.

Definition of LookZones

In addition to logging the clickstream and Web page data of subjects, the script also constructed 'LookZones' around key content regions. The script utilized a feature inherent to the GazeTracker software system that automatically creates LookZones around links and pictures, which the software recognizes within the HTML tags. (For more information on the GazeTracker software system and the eye tracking apparatus itself, see Appendix A.) Thus, the script enabled the creation of distinct LookZone regions around each of the ten displayed results (Figure 1). For the analysis, each of these displayed results on Google—abstracts in rank #1, rank #2, rank #3, to rank #10—is given its own set of LookZones, from which we can then compare eye tracking behaviors across all queries, relative to these zones. LookZones were not visible during the time participants were engaged in the experiment.

Judged Relevance

As stated above, we considered rank, position, and judged relevance in this study. For all queries and results pages that were encountered in the study, we gathered

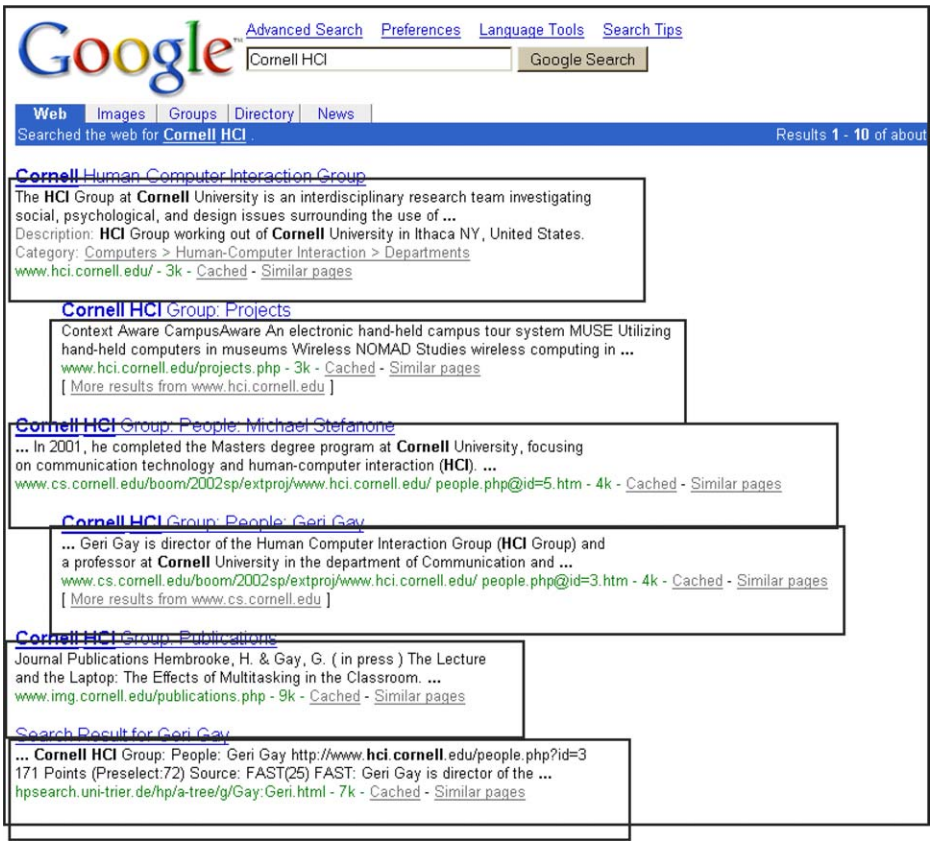

Figure 1 LookZone division on a Google result page

relevance assessments of the abstracts, which allowed us to look at the choices made by subjects as a function of the positions and judged relevance of the page chosen by the subject in case Google's rank did not reflect what other *humans* might consider relevant. Five non-participants were chosen as the judges in the study. For each results page, we randomized the order of the abstracts and asked judges to weakly order the abstracts (ties were allowed) by how promising they looked for leading to relevant. Each of five judges assessed all results pages for two questions, plus 10 results pages from two other questions, for inter-judge agreement verification. The set of abstracts/pages we asked judges to weakly order were not limited to the (typically 10) hits from the first results page, rather the set included all results encountered by a particular subject for a particular question. The inter-judge agreement on the abstracts was 82.5%. Furthermore, we also collected relevance judgments for the actual Web pages those abstracts represent. The inter-judge agreement on the relevance assessment of the pages was 86.4%.

