

EYE-R:
EYE-TRACKING ANALYSIS OF USER BEHAVIOR IN ONLINE SEARCH

A Thesis
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Master of Science

by
Laura Ann Granka
May 2004

© 2004 Laura Ann Granka

ABSTRACT

This thesis research is a behavioral and cognitive exploration into how users search for information online. Through a controlled eye-tracking experiment, the research presented here analyzes eye movements to assess how users evaluate, select, and retrieve information from an online search engine. Very few published information retrieval evaluations have specifically evaluated search and retrieval from an eye-tracking perspective, and this research provides the necessary exploratory insight to develop a greater understanding of online searcher behaviors.

Participants in the study were asked to search for ten different tasks using the Google interface. Search tasks were of two types – informational and navigational – and encompassed two levels of difficulty – low and high. Search behavior was measured through ocular indices, namely fixation duration and pupil dilation, as well as more overall measures of viewing behavior, including the total number of abstracts viewed, the total time spent searching, the average time spent viewing each abstract, and several others.

During this search process, subjects' eye movements were recorded by the ASL 5000 504 eye-tracking system. A proxy server was also established to log and cache all Web activity during the search sessions – including the unique search queries typed by subjects, as well as all the Web pages that were viewed and clicked. After the search and eye-tracking portion concluded, participants completed a post-questionnaire regarding their experiences with the system for the ten search tasks, as well as questions related to typical search behaviors. The study took place in the Usability Lab of Cornell's Information Science Building, located at 301 College Ave.

Key findings indicate that user search behavior varies as a function of task type and task difficulty, in addition to gender. Task type significantly influences pupil dilation,

fixation duration, and the total number of documents viewed, while task difficulty impacts the total time to make a selection as well as the number of abstracts viewed below the selected document. Furthermore, males appear to be more active when scanning the results, as they viewed significantly more abstracts within the results set than females.

BIOGRAPHICAL SKETCH

Laura Ann Granka completed her Bachelor of Science and Master of Science in the Communication Department at Cornell University. In the summer of 2004, she will intern at Google, Inc., and will be matriculating as a PhD student at Stanford University in fall 2004. She is looking forward to the California sunshine even though Ithaca really is Gorges.

To my family – Mom, Dad, Julie –

Thank you for everything!

ACKNOWLEDGMENTS

A number of people have helped me throughout my graduate career, and I wish to thank them all for their advice, support, and encouragement.

My committee – Geri Gay, Thorsten Joachims, Jeffrey Hancock, and Michael Spivey have given me incredible advice and support throughout this thesis. Francoise Vermeylen in the Office of Statistical Consulting offered a great deal of help with the statistical analyses in this thesis. Matthew Feusner helped with the invaluable task of formatting the eye-tracking data into something that could be analyzed. Helene Hembrooke, who I have worked with in the Human-Computer Interaction Lab, has been a great mentor for research. Finally, I would also like to thank Geri Gay and Thorsten Joachims for being outstanding advisors and mentors throughout my graduate career at Cornell. You have provided me with the opportunity to be creative, think big, and explore a number of research issues in HCI.

TABLE OF CONTENTS

Chapter 1: Introduction	1
Why use eye-tracking for information retrieval?	2
<i>Nature of online information retrieval</i>	2
<i>Existing information retrieval research</i>	4
Eye-tracking indices	9
Existing eye-tracking research	13
Chapter 2: Hypotheses and Research Questions	16
Overall viewing behavior	16
<i>Comprehensiveness of evaluation</i>	16
<i>Presentation of results</i>	19
<i>Salient regions</i>	21
Task dependent variables	22
<i>Task difficulty</i>	22
<i>Task type</i>	23
Subject factors	26
<i>Gender</i>	26
<i>Expertise</i>	28
Chapter 3: Methods	29
Experimental design	29
<i>Participants</i>	29
<i>Data Capture</i>	30
Experimental procedure	33
Analysis and justification of search tasks	34

Chapter 4: Analysis	40
General Descriptive Measures	40
<i>Total time</i>	40
<i>Total number of abstracts viewed</i>	41
<i>Order in which abstracts are viewed</i>	42
<i>Time spent viewing each abstract</i>	42
<i>How thoroughly is the results-set viewed?</i>	45
<i>What information in the abstract is most useful?</i>	46
Linear mixed models	46
<i>Data analysis structure</i>	48
<i>Task difficulty and task type</i>	50
<i>Total time</i>	50
<i>Number of abstracts viewed below selected document</i>	52
<i>Total number of abstracts viewed</i>	53
<i>Pupil dilation</i>	54
<i>Fixation duration</i>	56
<i>Gender and search behaviors</i>	57
<i>Number of abstracts viewed above selected document</i>	58
<i>Rank of selected document</i>	59
Survey Measures	62
<i>Online search usage</i>	62
<i>Task difficulty and satisfaction</i>	63
<i>Number of queries per task</i>	65
Chapter 5: Discussion	66
Descriptive measures of search behavior	66

<i>Viewing order</i>	66
<i>Time spent in each abstract</i>	67
<i>Click rate versus view rate</i>	67
<i>Number of queries per task</i>	68
Task Measures	69
<i>Task difficulty</i>	69
<i>Task type</i>	71
Subject Variables	74
<i>Expertise</i>	74
<i>Gender</i>	75
Chapter 6: Conclusions	78
<i>Information Visualization</i>	78
<i>Interpreting relevance from multiple clicks</i>	81
<i>Popularity of specialized, vertical portals</i>	82
<i>Impact on advertising</i>	83
Chapter 7: Future Research	85
<i>Relevance judgments</i>	85
<i>Search Success</i>	86
<i>Perceived trust and credibility in search engine</i>	86
Appendix A: Most Popular Search Queries	88
Appendix B: Sample Eye-tracking Output	89
References	91

LIST OF FIGURES

Chapter 1: Introduction

Figure 1.1: Overall scanpath depicting eye movements	10
Figure 1.2: Contour map of fixation intensity	11

Chapter 3: Methods

Figure 3.1 Eye-tracking apparatus and camera unit	30
Figure 3.2 Sample results page displaying “Lookzones”	32
Figure 3.3 “Hidden” link to CMU graduate housing	36

Chapter 4: Analysis

Figure 4.1 Total time spent searching, by search task	40
Figure 4.2 Number of abstracts viewed p/page, by search task	41
Figure 4.3 Total number of abstracts viewed per page	42
Figure 4.4 Arrival times to each abstract	43
Figure 4.5 Amount of time spent viewing each abstract	43
Figure 4.6 Abstract click rate to view rate	44
Figure 4.7 Number of documents viewed above/below abstract	45
Figure 4.8 Proportion of fixations to each portion of abstract	46
Figure 4.9 Overview of nested data structure	49
Figure 4.10 Interaction between task type & difficulty for total time	51
Figure 4.11 Differences in pupil dilation by rank of abstract	55
Figure 4.12 Interaction between task type & difficulty for fixation duration	56
Figure 4.13 Interaction between task attempt & task difficulty for abstracts viewed above	58
Figure 4.14 Search task difficulty and satisfaction	64

Figure 4.15 Number of queries formulated per search task 64

Chapter 5: Discussion

Figure 5.1 Proportion of clicks to gaze time 68

Chapter 6: Conclusions

Figure 6.2 Number of abstracts selected per page 81

LIST OF TABLES

Chapter 4: Analysis

Table 4.1 Total time – mixed model statistics	50
Table 4.2 Residual estimates for total time	51
Table 4.3 Abstracts viewed below – mixed model statistics	52
Table 4.4 Residual estimates for abstracts viewed below	52
Table 4.5 Total number of abstracts viewed – mixed model statistics	53
Table 4.6 Residual estimates for total abstracts viewed	53
Table 4.7 Pupil dilation – mixed model statistics	54
Table 4.8 Residual estimates for pupil dilation	55
Table 4.9 Fixation duration – mixed model statistics	57
Table 4.10 Residual estimates for fixation duration	57
Table 4.11 Abstracts viewed above – mixed model statistics	58
Table 4.12 Residual estimates for abstracts viewed above	58
Table 4.13 Rank of the selected abstract – mixed model statistics	59
Table 4.14 Residual estimates for rank of selected document	59
Table 4.15 Rank of selected document, by individual task	60
Table 4.16 Overall significant effects for task and gender	60
Table 4.17 Frequency searching for specific information	61
Table 4.18 Self-reported characteristics of search behavior	62
Table 4.19 Expertise, difficulty, satisfaction for search tasks	63
Table 4.20 Overview of significant findings	65

Chapter 5: Discussion

Table 5.1 Gender differences in “Time Machine actor” task	74
Table 5.2 Gender differences in undergraduate major	75

CHAPTER 1

INTRODUCTION

This thesis is a behavioral and cognitive exploration into how users search for information online. Through a controlled eye-tracking experiment, the research presented here analyzes ocular indices to assess how users evaluate, select, and retrieve information from an online search engine. To date, almost no published research has evaluated information search and retrieval interfaces from an eye-tracking perspective. The research presented here uses eye-tracking to provide the necessary insight towards developing a greater understanding of online searcher behaviors.

Although Web-based searches are becoming increasingly common, numerous studies have emphasized that searchers are still largely unsuccessful in finding their desired information, with failure rates often approaching 50% (Sherman, 2002, Hearst, 2000, Nordlie, 1999). Clearly, this presents a significant dilemma for online searches – why are users only modestly successful in formulating their search queries, and what can be done to improve the situation? Because the success rate of Web searches is still far from perfect, several attempts have been made to better understand user behavior during the search process, with the ultimate goal of drawing from these findings to enhance the performance of the search engine. In this thesis, I will review the relevant information retrieval literature, emphasizing where the current gaps lie and how this study can help to construct a more comprehensive account of online search behavior. Subsequently, I will present a review of the relevant eye tracking literature and present several hypotheses and research questions that this research will address.

While this research is unprecedented, it should still be placed in the context of existing literature in the areas of eye-tracking and information retrieval. The conclusions and methods from the most relevant research in these two areas can then be incorporated to help inform this present evaluation.

WHY USE EYE-TRACKING FOR INFORMATION RETRIEVAL?

Understanding how users evaluate and read the abstracts that are presented to them in an online search is important for a number of reasons. First, understanding how a searcher skips, scans, or critically reads the retrieved abstracts enables us to gain more insight into the underlying user motivations and cognitive processing involved in the online document selection process. Secondly, knowing how a user reads through the retrieved documents can provide insight for enhanced interface design: if we can capture user viewing patterns, then we can conceivably generate new visual displays that facilitate and accommodate a searcher's natural viewing behaviors. Finally, eye-tracking measures can lead to more robust understanding and interpretation of implicit feedback, such as click-through measures and log data from an online information retrieval (IR) session. Click-through and log measures are relatively robust in that they record the Web pages that searchers clicked, as well as the amount of time spent on each page. Although click data is representative of the pages and links viewed and clicked on, alone it cannot fully inform us of a user's behavior and cognitive processing. Eye-tracking can supplement this log data to provide us a better indication of what these clicks actually represent.

Nature of Online Information Retrieval

Web-based information retrieval differs from traditional information retrieval contexts, such as physical libraries, and even digital libraries, in a number of ways.

One of the most distinct differences is that online, searchers are not only able to search for specific bits of information, but for specific Web pages. For instance, a searcher may use an online search engine to query “Emeril the Chef” (for the Food Network Chef) with the attempt of finding Emeril’s homepage, <http://www.emerils.com>. However, if “Emeril the Chef” was attempted in a digital library catalog, a personal homepage result would not be the desired result; instead, the results would be much more informative in nature, possibly about Emeril’s recipes, or periodicals with news stories featuring Emeril.

Yet another distinction with online queries is that they may also have a “transaction” as the end goal. For instance, searchers have the opportunity to structure online queries with the intent of selling or purchasing something, downloading music, retrieving a map, among other things. Conversely, in traditional IR, searchers may more often desire books and periodicals, and will more frequently search through author and title keywords.

Because the aforementioned examples all represent types of queries that are best conducted online, Broder (2002) developed a taxonomy of web searches to categorize the most common types of queries that are possible in an online networked environment. When using a Web search engine, the “need behind the query,” as Broder states, is not always to find a specific bit of information, but it can also be to find a specific page or to “perform some web-mediated activity.” Thus, Broder developed a classification scheme for the three types of queries that users may perform online – informational, navigational, and transactional. Informational queries are designed to retrieve information, similar to in traditional non-Web information retrieval. Navigational queries are to find a specific homepage, and transactional queries are to perform Web-based “transactions.” Through a user survey and an analysis of Web query logs, Broder found that the total percentage of navigational

queries was between 20-24%, the percentage of informational queries was between 39-48%, and the number of transactional queries was between 22-36%.