Hypotheses and Analysis

We analyzed a combination of mouse click behaviors and the various ocular indices on three levels. On the *task level*, the analysis aggregates the behavior of a subject over all queries corresponding to one task; on the *page level*, the choice of clicks and ocular behavior on all pages were analyzed; on the *abstract level*, we analyzed the determinants of whether an abstract was viewed or clicked.

Our analyses were guided by the following hypotheses:

H1: At the page level of analysis, ocular data would differ among the three conditions.

Particularly, we expected that participants in the Reversed condition would demonstrate greater scrutiny of returned results when compared with controls, and this would manifest itself through longer fixation times overall, greater overall total fixations, how many abstracts were looked at per page, and visual backtracking or regressing to earlier viewed abstracts. In addition:

H2: At the abstract level of analysis, the eye data from participants in the Reversed condition would indicate explicit trust for Google's ranking, as evidenced by a lack of significant difference among the three conditions in the number of fixations per abstract on the top two positioned abstracts. Furthermore, subjects would look at the last two positioned abstracts (the number one and two Google ranked abstracts) more than in the other two conditions, indicating an implicit awareness of their significance, either from confusion or interest. This implicit awareness would also be demonstrated in subjects' pupil dilation data.

However, we suspect that the subjects tend to trust Google as an authoritative search engine since users mostly clicked on the first returned result in the first eye tracking study (Granka et al., 2004) They may still click on the abstracts higher on top even though the real relevance of those is low. Thus:

H3: Participants in both the Swapped and Reversed conditions would still choose abstracts of actual lower rank more often than subjects in the control condition (those who viewed Google results in their actual ranked order).

Thus, we anticipated dissociation between the ocular data, which would indicate some implicit conflict between the position and the actual Google rank, yet that the subjects would still choose a higher positioned abstract based on a greater trust in Google's algorithms than in their own judgment. This can be validated through regression analysis of whether or not an abstract was clicked on, the relevance, and the position of all abstracts.

Research Results

Task Level Analysis

On the task level, the success rates were significantly different among the three conditions. In the Reversed condition, subjects completed search tasks successfully only 62% of the time, compared with 85% and 80% in the Normal and Swapped conditions respectively ($p < .05$). On average, 43.1% of queries resulted in reformulation of the query right away without clicking on any results. The rates of query reformulation without clicking were not significantly different among the three conditions.

Page Level Analysis

The researchers obtained eye tracking data from Gazetracker on all of the Web pages the subjects viewed, along with the cache of those pages. Since our focus is on the evaluation process of Google results pages, we specifically selected those Web pages that are the first result pages returned by Google (containing abstracts 1–10) and on which the subject clicked on at least one abstract, so that the evaluation process can be obtained. The results show that subjects in the Reversed condition spent more time checking each page (10.9 seconds vs. 6.5 seconds in the Normal condition and 5.8 seconds in the Swapped condition, $p<.05$), made more fixations (30.0 fixations vs. 18.3 in the Normal condition and 17.9 in the Swapped condition, $p<.05$), and checked more returned results (3.8 abstracts vs. 2.5 in the Normal condition and 2.7 in the Swapped condition, $p<.05$).

In addition, the viewing sequences of the abstracts were calculated based on the subjects' scanpaths. A regression was defined as an instance in which a subject returns to a previously viewed abstract (see Figure 2 for an example). The number of regressions for subjects in the Reversed condition is significantly higher than those in the Normal and Swapped conditions (3.4, 2.2, and 1.9 in Reversed, Normal, and Swapped conditions respectively, $p<.05$). In all these results, there are no significant differences between Normal and Swapped conditions.[4] All of these indices suggest that less optimal abstracts resulted in more scrutiny. Furthermore, after the experiment session was completed, subjects in the Swapped and Reverse condition were explicitly asked if they had any feedback, followed up by a question regarding whether everything seemed normal with the Google interface, in order to assess whether they suspected the ruse. Comments were made such as, "I didn't have much