It is important to note, however, that although there is a classification scheme for the most distinct types of online queries, there is not always a definitive boundary between these three groups. Transactional queries are often the most difficult to explicitly distinguish because queries leading to transactional sites such as ebay.com or amazon.com could potentially also fall into the navigational category. The two main differences – querying for information versus querying for a Web page – were the two primary differences that we wanted to address with the present research, and as such, only informational and navigational queries were specifically included in this study.

Existing Information Retrieval Research

The information retrieval literature can largely be split into two distinct classes of research – those that investigate the effectiveness and efficiency of the search engine, and those that evaluate the behaviors that users employ when engaged in an online search. The present research focuses primarily on the latter component of user behaviors, though the findings are also intended to also provide direct recommendations for retrieval performance, particularly with respect to relevance judgments that are increasingly being determined by machine learning through implicit feedback (Joachims, 2002; Drucker, 2002).

A majority of current information retrieval evaluations focus on evaluating the functionality of the actual search *system*, in terms of its efficiency, precision, and ranking of results. These retrieval evaluations focus on assessing the precision and recall of the selected documents as a measure of the information retrieval system itself. Less focus has actually been placed on the *user*, towards addressing the explicit

and cognitive behaviors through which users evaluate online documents (Spink, 2002; Hembrooke, Granka, & Gay, under review; Hsieh-Yee, 2001). In this research, although we will also evaluate the relevance of the results set generated by the search system, our primary concern is to instead highlight the behavioral processes through which searchers assess and evaluate the online documents they are presented with and selectively single out a relevant result from its counterparts. Furthermore, the IR studies that do assess retrieval behaviors primarily do so by using Web log files and click-stream measures to determine the pages and documents that a searcher clicks on. Although these measures can certainly lead to informative evaluations, click-through data serve primarily as *outcome* measures, specifying only the selection that a searcher has made. To effectively supplement these outcome measures, it is desirable to gain insight into users' retrieval behaviors *before* an actual document is selected. In addition to knowing only what pages a searcher views, more specific eye-tracking metrics can inform us of which abstracts on the results page are viewed, as well as the amount of time a searcher spends reading each of the displayed results.

Capturing eye movements during search will address pre-click user behaviors and provide a comprehensive account of how a user views and finally selects an online document. An analysis of eye movements will enable us to recognize what a searcher reads, skips, or scans; and it is these ocular indices that can provide critical insight for comprehensively analyzing current information retrieval systems to ensure that they are optimized to match and accommodate actual user viewing behaviors.

User Studies of Information Retrieval Behavior

A number of studies have addressed user behaviors in Web-based information retrieval, and for the purposes of this study, I have further classified this group of research into three sub-categories. The largest of these three sub-categories contain

studies centered on issues of navigation and query formation, and the primary methodology employed in these studies is click-through and log file analysis (Jansen, Spink & Saracevic, 2000; Jansen & Pooch, 2000; Silverstein, 1998). These studies have led to such findings as the average number of query terms in an online search, the most common types of online queries, and the number of search tasks per session. Some of the primary findings indicate that online search queries typically contain only two to three terms, users primarily click on only one document in the results set, and approximately 90% of the users never go beyond the first page of results.

Another large body of user-centered literature relates to data visualization, and several researchers have developed alternate representations to display the documents listed in the results-set. As the primary retrieval display involves only a linear listing of results, ranked by system relevancy judgments, alternate visualizations have involved clustering items, hierarchical graph displays, and categorical presentations equipped with sliders and interactive manipulation features (Schneiderman, et al, 2000; Hearst, M., 1999, Pirolli, Card, & Van Der Wege, 2001).

Finally, a much smaller body of research evaluates search system usability, and looks at the extent to which users use the graphical user interface (GUI) features, such as the number of clicks on layout buttons, time spent scrolling, and conducting visual searches (Spink, 2002). Although this last group of research comes closer to matching the procedures used and objectives in the present study, the scope is still somewhat limited when compared to the utility of eye-tracking information. To the best of my knowledge, only one study has previously used eye-tracking in the context of information retrieval (Salogarvi, Kojo, Jaana, & Kaski, 2003). This study used measures of pupil dilation to infer the relevance of online abstracts, and found that pupil dilation increased when fixated on relevant abstracts. However, this study only

collected eye movements from three subjects, so the generalizability is a bit weak, and furthermore, no other measures of searcher performance were addressed.

Therefore, it is important to add to the existing body of information retrieval by providing insights gained through eye tracking. Eye tracking provides an account of the users' subconscious behavior and cognitive processing, which is an important component that is often missed from the existing user studies. Even in user-centered studies, participants do not always exactly know or clearly articulate what they want, and thus subconscious behavioral indices can effectively supplement measures of user responses.

In sum, the common trend throughout many of the existing user studies is that they look primarily at the outcome measures of user behavior, quantified in such ways as query wording, time spent searching, and the rank of the selected document. These measures all are produced *after* the subject has selected a document. To supplement these measures, it is also important to understand the actual process whereby users reach a decision about which document to select. Much less research has been placed on evaluating the behaviors leading up to a document selection, most specifically, with eye movements.

Online Relevance Evaluations

Traditional IR evaluations primarily rely on manual relevance judgments. In these evaluations, human experts individually assess the relevance of the documents that a user is presented with – this includes the documents that a searcher selects, and those that the searcher overlooks. Because this manual procedure is inevitably time-consuming and costly, researchers have begun to use more implicit measures as an assessment of relevancy (Joachims, 2002; Drucker, 2002). Automating this evaluation process is desirable in order to facilitate and accelerate the evaluation process, as well

to enrich and expand the potential scale of the evaluation. For example, Joachims has used machine learning techniques¹ to learn relevance judgments from implicit clickstream measures. Machine learning offers a number of advantages to predicting relevance, chiefly in that click-through data can be readily recorded in large quantities, offering a large sample training set for the researcher.

However, one of the main problems of using clickthrough data is that in order to fully maximize its utility, several assumptions about searching behavior need to be made. For instance, when initial studies evaluating document relevance emerged, relevance judgments were often made amongst the top ten results displayed by the search engine. But, do users actually view and assess all ten? The frequency with which a user scrolls below the first screen shot to view all ten documents in their results set is unknown. It may be, for example, that a searcher does not even scroll down the page to view the documents listed below the first screen shot.

Furthermore, current evaluation algorithms traditionally assume that individuals view documents in a linear order. Based on these models, top-ranked documents are given a higher weighting in many retrieval algorithms because it is assumed that they are more likely to be viewed and relevant to the searcher than those ranked lower in the set (Joachims, 2002). If a user clicks on documents ranked 1, 5, and 6, it is assumed that the results nested between the selected documents (in this case, 2, 3, and 4) are also viewed, but are considered less relevant by the searcher. Although this assumption is reasonable, the fact that users view documents linearly has not been empirically or systematically tested. Furthermore, it is debatable whether a user's viewing behaviors are truly regular and consistent across all queries within an online search context.

¹ Machine learning is a computer science technique that leads to improved performance by learning from previous results.

One final assumption of evaluation algorithms is that users view all of the abstracts above the clicked result (Joachims, 2002). Although plausible, it may be that some abstracts are viewed more carefully than others, or that users randomly skip over abstracts. Thus, not much tangible evidence exists for describing the users' information retrieval process *before* a document is selected, and it is necessary to use eye tracking to develop this understanding. Because these assumptions have not been systematically or empirically tested, it is important to address some of these concerns through eye-tracking. The results of eye-tracking evaluations can be used as behavioral and empirical evidence to corroborate and potentially supplement the algorithms used, with the ultimate goal of offering enough behavioral evidence to verify their accuracy and effectiveness.

EYE TRACKING INDICES

The research presented here seeks to obtain a more comprehensive understanding of *what* the searcher is doing and reading before actually selecting an online document, and one of the most comprehensive ways to approach this is through eye tracking. Ocular indices will enable us to determine what documents a user is viewing and reading, for how long, and in what order. An eye tracking evaluation will enable us to examine some of the assumptions that have limited the traditional evaluations of information retrieval systems. In particular, eye tracking offers a window into how individuals read and scan the information displayed to them, and particularly by relating eye movement behavior with subsequent questionnaires and logged clickstream data, one can obtain a more comprehensive picture of the process through which online information acquisition actually occurs.

Throughout the history of eye tracking research, several key variables have emerged as significant indicators of ocular behaviors, including fixations, saccades, pupil dilation, and scan paths (Rayner, 1998; McConkie & Loschky, 2002) (Figure 1.1).

Fixations

Eye fixations are defined as a spatially stable gaze lasting for approximately 200-300 milliseconds, during which visual attention is directed to a specific area of the visual display. Fixations are traditionally understood to be indicative of where a viewers' attention is directed, and represent the instances in which information acquisition and processing is able to occur (Rayner, 1998). Based on existing literature, a very high correlation has been found between the display item being fixated and that being thought about; similarly, there is a close connection between the amount of time spent fixated on certain items and the degree of cognitive processing (Just & Carpenter, 1980, Rayner, 1998). Eye fixations are the most relevant metric for evaluating information processing primarily because other indices, such as saccades,



Figure 1.1 Overall scanpath depicts pattern of movement throughout screen. Eye fixations are represented by the black circles, and saccades are indicated by the adjoining lines.

occur too quickly to absorb new information (Rayner, 1998).

According to Viviani (1990), at least three processes occur during an eye fixation: encoding of a visual stimulus, sampling of the peripheral field, and planning for the next saccade. Research has shown that information complexity, task complexity, and familiarity of visual display will influence fixation duration (Duchowski, 2002).

The length of eye fixations is also largely dependent on a users' task. The average fixation duration during silent reading is approximately 225 ms, while other tasks, including typing, scene perception, and music reading approach averages of 300-400 milliseconds. From an eye tracking perspective, information retrieval seems to encompass both a visual search as well as reading, so it is expected that the average fixation duration will fall within the range of these two groups. The differences in fixation length can be attributed to the time required to absorb necessary information, and the speed at which new information should be absorbed. It is necessary for the eye to move rapidly during reading, while in visual search and scene viewing, it is less imperative that the eye quickly scans the entire scene, but rather that the user can absorb key information from certain regions.

To analyze the relative popularity of key content areas, Wooding (2002) first



Figure 1.2 Contour map of Google showing the intensity of eye fixations. Higher peaks indicate regions of more fixations (Feusner, 2004).

developed the concept of “fixation maps,” which generally depict a contour analysis of the intensity of eye gaze. Research done at Cornell’s Human-Computer Interaction lab has used these contour fixation maps to generate a comprehensive aggregate evaluation of eye movements (Figure 1.2). By understanding what regions of a page generate the most attention, convergent measures can be introduced to offer a more reliable indicator into *why* certain regions of a page attract attention.

Saccades

Saccades are the continuous and rapid movements of eye gazes between fixation points. Because saccadic eye movements are extremely rapid, within 40-50 milliseconds, and approaching velocities of nearly 500 degrees per second, information acquisition is unable to occur during this time. This lapse of information intake is traditionally referred to as “saccadic suppression,” but because saccades represent such short time intervals, individuals are unaware of these breaks in information perception (Rayner, 1998). An analysis of saccadic movement has been conducted in the context of information processing and reading, and the literature evaluates and distinguishes several properties specific to saccades.

Pupil Dilation

Pupil dilation is a measure that is typically used to indicate an individual’s arousal or interest in the viewed content matter, with a larger diameter reflecting greater arousal (Duchowski, 2002; Rayner, 1998; Hess & Polt, 1964). Studies can compare the average pupil dilation that occurs in a specific area of interest with the average pupil dilation of the entire site to gain insight into how users might cognitively understand or process the various content matter (Hess & Polt, 1960).

Scanpath

A scanpath encompasses the entire sequence of fixations and saccades, which can present the pattern of eye movement across the visual scene. User scanpath behavior provides insight into how a user navigates through the visual content. Studies analyzing properties specific to scanpath movement have enabled researchers to create a more comprehensive understanding of the entire behavioral processes during a visual search or scanning session (Josephson & Holmes, 2002). Existing literature suggests that scanpath movement is not random, but is highly related to a viewer's frame of mind, expectations, and purpose (Yarbus, 1967). When looking at particular content areas, several studies exploring eye movement locations determined that unique regions of a visual piece are fixated on sooner than others (Antes, 1974).

Existing Eye Tracking Research

Although no eye-tracking evaluations have closely addressed the questions addressed in this research, several findings from related eye-tracking applications can be incorporated and adapted to our information seeking context.

As previously mentioned, only one existing study has used eye-tracking in the context of information retrieval evaluation (Salogarvi, et al, 2003). The researchers linked relevance judgments to increases in pupil diameter, as a larger diameter typically signifies high interest in the content matter. However, the sample size and search tasks in this experiment were not robust enough to generate predictable patterns of user search and scanning behavior. In order to offer more substantial conclusions, the present study includes a much larger sample size to generate more valid measures from which to understand these issues. Furthermore, this thesis will also investigate a number of other ocular indices and measures of search behavior in addition to pupil dilation, to provide a more comprehensive account of the user search process.

The greatest strength of eye-tracking is to highlight *what* a user is looking at, and in the context of information retrieval, eye-tracking can depict the process whereby individuals view and read the results presented to them. From this, it follows that eye-tracking can best be used to assess the actual behaviors that users employ when reading and making decisions about which documents to select and retrieve online.

More specifically, eye-tracking can depict how much time searchers spend viewing each abstract, whether subjects view documents in a linear order, and overall, how comprehensively users evaluate online documents. This present study can clue us in as to whether users immediately click on the first relevant link they see, or whether searchers prefer to make more calculated judgments by evaluating several possible abstracts.

Online Decision Making

Although no present studies have used eye-tracking to assess online search decision-making, perhaps the most relevant study to draw from is one that uses eye-tracking to evaluate the process of consumer choice when purchasing consumer goods in a simulated shopping experience (Russo & LeClerc, 1994).

In this study, Russo and LeClerc drew from existing theories related to consumer choice, which pinpoints the stages or phases in which individuals make a decision about what product to purchase. There are multiple theories depicting the process of consumer decision-making, but one primary debate questions whether decisions are made within a two-stage choice framework, or a three-stage framework. Russo and LeClerc's eye-tracking investigation offered evidence in support of a three-stage model of consumer choice – orientation, evaluation, and verification. Through eye movements, the researchers pinpointed several key behavioral differences between the three stages, particularly that the first and last stages of this model were relatively

short, and were marked by rapid eye movements, while the second stage is where the primary decision-making and evaluative decisions emerge.

CHAPTER 2

HYPOTHESES & RESEARCH QUESTIONS

The following chapter will address the hypotheses and research questions that were analyzed in this study. The hypotheses all fall into three distinct categories, each addressing a unique aspect of the search experience. The first section will address overall descriptive measures of user searching behavior, including elements such as total search time and the order in which abstracts are viewed. The second section will address the task factors influencing search success, including task type (informational and navigational) and task difficulty. Finally, the user characteristics that affect the search experience will also be addressed, including gender and expertise. A review of the related literature will be provided for the justification and support of all hypotheses.

OVERALL VIEWING BEHAVIOR

Comprehensiveness of Evaluation

One of the most fundamental issues to address is how comprehensively users evaluate the list of abstracts that is presented to them when selecting an online document. This sort of analysis will enable us to determine whether users skip links, make choice comparisons against nearby abstracts, and in general, assess how searchers evaluate each of the abstracts presented to them. Some of the most fundamental questions to address are how much time users spend before selecting an abstract, as well as how many documents are viewed in this time. These measures can assess how much critical evaluation and thought users may devote towards making a selection. Thus:

RQ1: How much time do users spend searching before selecting a document?

RQ2: How many abstracts do users look at before selecting a document?

Other aspects regarding the comprehensives of evaluation can more specifically compare users' eye movements to existing literature regarding decision-making. In addressing this issue, it will be useful to compare user behaviors from this online search to existing literature that uses eye-tracking to assess other types of decision-making. More specifically, we can analyze these findings within the context of a decision making framework, such as Russo and LeClerc's (1994) three-stage model, which was addressed in Chapter 1 of this thesis.

As noted previously, Russo and LeClerc pinpointed several key behavioral differences between the three stages, particularly that stage one was marked by eye movements indicating that the viewer is rapidly constructing an overall impression of the products available. The second stage is where the buyer critically compares product "A" with Products "B" and "C," making two-way and three-way comparisons about the relative utility of their alternatives. Finally, the third stage is marked by verification – the buyer again evaluates the properties specific to the document that they are planning to purchase to ensure this is indeed the desired product.

However, as interesting and useful as this model may be to consumer product selection, it is unlikely that we can critically evaluate the online decision-making process in quite the same way that we are able to assess consumer choice. It is unlikely that the behaviors represented in the consumer choice three-stage model will fully emerge in the context of online information retrieval. First, the decision-making and purchase of a tangible product is substantially different than the decision-making and selection of an online document. Two of the most obvious differences relate to the cost and permanence of the selection. In an online context, the user has nothing to lose (expect for a few seconds of time) by immediately clicking a link without a

comprehensive assessment. If the document does not match expectations, the user needs only click the “back” button or reformulate their query to find a better selection. Conversely, if a shopper makes an impulse purchase, there is the chance that they have wasted their money, or that they will have to later return to the store for an exchange or refund of the item.

Furthermore, when in a store, there is a finite selection of products to choose from. Although an online search presents results in a finite fashion (ten results per screen), it is possible for the search engine to retrieve hundreds of thousands of documents that match the search query. The searcher has considerable opportunity to type in a new query and generate another expansive set of results.

Thus, based on these distinct differences it may be difficult to apply a three-stage choice framework to the context of online document selection. Other two-stage frameworks have emerged in the consumer and marketing literature, which first involve an initial screening of alternatives, followed by a more thorough evaluation of the remaining options (Biehal & Chakravarti, 1986; Payne, 1976). Although these have not been evaluated through eye-tracking, they may more closely match the actual behaviors that searchers employ in an online context, simply because a decision in online search is made more quickly than in the context of consumer non-durables.

A final alternative is that because the cost of selecting an online document is so low, and because the speed and efficiency of search engines is so high, it is also probable that searchers’ behaviors are not extensive enough to be partitioned into separate stages, but rather that their behaviors are best described through one general selection process.

Because of these differences, it will still be interesting to evaluate the eye-tracking metrics from this study in the context of a choice framework, to determine if there are indeed distinct stages to the online decision-making process. If there are distinct

stages, this may be useful for advertisers on search engines, as one stage or behavior may make searchers more susceptible to advertising messages. Therefore, several research questions can be formulated. First, to determine whether searchers make choice comparisons with nearby abstracts:

RQ3: On average, how many abstracts above the selected document are viewed?

RQ4: On average, how many abstracts below the selected document are viewed?

RQ5: Do eye movements in the first few moments of page viewing significantly differ from the latter movements?

Presentation of Results

The presentation of the retrieved abstracts is expected to significantly influence ocular indices and the manner in which users navigate through and select an online document. In his article, *The Intelligent Use of Space*, Kirsh (1995) outlines several key features about managing and understanding space that may be relevant to the context of document selection in an online information search system. Kirsh points out that one of the goals in organizing both physical and information spaces is to structure the space so that a user's "option set" (the number of viable alternatives available to the user) is reduced and more effectively managed. A reduction in the users' option set is typically accomplished in an information retrieval interface by rank-ordering the retrieved results. By having the system highlight the opportunistic actions that a user should take, the amount of effort that the searcher ultimately needs to expend making a decision is reduced. Because most searchers understand that the top-ranked result has been rated most highly by the information retrieval system, one of the steps to decision-making – organizing the information – is thus eliminated.

Rank ordering the list of retrieved documents is an organizational strategy that is expected to reduce the cognitive effort (what Kirsh refers to as “internal computation”) of the searcher. By analyzing the eye movements in the list of retrieved documents, we can detect whether searchers are using the system’s presentation of results to facilitate their search. If users use the linearity of the results set to assist their searching, we will be able to detect this through eye movements. Therefore:

RQ6: Do users view documents in a linear order?

The linear presentation and rank-ordering of results may eliminate or reduce the initial “orienting” stage which usually occurs when a viewer is presented with a display. If the user knows that they will consistently be presented with ten documents per page, ranked in descending order of importance, there is much less of a need to first scan and evaluate the entire space, but rather view the documents sequentially.

This literature raises a number of questions to be addressed in the context of an online search, particularly related to the manner in which subjects view the results presented to them, primarily:

RQ7: Do users first scan the entire display before more careful processing and subsequent document selection?

Interestingly, because location does affect scene viewing, it is also of interest to determine whether abstract placement alone can affect ocular indices. Click-through data from online information retrieval studies demonstrate that the links most likely to be selected are located at the very top or very bottom of the results set. It seems that the middle portion of results does not generate as much attention or interest. Thus, it would be interesting to use ocular indices to ascertain whether the middle-ranked abstracts really do receive less attention and interest. To do this, pupil dilation and total time spent in each abstract can be used as indirect measures of interest and cognitive processing. Based on the initial click data studies, one can hypothesize:

H1: The abstracts displayed in the middle of the results set will be viewed for less time than the results on the periphery (the top and bottom two or three).

H2: Pupil dilation will be lower for abstracts in the middle of the results set than for those on the periphery.

Salient Regions

Another technique that is also used to manage the online space in information retrieval relates to the issue of salience. Much of the visual search literature emphasizes salience, in attempts to understand the key features that will attract a viewers' attention to key regions of a scene or display. (Henderson, 2003; Wolfe, 2003; Loftus & Mackworth, 1978; Lohse & Rosen, 2001). Much of the work related to salience and attention has been done in the fields of marketing and advertising, as well in traditional psychological visual search studies. Less work has been conducted within the context of Web pages, but it has been found that color, size, and location are all prime indicators of whether or not a user initially responds to an object within a scene (Lohse & Rohse, 2001, Granka, Hembrooke, Gay, & Feusner, 2004). The Google interface, which was used for the present research, effectively provides salient cues for users, particularly by using color and size as effective tools within the displayed abstracts. On the Google interface, the titles are displayed in a blue text with a somewhat larger and bold font; the URL is presented in a green color; and the snippet is presented in black text with bolded key terms. Employing strategic use of color and size is likely to attract a users' attention to these key regions. For instance, do users scan the titles of the abstracts first, and then only further evaluate the snippet and other portions of the abstract if the title looks promising? Other questions of interest are which of these regions does a user spend the most time in, and which is the most useful for determining document relevance. Thus:

RQ8: On which parts of the abstract do users first fixate?

RQ9: Which parts of the abstract do users spend the most time viewing?

TASK DEPENDENT VARIABLES

Task Difficulty

Total Search Time

Task difficulty was expected to affect eye movements in several ways; one of the initial assumptions is that difficult tasks will require users to spend more time searching. Foremost, it is expected that for more difficult tasks, searchers will have a more ambiguous interpretation of what is needed to fulfill the task. Because of this, it may be necessary for searchers to spend more time evaluating the retrieved results to more fully understand the information seeking need. Thus, it is speculated that total time taken to select a document will vary by difficulty of the task. Likewise, it is also probable that for difficult tasks, the user will look at more abstracts ranked below their selected document, as they may need to more extensively verify that they are selecting the most appropriate abstract. Furthermore, because the total number of fixations are inevitably highly correlated with the total time spent viewing a particular scene, it is also postulated that the total number of fixations will also increase as the task difficulty increases.

H3a: As task difficulty increases, the time taken to select a document will increase.

H3b: As task difficulty increases, the total number of abstracts viewed will increase.

H4: As task difficulty increases, searchers will view more abstracts below their selected document.

Pupil Dilation

Furthermore, much research has shown that pupil dilation increases with task difficulty (Pelz, et al, 2000, Hess & Polt, 1960). Pupil dilation has been measured to be an indicator of interest, arousal, and processing, and our pupil becomes larger when we need to expend more cognitive effort or energy towards performing a task.

Therefore, it would follow that:

H5: Pupil dilation is expected to increase as task difficulty increases.

Task Type

Total search time

As previously discussed in the introduction, the two main task types to be analyzed in this experiment include informational and navigational queries. It is hypothesized that these two types of tasks will influence user responses in slightly different ways, particularly with regards to the total time spent searching, the number of abstracts viewed, and fixation duration and pupil diameter. Because navigational queries lead to one primary URL as the search outcome, the user may have a better estimation of what to expect for these types of queries. For a number of tasks, URLs can be guessed fairly consistently (for instance the homepage for Cornell University is www.cornell.edu). For these types of queries, the title and URL within the abstract may offer the most relevant information, and searchers may find their search goal more efficiently by scanning through URLs and titles without needing to carefully read the information in the snippets (the portions of article text that Google presents). Conversely, the desired results for informational queries can be found on multiple pages, and in these instances, while the URL may provide an indication of credibility, reading the snippet to understand the context of the query terms may be the most effective way to accomplish this type of task.

Furthermore, because there is only one expected outcome for the navigational tasks, it is likely that searchers may find the result that they need and immediately click on it, without evaluating any of the alternatives listed below. Alternatively, informational queries may require that the searcher evaluate and make a comparison between several abstracts, as it is possible for each of them to contain the desired information. Based on the differences between these query types, several hypotheses related to total searching time and the total number of abstracts viewed, can predict how user behaviors may differ between the queries.

H6: Users will spend more time searching for informational queries than navigational queries.

H7: Users will look at fewer documents per results set for navigational queries than informational queries.

H8: Users will be more likely to look at documents presented below their selected abstract for informational queries than navigational queries.