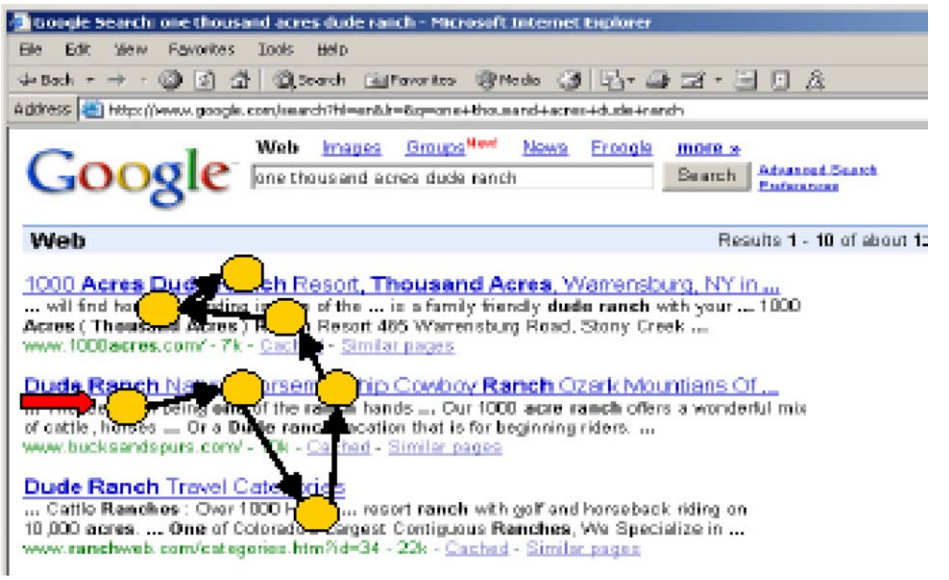

Figure 2 An example of regression on a Google page

luck with several of the questions," and "I just could not think of the right search terms," which reflected a tendency for subjects to blame themselves rather than the inferior Google results.

Abstract Level Analysis

Abstract level analysis was performed using two methods. First, on a global level, we made comparisons between views and clicks on the 10 abstracts across three conditions, followed by regression analysis of the determinants of viewing and clicking behavior. Next, we compared the three groups on a more granular level, looking at the number of fixations and pupil dilation on a per abstract basis. In this way we were able to infer differences among the groups in how they might be processing the displayed rank of Google's output.

Corresponding to the three conditions, Figure 3 illustrates which abstracts the subjects viewed before the first click and where the first click was placed. In particular, the bars in the graphs show the percentage of result pages with at least one fixation or click at the respective rank. In the Normal condition, the result is similar to the study by Granka et al. (2004): The subjects viewed the two top-ranked abstracts with the highest frequency and clicked on the number one abstract most of the time. In the Swapped condition, the subjects also viewed the two top-ranked abstracts almost equally. However, they still clicked on what they perceived to be the number one ranked abstract almost three times more often than the abstract truly ranked highest by Google (now in the number two position). It appears that respondents were heavily influenced by the position of the abstract. In the Reversed condition, the percentage of clicks on the abstracts positioned at rank one and two