Fixation Duration

Fixation duration is another ocular index that has been linked to task difficulty. Interestingly, changes in fixation duration are largely dependent on the type of task with which an individual is involved. In visual search, fixation duration, saccade length, and saccade occurrence have also been found to decrease as tasks become more difficult (Nakayama, Takahashi, 2002). When a task becomes increasingly difficult, fixation duration decreases because more frequent and shorter fixations are often required to more fully process the displayed information. Fitts et al. (1950) has also concluded that fixation frequency in an area of interests an indication of the degree of importance whereas fixation duration is an indication of the complexity and difficulty of visual display.

Conversely, in reading tasks, as tasks become more difficult, fixation duration increases (Rayner, 1998; Just & Carpenter, 1980). When readers are presented with material that is more difficult to comprehend, they will spend more time fixating on difficult words, as well backtracking and re-reading over earlier portions of the sentence.

As both bodies of literature – reading and visual search – offer different implications for the interpretation of fixation duration, it may at first appear difficult to interpret these findings to the context of online information retrieval. This thesis approaches online information search as a mediary between visual search and reading. Users are searching for information, yet in order to discover their desired source, it is likely that they will need to read some information in the abstracts. However, even with this interpretation of online search, it is still unclear as to whether increases in task difficulty will contribute to shorter or longer fixation durations. Therefore, in order to further explore this issue, it becomes necessary to further specify the effects of difficulty within the context of task type.

As previously addressed, it is possible that users will have different scanning and information retrieval styles based on the type of task that a user is performing. If a search is for a specific Web page or homepage, the searcher may be more likely to scan titles and key words; if the search task is informational, the searcher may be more likely to read the snippets of text displaying the context in which their search terms appeared. Based on this assumption, it may be the case that navigational queries induce eye movements that are more characteristic of a visual search, while informational queries generate behaviors more similar to reading. Thus:

H9a: For navigational tasks, fixation duration will decrease as task difficulty increases.

H9b: For informational tasks, fixation duration will increase as task difficulty increases.

Finally, fixation duration has been found to be shorter in the first few seconds of scene viewing, likely due to an orientation response in which the user is attempting to generate a “gist” of what the entire display contains (Henderson, 2003). Likewise, it is expected that in the first few moments of being presented with the list of retrieved results, searchers will also exhibit more frequent and shorter fixations to best acquaint themselves with the results set. Therefore:

H10: The first few fixations on each results page will average shorter durations.

SUBJECT FACTORS

Gender

Gender has not been extensively addressed as a potential factor influencing online search behaviors. A large portion of the research exploring gender differences has investigated the impact of gender on behaviors toward technology, Web page viewing, and search query structure. Interestingly, most of the studies finding gender differences in the context of Web search were conducted with populations of elementary and middle-school aged children (Large, Beheshti, & Rahman, 2002; Roy, Taylor, & Chi, 2003). Because school age children have unique attributes, it is somewhat difficult to generalize these results to older populations.

One study involving information search focused on the issue of gender, and found that boys were able to gain more knowledge than girls after a Web-based search session. Furthermore they found that boys scanned significantly more of the presented abstracts while girls clicked on and accessed significantly more than boys (Roy, Taylor, & Chi, 2003). Another study addressing gender in information retrieval

offered evidence that boys are generally more “active” on the Web when compared to girls, as boys clicked on more hyperlinks, formulated more queries, performed more page jumps per minute, and also gathered and saved information more frequently. Also, boys were more likely to type in queries containing only one word, while girls used more natural language queries than boys (Large, Beheshti, & Rahman, 2002).

Other gender differences in perceptual processing have also been reported in non-online search contexts. Meyers-Levy and Maheswaran (1991) demonstrated that males and females differ with respect to the selection processes, with females more often engaging in comprehensive processing of all available information, while males tend to focus their attention on a fewer number of areas. Furthermore, an eye-tracking study analyzing eye movements across popular Web pages found that there were significant gender differences in Web browsing, with females exhibiting shorter fixation durations than men, suggesting more comprehensive scanning of the entire Web page (Pan, Hembrooke, Gay, Granka, Feusner, & Newman, 2004). Based on this literature, research questions were formulated to address the issue of gender in a Web-based information retrieval context:

RQ10: Does the total number of abstracts viewed differ significantly by gender?

RQ11: Does the rank of the selected abstract differ significantly by gender?

RQ12: Does the total time taken to select a document differ significantly by gender?

Expertise

Expertise in information retrieval has been explored on two levels: familiarity or experience with the search system, and expertise with the search topic (domain knowledge) (Fenichel, 1979; Harter, 1984; Hoelscher & Strube, 1999; Hsieh-Yee,

1993; Jansen, et al., 1998 Lazonder, et al, 2000; Marchionini, 1995). Many studies have concluded that experts and novices approach the task of information search from different perspectives, particularly as they each contribute varying degrees of existing knowledge to the search process. In turn, these factors reflect both the sources that a user will look for, as well as a searcher's reliance on browsing behaviors.

In this study, only the issue of domain expertise is addressed, as prior research suggests that a user's system-based search expertise is typically less influential in determining success and failure than is domain knowledge (Nordlie, 1999; Marchionini, 1995). Furthermore, all subjects reported at least a general familiarity with the Google search interface, so we did not anticipate that system expertise would be particularly variable across subjects. Several studies have addressed how a searcher's domain expertise can affect query formation and the information seeking performance (Jansen, et al, 1998; Hsieh-ye, 1993; Lazonder, 2000). Some of these studies have evaluated this issue by examining the amount of time that users take to construct their queries, indicating that experts are often able to conduct faster searches or formulate search queries more quickly. Furthermore, it is also possible that because novices have a more ambiguous expectation of their search needs and goals, they may often spend more time evaluating the abstracts presented to them, possibly in attempts to learn from the displayed results. Therefore:

H11: Subjects who report to be experts on particular query tasks will take less time to find a relevant document than non-experts.

CHAPTER 3

METHODS

EXPERIMENTAL DESIGN

There are primarily two main types of information retrieval query designs – the interactive query task, and the batch design (Baeza-Yates& Ribeiro-Neto, 1999). Each of these methods highlights and assesses a unique aspect of the information retrieval experience. The batch design is primarily structured to emphasize and evaluate the performance of the search engine, and does not involve a true user evaluation. In contrast, this study used an interactive query design, which enables users to search through a live interface, constructing their own query terms and evaluating the unique results generated through their personal queries. This approach provides for a more robust evaluation of the interface and user, in contrast to solely assessing the document selection quality.

Furthermore, because the eye-tracking was an integral component to this research, a laboratory setting was necessary to capture all aspects of the search session and related eye movements. Although some may argue that external validity is comprised in a laboratory condition, this concern is less of an issue in the context of this research, as the experimental setting is necessary for user behaviors to be systematically compared across all subjects.

Participants

Participants were undergraduate students of various majors at a large university in the Northeast. In total, 36 participants were recruited. Due to recording difficulties and the inability of some subjects to be precisely calibrated, comprehensive eye movement data was recorded for 29 of the subjects. The majority of students were

given extra credit in communication courses for their participation. All subjects were between 18 and 23 years old, with a mean age of 20.3. The gender distribution was split between 19 males and 15 females, and all subjects indicated at least a general familiarity with the Google interface, as 31 of the subjects reported that Google is their primary search engine. Search queries and questionnaire data were recorded and captured for 34 subjects.

Data Capture

Eye-Tracking Apparatus

The subjects' eye movements were recorded using an ASL 504 commercial eye-tracker (Applied Science Technologies, Bedford, MA) which utilizes a CCD camera that employs the Pupil-Center and Corneal-Reflection method to reconstruct a subject's eye position (Figure 3.1). A software application accompanying the system was used for the simultaneous acquisition of the subject's eye movements (Lankford, 2000). Websites were displayed on a 13 inch flat panel monitor with a resolution of 1024 by 768 pixels. The camera used, heretofore referred to as the pan/tilt unit, was an ASL Model 504 unit with the Model 5000 control unit. The pan/tilt unit is a remote



Figure 3.1 Subject with eye tracking apparatus (left); close-up of pan/tilt camera unit.

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, communicated 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 in the instance there is a lot of 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 housed the ASL and head tracker software, while the participant's computer stored the web pages to be viewed during the experiment (Lankford, 2000).

Clickstream

Because Google is continuously updating its search algorithms, it is probable that one specific query will not produce the same exact results on two separate occasions. Because much of the data analysis and relevance judgments were to occur after the experimental sessions, it was necessary to cache the Web pages that the subjects saw instead of simply recording URLs of the pages. To accommodate this, a proxy server was established to log all clickstream data and store all Web content. The proxy script was run on the subjects' 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.

LookZones

In addition to logging the clickstream and Web page data of subjects, the script was also generated to automatically construct “LookZones” around key content regions. The script utilized a feature inherent to the GazeTracker software system which automatically creates LookZones around links and pictures, which the software recognizes within the HTML tags. Thus, the script exploited this feature, and enabled the creation of distinct LookZone regions around each of the ten displayed results (Figure 3.2). Furthermore, through the script, we were able to further distinguish in which part of the abstract a fixation landed, whether it be in the title, snippet, URL, or elsewhere. Thus, for the analysis, each of these displayed results – rank #1, rank #2, rank #3... rank #10 – was given its own set of LookZones, from which we can then compare eye tracking behaviors across all queries, relative to these zones.



Figure 3.2 Sample results page with LookZones.

Experimental Procedure

All participants were required to give informed written consent prior to the start of the experiment. Participants were instructed to search for a total of 10 different queries through the Google interface. All search queries, Web pages, and related activity were recorded and stored on a proxy server, to allow for further post-experiment analysis. A script was generated to automatically eliminate all advertising content, so that the results pages of all subjects would look as uniform as possible, with approximately the same number of results appearing within the first scroll set.

With these pre-experimental controls, subjects were able to participate in a live search session, generating unique search queries and results from the questions and instructions presented to them. Subjects were told to view the webpage and search as they typically would under normal conditions, with the opportunity to scroll up and down the page at their leisure. The experimenter sat to the right of and behind the participant, 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 subjects' eye due to extreme movements, she could re-center and if appropriate, perform a quick recalibration fix.

The ten search tasks were read aloud to the subjects by the experimenter to eliminate unnecessary eye movements away from the computer monitor that could potentially hinder the accuracy of the ocular calibration. To eliminate the potential bias from question order effects, all search questions were completely randomized for all subjects. Similarly, a subset of the questions on the post-questionnaire related directly to the search tasks given in the experiment, specifically to address subjects' perceived difficulty, satisfaction, and expertise with the experimental search tasks. The post-questionnaire questions relating to the specific tasks were presented in the

same random order that the subjects actually searched for the queries, in an attempt to facilitate memory and recall of the search experience.

Analysis and Justification of Search Tasks

Ten search tasks were included in this study, which represent different facets of an information seeking 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. These searches were more ambiguous because the correct answer could potentially be found on multiple sites.

Before selecting the final search tasks to include for the study, tasks were cross-checked with the top searches listed on the “Google Zeitgeist.” The purpose of this 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. Furthermore, the 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 were on average, more difficult than a user’s typical query.

Following is a brief description of the ten search tasks included in the experiment, and a justification and explanation for why these specific queries were included. Again, all subjects received these queries in a completely randomized fashion, to minimize or eliminate any potential biases that could occur through order effects.

Find the homepage of Michael Jordan, the statistician.

Correct Answer: <http://www.cs.berkeley.edu/~jordan/>

Although this query appears to be a straightforward navigational query – find a personal homepage – this task actually has several underlying levels of ambiguity. Most apparent is that the name of this statistician is also name of the famous NBA basketball player, who is also likely to be ranked highly in results. Secondly, the word “statistician” is not written explicitly on Michael Jordan’s homepage. Dr. Michael I. Jordan, to whom this search pertains, is a professor of Statistics and Computer Science at UC Berkeley, and although the word “statistics” is written on his homepage, “statistician” was not. Because many participants included the word “statistician” in their queries, many were unable to find this page as a highly ranked result, making this query rather difficult for most subjects.

Find the page displaying the route map for Greyhound buses.

Correct answer: <http://www.greyhound.com/maps/>

This task was a straightforward navigational query, yet even by typing in the key words included in the question, it was difficult for subjects to immediately recognize the correct page. The correct page had the title of, “Greyhound Lines, Inc.: Terminal Information,” which even when ranked highly on Google, may not have appeared as a direct answer to the search task.

What is the name of the researcher who discovered the first modern antibiotic?

Correct answer: Alexander Fleming

This informational query was included because it incorporated a process of exploration and discovery. Some participants may have gone into the task knowing that penicillin was the first discovered antibiotic, yet other subjects might first have to discover that penicillin was the antibiotic in question. They might then use this

newfound information to better structure a subsequent query for the researcher who discovered it.

Find the homepage for graduate housing at Carnegie Mellon University

Correct answer: <http://www.housing.cmu.edu/graduatehousing/>

Again, although this appears to be a straightforward navigational query, most Google searches did not include the correct link within the top few results. Furthermore, many participants went to the general CMU housing Web page (<http://www.housing.cmu.edu>), and did not scroll down to the bottom of the page, where the only direct link to graduate housing was located (Figure 3.3).



Figure 3.3 “Hidden” link to graduatehousing on housing.cmu.edu. This is the only direct internal link and icon to graduate housing at Carnegie Mellon University, as listed on the page: <http://housing.cmu.edu>.

Where is the tallest mountain in New York located?

Correct Answer: The Adirondacks OR High Peaks Region

Again, this task also addresses a facet of exploration and discovery, in the instance that the searchers do not know that the tallest mountain in New York is Mt. Marcy.

Find the homepage of Emeril – the chef who has a television cooking program.

Correct Answer: <http://www.emerils.com/emirilshome.html>

The Google Zeitgeist included many entertainment-related queries, relating to television, movies, and celebrities, and it was important to include aspects of these popular queries into this particular experiment. Although this was a straightforward navigational search, Google displays relevant-looking titles among the top ten results that are not actually the correct solution to the task.

With the heavy coverage of the democratic presidential primaries, you are excited to cast your vote for a candidate. When can you do so in New York?

Correct Answer: March 2, 2004

The Google Zeitgeist also listed several queries that were specific to current events and news information. Drawing from this (the experiment took place during the time of the Democratic Presidential Primaries), this question is actually one that many voters would need to ask if they intend to vote. Because this is an informational search, the answer was not always explicitly stated in the titles and descriptions of documents. After a first search attempt and finding irrelevant information, several subjects refined their query by including “2004”.

Which actor starred as the main character in the original Time Machine movie?

Correct Answer: Rod Taylor

This question also drew from the relative popularity of entertainment queries on the Google Zeitgeist. Furthermore, this informational query was also selected because there was recently a 2002 remake of “The Time Machine,” linking many of the top-ranked results to the modern remake instead of the 1960 original. Furthermore, there

was also a process of discovery and refinement involved in this task, as subjects might find that the original movie was made in 1960; this date could then be included into subsequent queries to obtain more precise and relevant results.

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?

Correct Answer: Members of Llenroc, the Cornell chapter of the Delta Phi Fraternity live in the mansion.

This was perhaps the most ambiguous query included in the search experiment. This query was meant to represent a question that has developed from casual conversation between two individuals. There is no precise address of the residency given in the statement of the task, which is typical of an issue or question prompted from a natural conversation. After hearing that Mr. Cornell used to live close to campus, it is probable that an individuals' curiosity has been piqued, and he or she may return home to search for more information online.

Find the homepage of the 1000 Acres Dude Ranch

Correct Answer: <http://www.1000acres.com>

Many individuals make travel plans on the Internet, and this search represents a query or question that a traveler may have. Interestingly, this navigational query also involved two levels of vagueness. First, subjects may have spelled the title of this ranch with either “1000” in numerical digits, or with the word “thousand.” The differences in spelling (the actual ranch name is spelled with ‘1000’) produce distinct differences in the rankings produced by Google. Furthermore, although 1000 Acres is listed under the heading of “dude ranch” on all other sites, its homepage does not

include the word “dude,” but only “ranch resort.” Because of this, the searchers who include the word “dude” in their query may actually be at a disadvantage.

CHAPTER 4

ANALYSIS

The following analyses and results are split into two primary sections. The first addresses overall descriptive measures of online search behavior, and the second includes more specialized statistical analyses to address the relationships between search tasks, searchers, and their related eye movements.

GENERAL DESCRIPTIVE MEASURES

Total Time

RQ1 addressed how long, on average, users spend selecting an abstract online. The results indicate that on average, participants took 7.78 seconds to select a document ($SE = .37$). However, the time did vary significantly across the 10 search tasks, from 3-5 seconds for the most straightforward questions, and up to 9-10 seconds for the most difficult and ambiguous tasks. Figure 4.1 presents the total time spent searching across all ten search tasks.

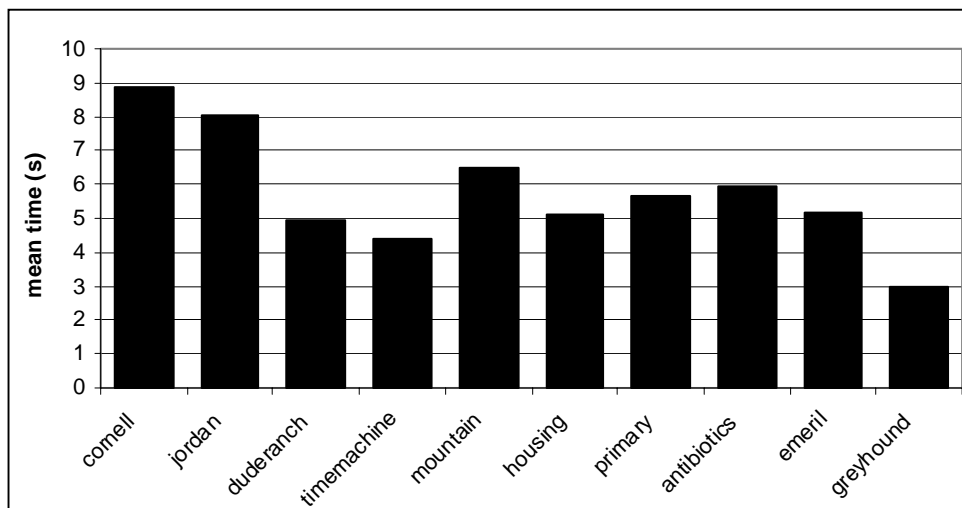


Figure 4.1 Total time spent searching, by search task

Total Abstracts Viewed

As addressed in *RQ2*, the total number of abstracts viewed during a search session can indicate how thoroughly users are evaluating all available information. The eye-tracking results indicated that on average, three abstracts were viewed on each page. It should be noted that because a script removed all non-document content (such as ads), approximately five abstracts fit on the screen at any given time. Results show that searchers rarely viewed all of the abstracts initially visible to them.

Furthermore, with this sort of rank-ordered data, it is often more appropriate to analyze the mode or median response rather than the mean, as this better reflects the behaviors that the majority of subjects are prone to do. Figure 4.3 displays that individuals are most likely to view a total of two abstracts per results-set before making a selection. In just over one-quarter of the cases analyzed, individuals only looked at one abstract on the page before making their selection. From the graph, it also becomes apparent that there is a significant drop in the total number of abstracts viewed after the page break (marked approximately by the fifth-ranked result). Figure 4.3 highlights that users rarely scroll below the first scroll set to evaluate results that are hidden from view at the outset.

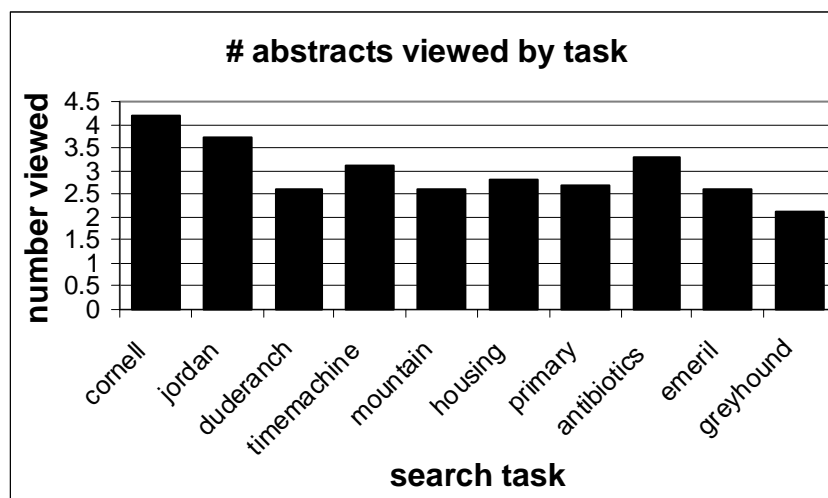


Figure 4.2 Number of abstracts viewed per page, by search task

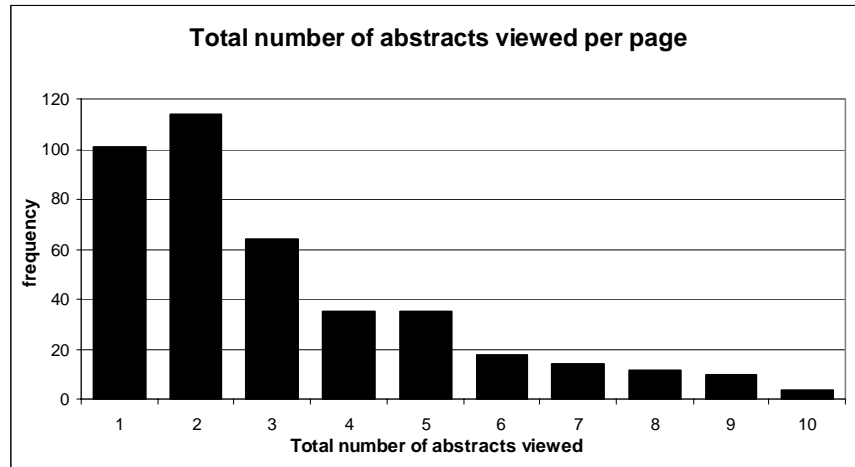


Figure 4.3 Total number of abstracts viewed per page

In what order are the abstracts viewed?

RQ3 addressed whether the linear presentation of the results would also promote users to view the documents linearly. Figure 4.4 is a graph depicting the instance of first arrival to each abstract presented on the screen. The arrival time is measured by fixations; ie, at what fixation did a searcher first view the n^{th} -ranked abstract. From the graph, we can tell that individuals tend to view the first and second-ranked results right away, within the second or third fixation, and there is a big gap before viewing the third-ranked abstract. Furthermore, the page break also manifests itself in this graph, as the instance of arrival to results seven through ten is much higher than the other six, likely due to the fact that they are displayed below the page break, and few searchers scroll below to read the abstracts.

Time spent viewing each abstract

Also of interest is determining how the relative positioning and layout of the abstracts on a page affects a users' subsequent attention to certain items. *H1* postulates that the relative rank of the abstract being viewed would significantly influence the amount of attention and interest it is given, as reflected by pupil dilation.

In other words, are individuals more or less likely to skim over abstracts based on their relative positioning? To determine the impact of placement, the total time spent in each abstract was compared across all subjects (Figure 4.5). Abstracts that users did not view were not included in the analyses.

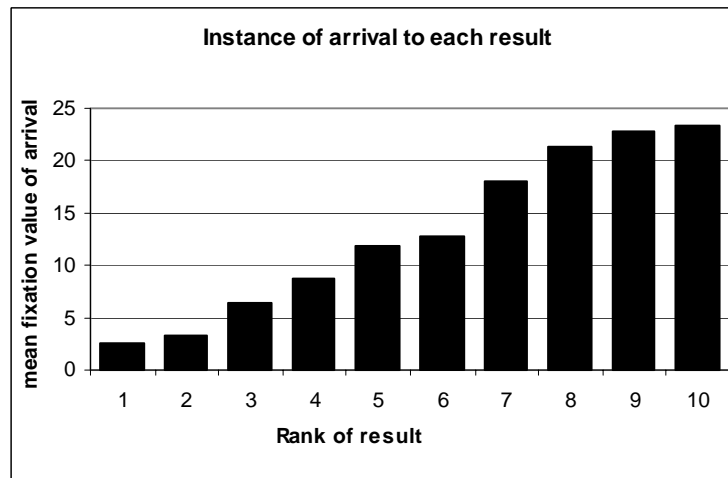


Figure 4.4 Instance of arrival to each result, based on fixation number

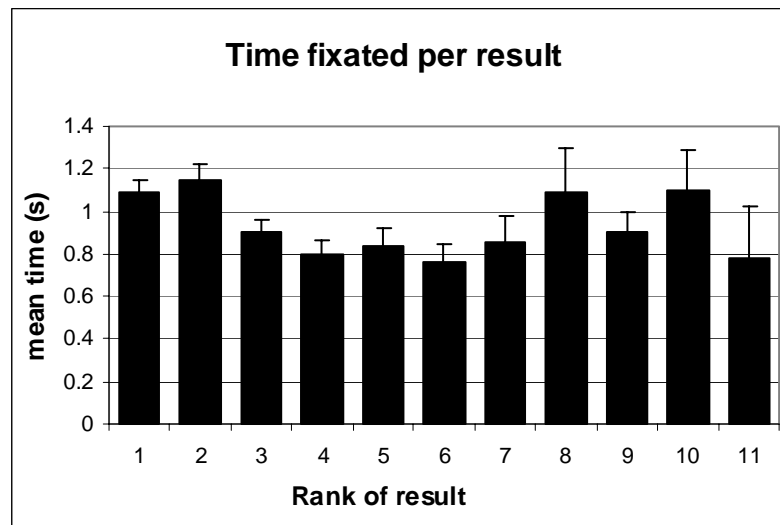


Figure 4.5 Time spent viewing each abstract

Figure 4.5 presents an interesting trend, in that the total time spent assessing the middle-ranked results is lower than abstracts presented in the very beginning or very

end of the results set (ie – the first and last few abstracts). Searchers appear to spend the most time viewing the first and second result, as well as the eighth and tenth result, and are most likely skim or scroll faster through the abstracts presented in the middle of the results set.