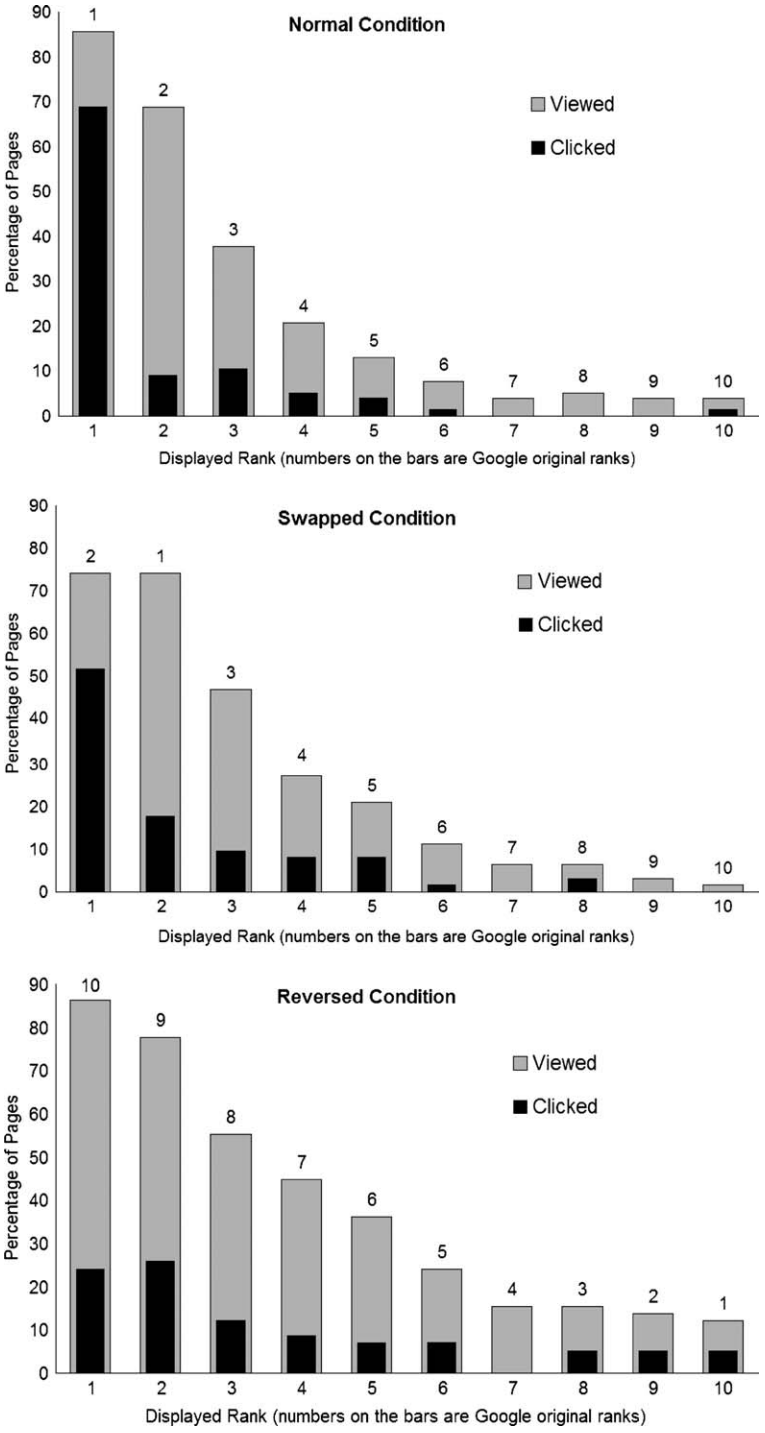

Figure 3 Percentage of views/clicks on different ranks under three conditions

was significantly less when compared with the other two conditions. This indicates some degree of explicit evaluation of Google results. However, the graph is far from being a "mirror image" of the Normal condition. A one-sample T-test indicated that, similar to the Normal condition, subjects chose the top five presented abstracts significantly more often than the bottom five abstracts (df=57, $p<.01$). While subjects in the Reversed condition were somewhat more likely to view the lower positioned abstracts, this still happened fairly infrequently.

We also compared the groups using fixation and dilation data to see how they differed as subjects viewed the top positioned abstracts versus the 9th and 10th abstracts. When we compare this to what abstract(s) they ultimately chose, we are able to gain further insight regarding the balance between trust and users' evaluations of relevance.

The first analysis we conducted compared the average number of fixations on the first and second positioned abstracts. The result of this analysis indicated that there were no significant differences among the three groups in fixation density between the top two abstracts. Thus, even among participants in the Reversed condition, who were actually viewing the 9th and 10th ranked abstracts in these positions, the abstracts were evaluated with equal attention. It appears, then, that these subjects contemplated these abstracts much as they would have under normal conditions.

Next, we made comparisons among the groups on the 9th and 10th positioned abstracts. Here we found significant differences among the groups, with subjects in the Reversed condition making more fixations on both of these abstracts than in either of the other two conditions ($F(1, 2) = 3.53$, $p<.03$ and $F(1, 2)=3.83$, $p<.02$, respectively, for the 9th and 10th abstracts). When we compared the three groups again on the average number of fixations for the number one ranked abstract (positioned number one in the Normal condition, number two in the Swapped condition, and number 10 in the Reversed condition), there were significant differences again. Participants in the Normal condition made more fixations on the number one positioned abstract than the subjects in the Swapped condition made on the number two positioned abstract. Furthermore, fixations in the Swapped condition were significantly greater than in the Reversed Condition in response to the 10th positioned abstract. This indicates that participants in both the Swapped and Reversed conditions were "fooled" by the manipulated results.

Finally, we looked at differences in pupil dilation for the subjects in the Reversed condition, specifically focusing on how the dilation measure did or did not differ between the first two positioned abstracts and the 9th and 10th positioned abstracts. No significant differences in pupil dilation between the 1st and the 10th positioned abstracts or the 2nd and the 9th positioned abstracts were noted. Again, this indicates that attention to and interest in the least relevant but positionally first two abstracts was equivalent to the most relevant but positionally last two abstracts (Figure 4).

How strong is the influence of position compared to the intrinsic relevance of an abstract? To address this, we regressed viewing and clicking behavior on position and judged relevance of the abstracts. We first used a mixed model analysis to explore the significant determinants of viewing and clicking. The results show that significant

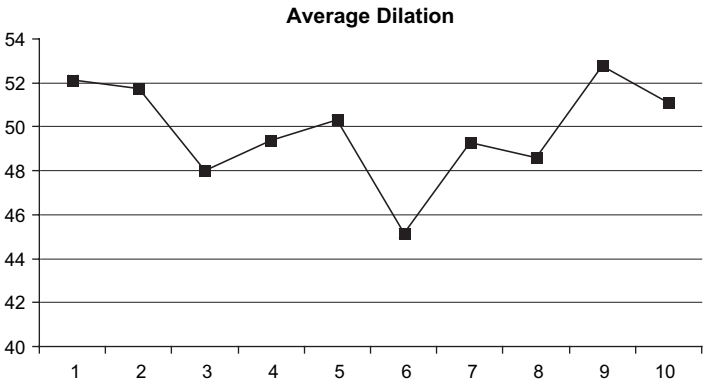

Figure 4 Average pupil dilation across abstracts for reversed condition

determinants of whether an abstract is viewed are the order of significance, position, relevance, condition, and query order (Table 2). Similarly, the order of significance, position, relevance, and task type (informational task or navigational task) significantly determines whether an abstract is clicked (Table 3). Relevance here was determined by the human judgments on the abstracts. In both models, the F values of displayed position are always greater than the F values of relevance. In other words, when all factors are considered, subjects trust Google's positioning more than their rational judgments based on the evaluation of different alternatives.

Conclusion, Limitations, and Future Research

In summary, the findings here show that college student subjects are heavily influenced by the order in which the results are presented and, to a lesser extent, the actual relevance of the abstracts. These subjects trust Google in that they click on abstracts in higher positions even when the abstracts are less relevant to the task. When looked at in combination, the behavioral data (clicked choices) and the ocular data indicate that while there might be some implicit awareness of the conflict between the displayed position and their own evaluation of the abstracts, it is either not enough, or not strong enough, to override the effects of displayed position.

Table 2 Test of fixed effects in the mixed model of predicting clicks from all abstracts

Fixed Effect	Levels	F Value	Probability > F
Gender	1	1.43	0.23
Condition	2	3.61	0.03**
Query Type	1	0.03	0.86
Query Order	5	4.65	<.01**
Position	9	137.38	<.01**
Relevance Value	9	5.75	<.01**

Notes: N = 4125, the total numbers of all abstracts retrieved. ** Significant at .05 level

Table 3 Test of fixed effects in the mixed model of predicting clicks from viewed abstracts

Fixed Effect	Levels	F Value	Probability > F
Gender	1	0.40	0.53
Condition	2	0.36	0.70
Query Type	1	8.31	0.01**
Query Order	5	0.58	.72
Position	9	12.32	<.01**
Relevance Value	9	12.22	<.01**

Notes: N = 1219, the total number of abstracts viewed. ** Significant at .05 level

Trust is one mechanism humans use to reduce the complexity of decision making in uncertain situations (Luhmann, 1989) and may be viewed as a fast and frugal heuristic that exploits the regularity of the information environment (Gigerenzer & Selten, 2002; Simon, 1956). Google's information retrieval algorithm sorts the results by an estimate of the probability that it will fulfill the user's information need, thereby potentially reducing the cognitive effort and time costs for searchers. In order to determine whether trust does, in fact, lead users to the most relevant documents, we further asked human judges to rate the relevance of the actual pages instead of the abstracts on Google results pages. Further analysis showed that under the Normal condition, 51% of the time participants clicked on the abstracts that represented the most relevant Web pages. If the subjects were to have simply clicked on Google's number one ranked abstract, in 43% of the cases that would have resulted in them finding the most relevant Web page. Thus, trust does help the subjects reduce time and effort costs to locate the most relevant abstracts successfully some of the time.