How does rank influence the amount of attention a link receives?

A related issue to investigate is to compare the amount of time that searchers spend evaluating and reading each presented abstract with the number of times that a document is selected. This analysis differs from the one previously described, in that it includes all valid cases (regardless of whether or not a particular abstract was viewed) to depict the relative likelihood of an abstract being viewed. Figure 4.6 shows the mean time users spend fixating on a given abstract compared with the number of hits that result generated. It is clear to see that the amount of time searchers spend viewing the first result very closely matches the number of clicks that result receives, while the subsequent abstracts receive much more attention than clicks.

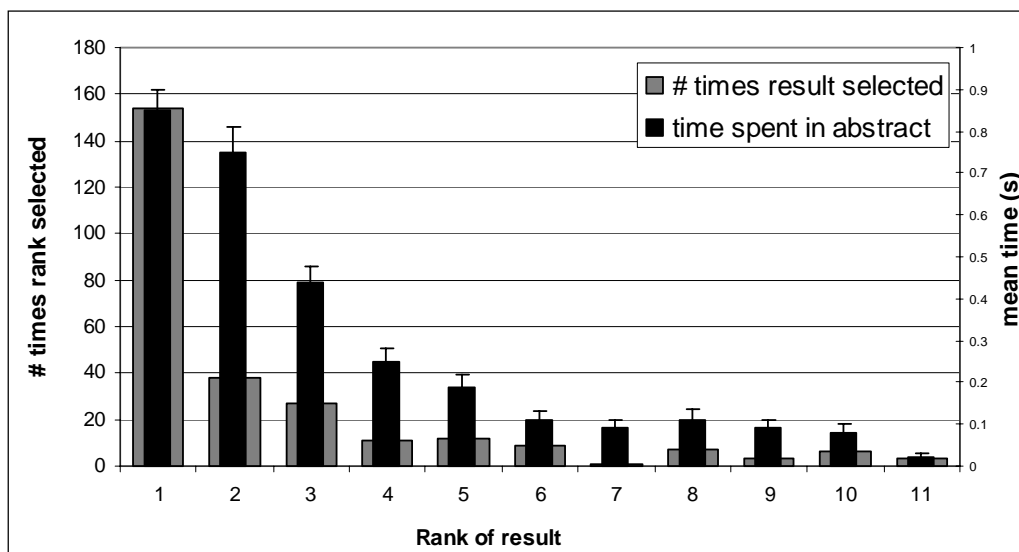


Figure 4.6 Time spent viewing abstracts compared with the frequency of each result being selected

Searchers view the first and second abstracts almost equally, yet click on the top result disproportionately more. Furthermore, beginning with the third-ranked result, fixation time drops off sharply. There is an interesting dip around the sixth-ranked result, both in the viewing time as well as in the number of clicks, as this represents the portion of the results-set not contained in the initial screen-shot. Unlike for ranks two to five, the abstracts ranked six to ten receive approximately equal attention.

How thoroughly is the results-set viewed?

RQ3 and *RQ4* questioned how comprehensively users would evaluate the results presented to them online when searching for a document. For instance, if a user clicks on the third-ranked result, did he or she look at the abstracts ranked one and two? And did the user explore any of the abstracts below their selected link? Figure 4.7 depicts how many results above and below the selected document users scan on average.

Again, this graph displays some interesting effects around and before the page break. First, in only one instance of the 397 cases analyzed did a user actually click on rank seven, which often fell directly below the page break. Secondly, users who

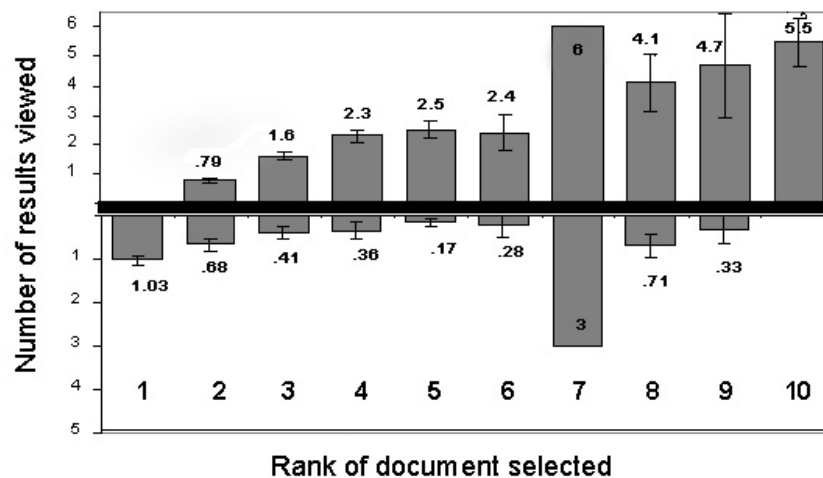


Figure 4.7 Number of abstracts viewed above and below selected document

selected the lower ranked documents viewed proportionately more abstracts overall. Finally, the number of links viewed below a click is low beyond rank 1, indicating that users do tend to scan the list from top to bottom, and select the first promising abstract that they encounter. Comparisons of lower-ranked abstracts are only likely to occur between the first and second-ranked result.

What information in the abstract is most useful?

In addition to investigating which of the presented abstracts were viewed, it is also useful to know which *portion* of the abstract users most rely on to make a relevance judgment. *RQ8* and *RQ9* addressed which portions of the abstract would receive the most attention, as well as which portion users would first view. Attention to each component of the abstract was measured by the number of fixations in each region. Figure 4.8 illustrates that overall, the snippet received the most attention, with the title and URL following close behind.

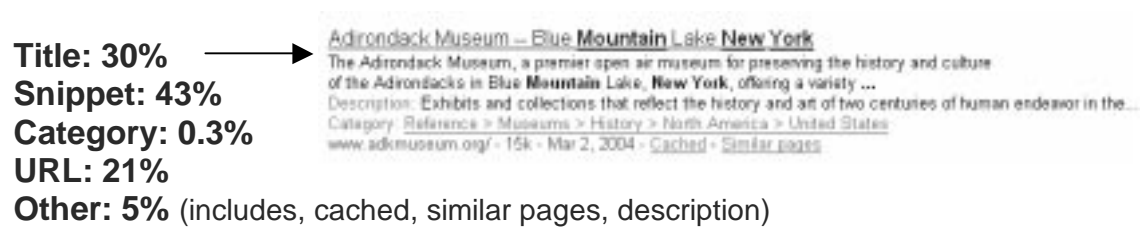


Figure 4.8 Proportion of fixations in each portion of the abstract

However, it should be noted that the snippet is proportionately larger than both the URL and title, and is therefore most likely to capture more fixations. Furthermore, because the information contained in the snippet is in the form of sentences and full-text content, it is presumably more difficult to cognitively parse than either the titles or URL. The titles and URL can be more readily perceived through an efficient visual

scan, while multiple fixations are required to make sense of the information contained in the snippet. Furthermore, the title is in a larger font and contains fewer words than the snippet, and the key portion of the URL is also much shorter in comparison to the snippet, which also make these elements of the abstract more likely to be scanned in fewer fixations.

LINEAR MIXED MODELS

While the previously described descriptive measures present a thorough overview of the trends in the data, a statistical model is desired to fully capture the complexities and interdependencies of the different measures included in the data. To fully account for all random and fixed effects in this experiment, and to generate appropriate estimates of error based on the nested data structure, three- and four-level linear mixed models were used to accurately analyze the data (MuCulloch & Searle, 2001). All dependent variables were analyzed within several three and four level mixed models with SPSS 12.0 using the MIXED command syntax.

For the purposes of this analysis, a linear mixed model offers a number of distinct advantages over the commonly-used general linear model. One of the primary advantages is that a mixed model better handles correlated repeated-measure data and unequal variances. Because mixed models enable the researcher to specifically identify random effects (in the case of this research, task and subject factors), appropriate error estimates and residuals are more accurately calculated independently of each other within a covariance matrix. Adjustments made to the covariance structure in a GLM are not always appropriate because it assumes the independence of the data; in the present research, a number of the variables and factors are highly correlated, and by using the analysis of variance measures (ANOVA) common to the GLM, unbalanced designs are not always properly explained. Although this was a

controlled experiment, some subjects were missing eye movement data for some queries (due to technical and calibration issues), and furthermore, some tasks were more difficult to find and required the searcher to employ more queries to find the solution. Because there were more units of measurement for some factors in the analysis and there is also some missing data, the present research is slightly unbalanced. Linear mixed models compute fixed and random effects in one model by using restricted maximum likelihood (REML); the key advantage of REML is that it accommodates data that are missing at random (Littel, 1996). Because this test uses efficient estimators to compute significance tests, the degrees of freedom associated with the tests often contain decimal values, unlike whole integers with GLM. Finally, it should be noted that the residuals are calculated independently for each random factor; in this case of this research, for the subject and task.

Data Analysis Structure

The experiment was designed so that the data was structured hierarchically, with several levels with nested factors. Figure 4.9 depicts an overview of the experimental structure. Twenty-nine subjects (male and female) were included in the analysis. Each of the 29 subjects searched for ten different tasks, some of which required numerous attempts to find the desired information. These tasks represented two of the three main query types – informational and navigational – and fell into two categories of difficulty (easy and hard). For each of these searches, subjects viewed a number of the presented abstracts on the page, and clicked on an abstract to further explore. Furthermore, all measures of eye fixations were nested within each query and were further classified according to the rank of the abstract that was observed by each fixation. Two dependent variables were also recorded at the fixation level of analysis – fixation duration and pupil dilation. Therefore, in sum: search tasks are nested

within subjects as a between-subjects measure. Each results page is nested within queries, and the abstracts are nested within each results page. The lowest level of analysis is on fixations (including pupil dilation and duration), which occur on the results page, and within the rank-ordered abstracts.

Prior to running the models, descriptive analyses of all the independent and dependent variables were first conducted. The subjects' demographic and background variables, such as the subjects' age, online search, and familiarity with various search engines, were dropped from the analysis because there was little variability in the sample population. Additionally, for the purposes of simplification in the multi-level mixed model, users' difficulty rankings of the ten search tasks were categorized into two groups – easy and hard – based on a median split. Unless otherwise specified, all analyses were done with three-level mixed models, including subject and task as random factors, and task type, task difficulty, and gender as fixed factors. Subject and

	Random Variables	Fixed Variables (with number of levels)
Level 1	Subject (R) ¹	Gender (F) (2 ² ; male, female)
Level 2	Search Tasks (R)	Difficulty (F) (2; easy, hard) Task type (F) (2; informational, navigational)
Level 3	Results Page	Total time (F) (10; 1-10) Total number of fixations (F) (continuous)
Level 4	Rank-ordered Abstracts (1-10)	Individual fixations (F) (continuous) Fixation duration (F) (continuous) Pupil dilation (F) (continuous)

¹(R) and (F) represent random and fixed factors.
² Numbers represent levels within the nested factors

Figure 4.9 Overview of nested data structure

task were used as random factors as they are representative of a random sample drawn from a much larger population of possible users and search tasks.

Task Type and Difficulty

Total Time

Earlier in the chapter, the overall mean time spent for users to select a document was presented. However, as stated in *H3* and *H6*, it was expected that both task difficulty and task type would affect the total time taken to select a document. Total time spent on each page before selecting the first document was modeled as a three-level mixed model. Gender and task attempt – whether it was the subjects’ first, second, third, or fourth try for finding the information – were also included in the model to control for additional variability that these factors may have produced. Note that only the first four task attempts were included in these models for analysis, as this accounted for 99% of the data. One subject did indeed search for a total of eight trials on the same page; however, these outliers were excluded from analysis.

Table 4.1 Total Time: Mixed Model Results

Fixed Factor	Degrees of Freedom	F	Sig. (p)
Task type	1, 257.152	1.090	.297
Task difficulty	1, 255.885	13.307	<.0001
Gender	1, 28.625	.676	.418
Interaction: task type & task difficulty	1, 252.070	3.296	.071
Task attempt	7, 455.723	.256	.970

Table 4.2 Residual Estimates for Total Time

Residuals	Estimate	Std. Error
Overall R	66.963	4.696
Subject Variance	2.752	1.841
Task Variance	4.127	2.784

Results indicated that task difficulty significantly contributed towards total search time ($F = 13.31$, $p < .001$), supporting $H3a$. The estimated marginal means indicate that the hard tasks take users approximately two seconds longer than easier tasks (easy: $\underline{M} = 6.02$, $SE = .54$; hard: $\underline{M} = 8.34$, $SE = .50$). Gender and task attempt were insignificant and did not affect the total time spent searching. Furthermore, task type did not affect the total search time, so $H6$ was not supported. However, there was a marginally significant interaction effect between task difficulty and task type ($F = 3.30$, $p = .071$). The interaction between task type and task difficulty indicated that the easier informational tasks took approximately two seconds longer to accomplish than the easier navigational tasks (Figure 4.10).