What happens when users' trust extends beyond their individual evaluation of a set of returned results? To the extent that the PageRank algorithm works well on any given query, the imbalance between trust and evaluation may simply mean greater or lesser search costs, in terms of, say, time and/or effort. More critical, perhaps, is an increased probability of misinformation, particularly in circumstances of topic naiveté. More insidious, however, is the potential for misguided trust to exacerbate what others already fear regarding the non-egalitarian distribution of information (Hindman et al., 2003; Introna & Nissenbaum, 2000), whether as a result of economic resources, indexing policies, or algorithms.

Combining users' proclivity to trust ranked results with Google's algorithm increases the chances that those "already rich" by virtue of nepotism get "filthy rich" by virtue of robotic searchers. Smaller, less affluent, alternative sites are doubly punished by ranking algorithms and lethargic searchers. A study conducted by Cho and Roy (2004) using simulation experimented with the popularity of new Web pages under different information access methods. They compared how one page can become popular under two conditions: assuming all users do random surfing online versus assuming all users access information through a popularity-based

search engine like Google. Their results demonstrated that it takes 66 times longer for a page to become popular under the search-engine model.

Users, as a whole, are not familiar with how search engines "find" what they are looking for (Introna & Nissenbaum, 2000). The present results suggest that some users might benefit from having more information regarding the mechanisms by which Google and other search engines "crawl" the Web and determine how a website is ranked. Raising awareness through design is a promising direction. This might be accomplished on the results page of a search through a short explanation provided to the user on how the results were ranked, or perhaps through the visualization of the inbound and outbound link relationships that a website has fostered. As in social network analysis, users would be able to view central and peripheral sites, and more importantly, trace the connections or lineage between sites to determine the interest and relevance of a site based on its similarity to other sites and provenance. However, this requires a delicate design balance between maintaining simplicity and adding information content on search engine interfaces.

The limitations of this study lie primarily in its experimental nature. Research shows that information searching in a lab setting can be generalized to a larger context (Epstein et al., 2001; Schulte-Mecklenbeck & Huber, 2003). However, the subjects in the current study were all undergraduate students, in the same age group, who used Google as their primary search engine, and who trusted Google greatly; they conducted information searches in a lab setting on artificially designed search tasks. The applicability of the study results to a broader range of users and contexts is thus limited.

Google was chosen as the test search engine for the reasons described earlier; the generalization to other search engines is also limited. In a follow-up study we intend to investigate the generalizability of these effects with different existing search engines as well as with one that we contrive on our own, with no known history or a-priori influence on subjects. In this way we will be able to explore the pervasiveness of this trust and whether (and if so, how much) user trust transfers to other search engine contexts. In addition, we intend to recruit participants from different age groups, cultures, and search expertise, since our current study was limited to a group of college students and may not extend outside of that group.

Another promising direction is more refined analysis of the eye tracking data within the framework of information foraging theory (Pirrolli & Card, 1999), especially with respect to the cost of clicking and viewing lower ranked abstracts, as well as the value of information sent. In the meantime, as one of the first studies to explore evaluation processes in Web search, this study makes a significant theoretical and empirical contribution.

Acknowledgments

Google Inc. provided partial funding for this research. We also would like to thank Matthew Feusner for data preparation and discussions.

Notes

1 Corresponding author.
2 The authors found out that Asian subjects more likely fell into this category. This may be due to the fact that Asians tend to have less bright pupil retro-reflection, which the eye tracking device relies on to track eye movements (Nguyen, Wagner, Koons, & Flickner, 2002).
3 The instruction regarding scrolling was necessary because users might mistakenly think they could not manipulate the visual display on the computer. The instruction regarding scrolling will unlikely affect the results of the study since searching and scrolling are very habitual behaviors for the subjects in this study.
4 Under the Swapped condition, the results suggested that the subjects had fewer fixations, less regression, and checked more abstracts compared with the Normal condition. We refrain from discussing these results since they are not statistically significant.

References

Applied Science Laboratories. (2001). *Eye Tracking System Instruction Manual: Model 504 Pan/Tilt Optics*. Bedford, MA: Applied Science Group, Inc. Retrieved February 25, 2007 from http://www.cis.rit.edu/people/faculty/pelz/research/manuals/asl_504_manual.pdf

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, **91**(2), 276–292.

Bilal, D., & Kirby, J. (2002). Differences and similarities in information seeking: Children and adults as Web users. *Information Processing & Management*, **38**(5), 649–670.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, **30** (1-7), 107–117.

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

Brooks, T. A. (2004). The nature of meaning in the age of Google. *Information Research*, **9**(3). Retrieved April 12, 2005 from http://informationr.net/ir/9-3/paper180.html

Cho, J., & Roy, S. (2004). Impact of search engines on page popularity. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 20–29). New York: ACM Press.

DeGraef, P., De Troy, A., & d'Ydewalle, G. (1992). Local and global contextual constraints on the identification of objects in scenes. *Canadian Journal of Psychology*, **46**(3), 489–508.

Duchowski, A. (2003). *Eye Tracking Methodology: Theory and Practice*. London: Springer.

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across Internet and paper-and-pencil assessments. *Computers in Human Behavior*, **17**(3), 339–346.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, **108**(3), 316–255.

Gerhart, S. (2004). Do Web search engines suppress controversy? *First Monday*, **9**(1). Retrieved April 12, 2005 from http://firstmonday.org/issues/issue9_1/gerhart/index.html

Gigerenzer, G., & Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., & Wichansky, A. M. (2002). Eye tracking in Web search tasks: Design implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (pp. 51–58). New York: ACM Press.

Google Inc. (2005a). *Google corporate information: Google milestones*. Retrieved February 25, 2007 from http://www.google.com/corporate/history.html

Google Inc. (2005b). *Google press center: Zeitgeist*. Retrieved February 25, 2007 from http://www.google.com/press/zeitgeist.html

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management*, **35**(2), 141–180.

Granka, L., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 478–479). New York: ACM Press.

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). Effects of semantic consistency on eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, **25**(1), 210–228.

Hess, E. H. (1965). Attitude and pupil size. *Scientific American*, **212**(4), 46–54.

Hindman M., Tsioutsiouliklis K., & Johnson, J. A. (2003). *"Googlearchy:" How a few heavily-linked sites dominate the Web*. Retrieved February 25, 2007 from http://www.cs.princeton.edu/~kt/mpsa03.pdf

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

Huey, E. B. (1908). *The Psychology and Pedagogy of Reading*. New York: MacMillan.

Introna, L. D., & Nissenbaum, H. (2000). Shaping the Web: Why the politics of search engines matters. *The Information Society*, **16**(3), 169–185.

Jansen, B. J., & Spink, A. (2004). An analysis of Web searching by European AlltheWeb.com users. *Information Processing and Management*, **41**(2), 361–381.

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

Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of Alta Vista Web searching. *Journal of the American Society for Information Science and Technology*, **56**(6), 559–570.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 132–142). New York: ACM Press.

Josephson, S., & Holmes, M. E. (2002). Visual attention to repeated Internet images: Testing the scanpath theory on the World Wide Web. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (pp. 43–51). New York: ACM Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, **87**(4), 224–244.

Kahneman, D. (1973). *Attention and Effort*. New Jersey: Prentice Hall.

Lankford, C. (2000). Gazetracker: Software designed to facilitate eye movement analysis. *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (pp. 51–55). New York: ACM Press.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2005). The influence of task and gender on search and evaluation behavior using Google. *Information Processing and Management*, **42**(4), 1123–1131.

Luhmann, N. (1989). *Vertrauen: Ein Mechanismsum der Reduktion Sozialer Komplexitat*. Stuttgart: Enke.

Nguyen, K., Wagner, C., Koons, D., & Flickner, M. (2002). Differences in the infrared bright pupil response of human eyes. *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (pp. 133–138). New York: ACM.

Ozmutlu, S., Spink, A., & Ozmutlu, H. C. (2004). A day in the life of Web searching: An exploratory study. *Information Processing and Management*, **40**(2), 319–345.

Pan, B., Hembrooke, H. Gay, G., Granka, L., Feusner, M., & Newman, J. (2004). The determinants of Web page viewing behavior: An eye-tracking study. In *Proceedings of the 2004 Symposium on Eye tracking Research & Applications* (pp. 147–154). New York: ACM Press.

Pandey, S., Roy, S., Olston, C., Cho, J., & Chakrabarti, S. (2005). Shuffling a stacked deck: The case for partially randomized ranking of search engine results. In *Proceedings of the 31st International Conference on Very Large Data Bases* (pp. 781–792). New York: ACM Press.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review*, **106**(4), 643–675.

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

Rayner, K., Rottello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, **7**(3), 219–226.

Schulte-Mecklenbeck, M. & Huber, O. (2003). Information search in the laboratory and on the Web: With or without an experimenter. *Behavior Research Methods, Instruments & Computers*, **35**(2), 227–235.