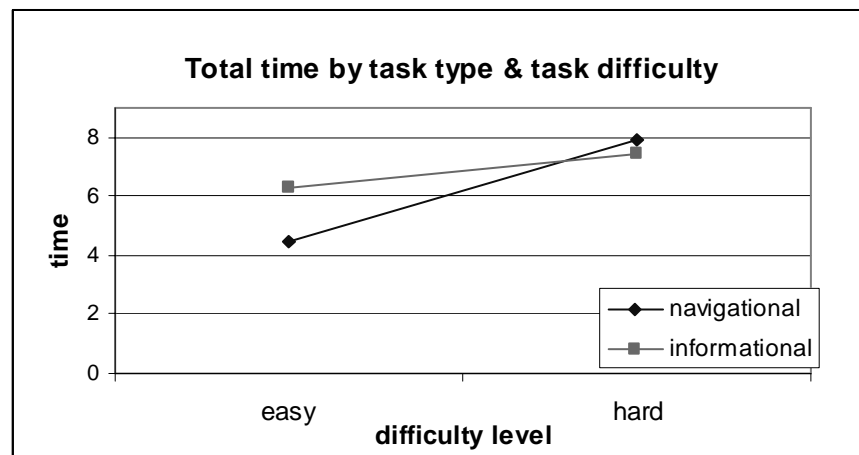


Figure 4.10 Interaction between task type and task difficulty for total time

Number of Abstracts Viewed Below Selected Document

The number of abstracts viewed below the selected document was expected to be greater for informational tasks as well as for more difficult tasks, as predicted through *H3a* and *H8*. Again, gender and task attempt were also included to control for additional variability. The linear mixed model showed that the number of abstracts viewed below the selected document varied significantly by task difficulty, supporting *H4* ($F = 9.14$, $df = 1$, 341.54 , $p = .003$). For “easy” tasks, searchers looked at approximately .84 documents below the selected abstract, while when searching for “hard” tasks, users looked at approximately 1.4 abstracts below. Task type, task attempt, and gender were not significant in this model, and thus *H8*, predicting that task type would have an effect, was not supported ($F = 2.35$, $df = 1$, 336.48 , $p = .126$).

Table 4.3 Number of Documents Viewed Below: Mixed Model Results

Fixed Factors	Degrees of Freedom	F	Sig.
Task type	1, 336.478	2.354	.126
Task difficulty	1, 335.674	9.053	.003
Gender	1, 28.825	.004	.951
Task attempt	6, 346.991	.817	.557

Table 4.4 Residual Estimates for Documents Viewed Below

Residuals	Estimate	Std. Error
Overall R	2.985	4.696
Subject Variance	.265	1.841
Task Variance	--	--

Total Number of Abstracts Viewed

The total number of abstracts viewed was expected to differ on account of task type and task difficulty, with informational tasks and more difficult tasks causing users to view more presented documents. Based on the analysis, the total number of abstracts viewed varied significantly across task type, supporting *H7* ($F = 4.70$, $df = 1, 278.3$, $p = .03$). Users looked at significantly fewer abstracts when engaged in navigational queries than when performing informational queries (navigational: $\underline{M} = 1.58$, $SE = .46$; informational: $\underline{M} = 2.01$, $SE = .48$). Furthermore, *H3b*, predicting that the total abstracts viewed would vary by task difficulty was not supported ($F = .421$, $df = 1, 284.66$, $p = .517$), and neither gender or task attempt significantly affected the model.

Table 4.5 Total Number of Abstracts Viewed: Mixed Model Results

Fixed Factors	Degrees of Freedom	F	Sig.
Task type	1, 278.301	4.698	.031
Task difficulty	1, 284.662	.421	.517
Gender	1, 27.684	.457	.505
Task attempt	7, 471.852	.780	.604

Table 4.6 Residual Estimates for Total Abstracts Viewed

Residuals	Estimate	Std. Error
Overall R	4.358	2.305
Subject Variance	.445	.021
Task Variance	.309	.116

For the two most difficult tasks – the Michael Jordan and Cornell mansion tasks – users viewed approximately 3.7 abstract views per page, while for the easiest tasks – find the routemap for Greyhound buses – searchers only viewed 1.97 abstracts per

page (Figure 4.2). However, these differences were insignificant when compared with task type.

Pupil Dilation

Pupil dilation was analyzed as a four-level mixed model, containing fixed factors of gender, task type, task difficulty, rank of the selected document, and rank of the abstract viewed. It was predicted that pupil dilation would differ on account of task difficulty and task type, with more difficult and informational tasks causing larger pupil dilations.

Based on the mixed model results, both task type and the rank of the abstract viewed emerged as significant, supporting *H2*. Interestingly, task difficulty had no effect on pupil dilation, so *H5*, which predicted that pupil dilation would increase with difficult tasks, was not supported ($F = 1.90$, $df = 1$, 147.64 , $p = .17$). Informational tasks averaged a significantly larger pupil diameter than navigational tasks ($F = 7.34$, $df = 1$, 144.85 , $p = .008$) (navigational: $M = 53.7$, $SE = 1.95$; informational: $M = 54.5$, $SE = 1.95$), and middle ranked abstracts also resulted in a smaller pupil dilation than the abstracts on the periphery ($F = 11.21$, $df = 1$, 2617.75 , $p = .001$)² (Figure 4.11).

Table 4.7 Pupil Dilation: Mixed Model Results

Fixed Factors	Degrees of Freedom	F	Sig.
Gender	1, 27.021	1.955	.173
Task Type	1, 144.849	7.354	.008
Task difficulty	1, 147.644	1.900	.170
Rank of selected doc	9, 174.151	.806	.612
Abstract viewed	1, 2617.75	11.213	.001

² Although the degrees of freedom specified in this model for the rank of the viewed abstract seems extremely high, because the rank of the abstract viewed was measured at the instance of every single fixation (and over 3,500 fixations are included in this data set), the measure is likely to be valid.

Table 4.8 Residual Estimates for Pupil Dilation

Residuals	Estimate	Std. Error
Overall R	3.658	.696
Subject Variance	104.59	1.841
Task Variance	2.30	--

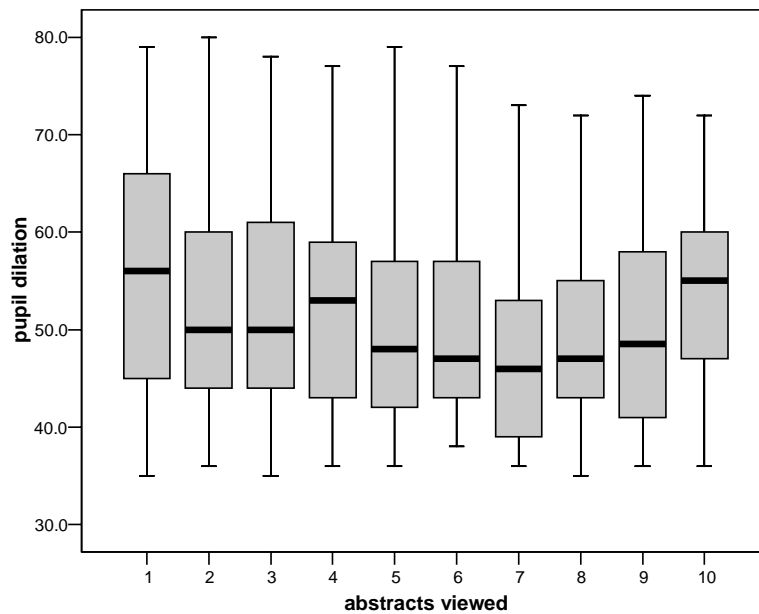


Figure 4.11 Difference in pupil dilation by the rank of the abstract viewed

Finally, one more model for pupil dilation was computed, as one of the key questions relating to pupil dilation is whether fixations to the selected document result in greater pupil dilation than all other documents. To address this, a dichotomous variable was created to indicate whether the abstract viewed in a particular fixation was also the rank of the selected document. Specifically, all values were coded as “1” if the abstract viewed matched the rank of the selected document, and all other values were coded as “zero.” Interestingly, pupil dilation was not significantly greater within the abstracts that were selected, and therefore a more detailed analysis is excluded.

Fixation Duration

With the present data, a three-level model was run to analyze the effects of gender, task type, and task difficulty on fixation duration. Fixation duration was expected to be greater for more difficult informational tasks, and shorter for more difficult navigational tasks. Based on the results, fixation duration emerged as significant for task type, with navigational queries averaging a longer fixation duration than informational queries ($F = 5.0$, $df = 1$, 82.179 , $p = .028$) (navigational: $M = 0.35$, $SE = .006$; informational: $M = .33$, $SE = .006$). Furthermore, there was also a significant interaction effect between task type and task difficulty, supporting *H9a* and *H9b* ($F = 5.287$, $df = 1$, 83.443 , $p = .024$) (Figure 4.12). For each of the task types, the most difficult tasks essentially averaged the same fixation duration (navigational: $\underline{M} = .341$; informational: $\underline{M} = .342$). However, with respect to the easiest tasks in each category, the navigational tasks averaged a much longer fixation duration than did the informational tasks (navigational: $\underline{M} = .36$, $SE = .009$; informational: $\underline{M} = .33$, $SE = .008$).

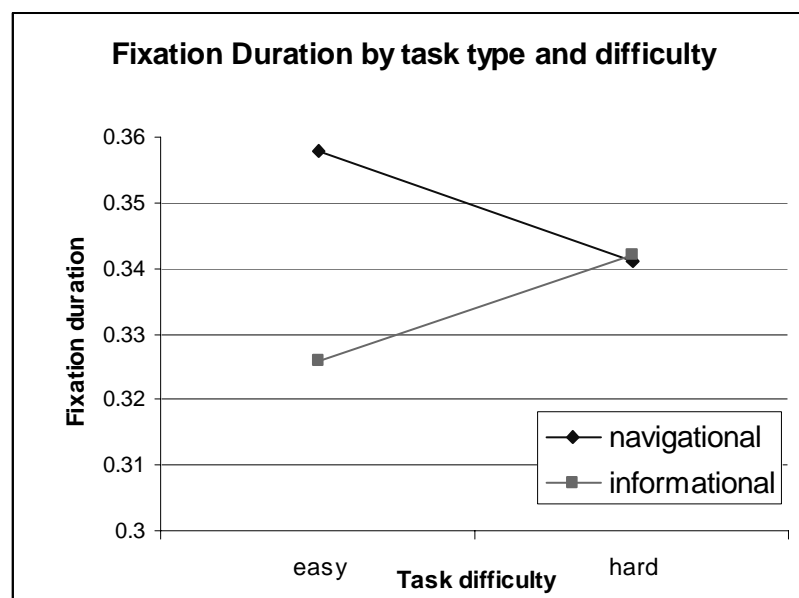


Figure 4.12 Interaction between task type and task difficulty for fixation duration

Table 4.9 Fixation Duration: Mixed Model

Source	Degrees of Freedom	F	Sig.
Gender	1, 21.190	.505	.485
Task type	1, 82.179	4.995	.028
Difficulty	1, 84.985	.013	.910
Interaction with task type & task difficulty	1, 83.443	5.287	.024

Table 4.10 Residual Estimates for Fixation Duration

Residuals	Estimate	Std. Error
Overall R	.0227	.0006
Subject Variance	.0004	.0002
Task Variance	.0006	.0003

GENDER AND SEARCH BEHAVIORS

Number of Abstracts Viewed Above Selected Document

Based on the existing literature, it was expected that gender would influence the overall thoroughness and activity in an online search. One way to determine the comprehensiveness of the search is to evaluate how many abstracts above the selected document were fixated, to assess how frequently users may skip links. Therefore, the number of abstracts viewed above the selected document was analyzed as a three-level model with gender, task type, and task difficulty as the fixed factors, and with subject and task as random factors. The number of abstracts viewed above the selected document was marginally significant for gender, task type, task difficulty, and task attempt. Furthermore, there was a significant interaction between the task difficulty and task attempt. Informational queries generated more looks above the selected document than did navigational queries (navigational: $\underline{M} = .52$ SE = .112;

informational: $\underline{M} = .79$, $SE = .11$), and males viewed more abstracts above their selected document than did females ($F = 3.90$, $df = 1, 28.70$, $p = .058$) (males: $\underline{M} = .81$, $SE = .11$; females: $\underline{M} = .50$, $SE = .13$).