Search Engine Watch. (2007). *U.S. search engine rankings and top 50 web rankings, January 2007*. Retrieved March 25, 2007 from http://searchenginewatch.com/showPage.html?page=3625081

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, **63**(2), 129–138.

Spink, A., & Ozmultu, H. C. (2002). Characteristics of question format Web queries: An exploratory study. *Information Processing and Management* **38**, (4), 453–471.

Stanford-Poynter Project. (2000). *Front Page Entry Points*. Retrieved February 25, 2007 from http://www.poynterextra.org/et/i.htm

Tatum, C. (2005). Deconstructing Google bombs: A breach of symbolic power of just a goofy prank? *First Monday*, **10** (10). Retrieved February 25, 2007 from http://www.firstmonday.org/issues/issue10_10/tatum/

Appendix A: Eye Tracking Apparatus

The subjects' eye movements were recorded using an ASL 504 commercial eye tracker (Applied Science Laboratories, 2005) that utilizes a CCD camera that employs the Pupil-Center and Corneal-Reflection method to reconstruct a subject's eye positions with a rate of 60Hz. A software application, GazeTracker, accompanying the system was used for the simultaneous acquisition of the subject's eye movements (Applied Science Laboratories, 2005; Lankford, 2000). Web pages were displayed on a 13-inch flat panel monitor with a resolution of 1024 by 768 pixels. The camera used, hereafter referred to as the pan/tilt unit, is an ASL Model 504 unit with the Model 5000 control unit. The pan/tilt unit is a remote eye tracker placed

underneath the flat panel display. It uses auto-focus with built-in illuminators to produce accurate reflections. The Ascension Flock of Birds magnetic head tracker was used in connection with the pan/tilt unit. A remote sensor, placed just above the participant's left eye, communicates with the sensor located on a post behind the subject's chair. Once calibrated, the sensor sends position information to the ASL pan/tilt unit to provide more accurate position information should there be any variability in a subject's head movements. The sensor enables the participant's eye to remain in the center of the pan/tilt field of view. The sensor is capable of making from 20 to 144 measurements per second and uses a pulsed DC magnetic field to communicate with the base unit. The tester's computer houses the ASL and head tracker software, while the participant's computer stores the Web pages to be viewed during the experiment (Lankford, 2000).

About the Authors

Bing Pan is Assistant Professor and Head of Research in the Office of Tourism Analysis and the Department of Hospitality and Tourism Management of the School of Business and Economics at the College of Charleston. He spent two years at Cornell University as a Post-Doctoral Associate at the Human-Computer Interaction Lab before joining the College. His research focuses on information search behavior, search engine marketing, online consumer behavior, and information technology use and development in the tourism industry.
Address: Department of Hospitality and Tourism Management, School of Business and Economics, College of Charleston, 66 George Street, Charleston, SC 29424 USA

Helene Hembrooke (Ph.D.) studies the use of technology and its social applications. Her current work focuses on using eye tracking devices as a method to investigate people's reactions to various aspects of websites—the visual and information density aspects—as well as information searching behaviors and decision making in online environments.
Address: 339 Kennedy Hall, Cornell University, Ithaca, NY 14850 USA

Thorsten Joachims is Associate Professor in the Department of Computer Science at Cornell University. His research focuses on statistical machine learning for intelligent information access and natural language processing.
Address: Cornell University, Department of Computer Science, 4153 Upson Hall, Ithaca, NY 14853, USA

Lori Lorigo is a Knowledge Manager in the Tuck School of Business at Dartmouth College. She obtained her Ph.D. in Information Science from Cornell University in 2006. Her research interests include information search and discovery, human-computer interaction, knowledge management, and information evolution and visualization.
Address: 1 South Park St. Apt. B, Hanover, NH 03755 USA

Geri Gay is the Kenneth Bissett Professor of Communication and Information Science at Cornell University. She is also the director of the Human Computer Interaction Lab at Cornell. Her current research interests include social networking, mobile computing and social interactions, collaborative computing, and interactive design and research.
Address: 339 Kennedy Hall, Cornell University, Ithaca, NY 14850 USA

Laura Granka is a User Experience Researcher at Google, Inc. She studies user behaviors in online search and uses eyetracking to build a comprehensive understanding of the search experience. She received her M.S. and B.S. from Cornell University.
Address: Google, Inc. 1600 Amphitheatre Pkwy Mountain View, CA 94043 USA