Table 4.11 Number of Abstracts Viewed Above

Fixed Factors	Degrees of Freedom	F	Sig.
Task Type	1, 333.508	2.934	.088
Task Difficulty	1, 338.409	3.154	.077
Gender	1, 28.695	3.896	.058
Task attempt	3, 343.276	2.226	.085
Interaction between task difficulty & attempt	3, 340.160	3.194	.024

Table 4.12 Residual Estimates for Abstracts Viewed Above

Residuals	Estimate	Std. Error
Overall R	1.776	.140
Subject Variance	.0623	.055
Task Variance	--	--

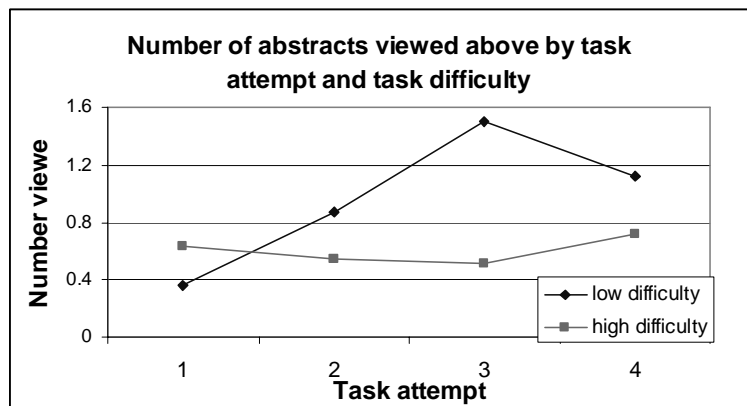


Figure 4.13 Interaction between task attempt and task difficulty for the number of abstracts viewed above

Rank of Selected Document

Another issue to assess by gender was the rank of the selected abstract. Linear mixed models are built on the assumption that the dependent variable is normally distributed. Although the rank of selected document follows a slightly non-normal distribution, a cursory analysis can still be performed with a mixed model, using the same task, subject, and task attempt as random factors. Based on the analyses, it becomes clear that the rank of the selected document differs significantly by subjects' gender and their search attempt. Based on estimated marginal means, males were significantly more likely to select a lower ranked document than females ($F = 6.17$, $df = 1, 27.97$, $p = .019$) (females: $\underline{M} = 3.062$, $SE = .506$; males: $\underline{M} = 4.065$, $SE = .491$).

Table 4.13 Rank of the Selected Document: Mixed Model Statistics

Fixed Factors	Degrees of Freedom	F	Sig.
Gender	1, 27.972	6.171	.019
Difficulty	1, 166.794	1.045	.308
Task type	1, 164.818	1.416	.236
Task attempt	4, 152.159	3.505	.009

Table 4.14 Residual Estimates for Rank of Selected Document

Residuals	Estimate	Std. Error
Overall R	4.6362279	.1128133
Subject Variance	.4701265	.3024346
Task Variance	1.4309497	.7722197
Clickset variance	4.2039916	.7374480

Because neither task difficulty nor task type emerged as significant in this analysis, another model was created using task as a fixed factor instead of a random factor. The purpose of this model was to determine whether there is significant variability between the individual tasks users were given. The results from this analysis are thus not as generalizable to a larger population; however, understanding how each of the different tasks impacts document selection is an intriguing question to address in the context of the present research. Based on the analyses, gender still emerged as significant, while the individual tasks were only marginally significant, indicating that there is a only small degree of random variability between the tasks themselves that cannot be attributed to their type or difficulty.

Table 4.15 Rank of selected document, by individual task

“Fixed Factors”	Denominator df	F	Sig.
Gender	1, 28.187	5.050	.033
Task ID	9, 207.595	1.890	.055

In sum, Table 4.16 presents the main factors of interest – task type, task difficulty, and gender – and highlights the key findings for each variable.

Table 4.16 Overall Significant Effects for Task and Gender

Fixed Factor	Significant Effects		
Task Type	Total abstracts viewed	Fixation duration	Pupil Dilation
Task Difficulty	Total Time	N abstracts viewed below	
Task type * Task difficulty	Total Time	Fixation duration	N abstracts viewed above
Gender	Rank of selected abstract	N abstracts viewed above	

SURVEY MEASURES

An online questionnaire was administered to all participants after the completion of the search session. The questionnaire included items to address the general frequency with which participants use online search engines, as well as items to specifically evaluate performance on each search task. These measures addressed subjects' expertise, satisfaction, and perceived difficulty in accomplishing the experimental search tasks.

Online Search Usage

Subjects were asked to rate, on a one (not at all) to ten (all the time) scale, the frequency with which they search for the following topics: research, employment, entertainment, news, directions, shopping. Table 4.17 displays the mean and standard deviations for each type of search.

Table 4.17 Frequency with which users search for selected information

Search Task	Mean	Std. Dev.
<i>Research</i>	7.74	1.81
<i>Employment</i>	4.62	2.74
<i>Entertainment</i>	6.56	2.36
<i>Directions</i>	6.35	2.88
<i>News</i>	6.88	2.69
<i>Shopping</i>	5.59	2.32
<i>N= 34</i>		

Table 4.18 presents rankings of users' information retrieval behavior and how they interact with search engines. The interesting findings that emerged from these questions are primarily that users are very unlikely to change their search engine if they are unsuccessful, but rather searchers would rather reformulate their query on the

search engine they are most familiar with. Study participants also ranked Google's credibility quite high.

Table 4.18 Self-reported characteristics of search behavior

Search Behavior	Mean	Std. Dev.
<i>Trust in Google's retrieved results</i>	7.71	1.38
<i>Change query if not successful</i>	9.38	1.18
<i>Change search engine if not successful</i>	2.91	2.33
<i>Google offers good description</i>	7.88	1.67
<i>N = 34</i>		

Furthermore, Table 4.19 also presents users' assessments of the expertise, difficulty, and satisfaction with each of the search tasks. From the chart, it is clear that for the majority of the queries, there is little variability for search expertise, while there are more apparent differences for difficulty and satisfaction. Because there was little variability for expertise, it was not included in subsequent analyses due to the little effect it would have.

How does search task difficulty affect satisfaction with the search engine?

Finally, correlations were drawn between satisfaction, difficulty, and expertise. Based on Figure 4.9, it is clear that search task difficulty is negatively correlated with a searcher's satisfaction with the search engine performance. For the most difficult queries (the Michael Jordan and Cornell mansion tasks), satisfaction ratings are the lowest, while for the easiest navigational tasks, the satisfaction is much higher.

Table 4.19 Expertise, Difficulty, and Satisfaction for search tasks

Search Task	Expertise		Difficulty		Satisfaction	
	Mean	SD	Mean	SD	Mean	SD
<i>1,000 Acres Dude Ranch page</i> (N=33)	1.00	0.00	6.06	2.78	6.18	2.93
<i>Emeril homepage</i> (N=35)	2.67	2.78	3.26	2.72	8.23	2.02
<i>Tallest Mountain in NY</i> (N=33)	1.82	1.72	4.18	3.16	7.79	2.76
<i>CMU Graduate Housing</i> (N=34)	1.85	1.84	4.15	2.60	7.00	2.71
<i>Original Time Machine actor</i> (N=35)	2.11	2.26	4.94	2.90	5.97	3.08
<i>NY Presidential Primary date</i> (N=35)	3.49	3.33	4.06	2.89	7.06	2.89
<i>Former residency of Cornell</i> (N=35)	2.77	3.13	8.06	2.53	3.77	2.78
<i>Researcher of first antibiotic</i> (N=35)	3.91	3.23	3.89	3.03	7.57	2.50
<i>Page for Greyhound routemap</i> (N=35)	4.49	3.26	3.06	2.50	8.37	2.14
<i>Michael Jordan, statistician</i> (N=34)	1.09	0.51	7.97	2.49	4.68	3.10

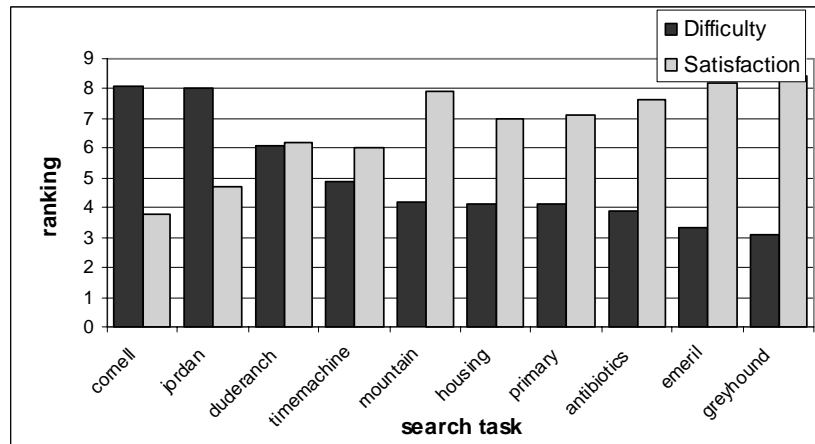


Figure 4.14 Search task difficulty and satisfaction.

Number of Queries per Task

The number of query attempts per search task was also analyzed. Query attempts from all subjects were aggregated for each search task, and are presented in Figure 4.11 from most difficult to least difficult. From this table, it is interesting to note that even though the Cornell mansion task was the most difficult, the total number of queries that users submitted to find the answer was much less than for the Michael Jordan task. The Michael Jordan task received many more query reformulations than any of the other search tasks.

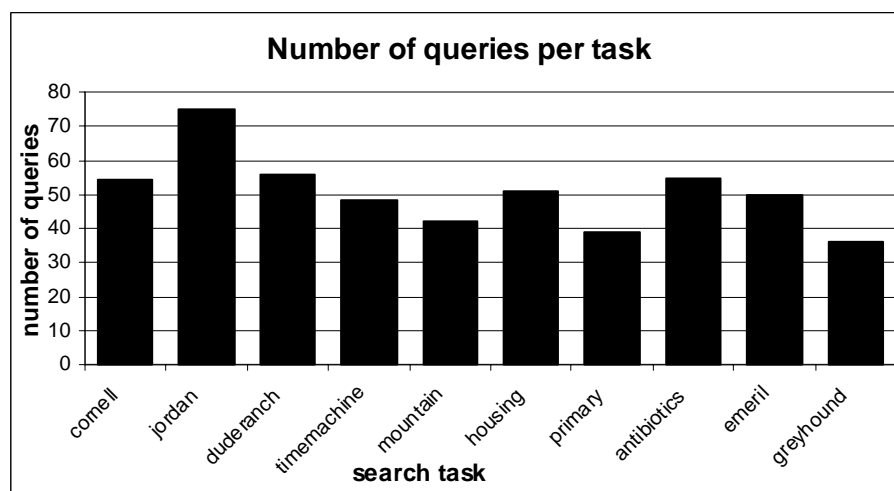


Figure 4.15 Number of queries formulated per search task.

CHAPTER 5

DISCUSSION

The following section will provide a review of the research questions and hypotheses introduced in the Chapter 3 of this thesis. The review will also include a discussion and explanation indicating which hypotheses were and were not supported.

DESCRIPTIVE MEASURES OF SEARCH BEHAVIOR

This thesis addressed several hypotheses and research questions related to the overall behaviors that users employ when searching for information online. These issues included the total time spent searching, the total number of abstracts viewed, the order in which results are viewed, among others.

Viewing Order

RQ7 addressed whether users view documents in a linear order. Based on the arrival time to each abstract on the results page, it appears that users do indeed view the abstracts from one to ten; however, arrival times to each abstract appear to follow a non-linear curve, specifically in the form of exponential decay. Users view the second abstract right after the first, yet the probability that a user will fixate on an abstract ranked lower than two drops sharply. Furthermore, users seem to dislike scrolling through results, and the instances of arrival to documents presented below the page fold are viewed much later than the abstracts on the first screen shot. Typically only the first five or six links were visible to the user without scrolling. Finally, because the instances of arrival to the documents listed below the page fold do not significantly vary, it seems that once a user makes the effort to scroll and view items below the page fold, the influence of rank has less of an impact on the document

selection. A sharp drop in viewing time occurs after the 10th link, as ten results are displayed per page, and few searchers venture beyond the top ten.

Time Spent in Each Abstract

The total time spent in each abstract peaked for the first and second results, decreased for the middle results, then increased slightly for the documents towards the end of the results set. This finding offered evidence in support of *H8*, which predicted that the time spent in the middle abstracts would be lower than those on the periphery. Essentially, the time spent in each abstract provides an approximate indication of how much attention users are giving to the selected content. Furthermore, pupil dilation was also smaller for these middle-ranked abstracts, also indicating less attention and cognitive processing. Based on the results of pupil dilation and time per abstract, it is very probable that the presentation of results affects and causes people to view the abstracts in this manner. It is quite possible that when the screen display is fixed – ie, when the user is unable to scroll up at the beginning of the results set, or to scroll down at the very end of the results set, they subconsciously take more time to process the information. However, when users scroll through the page, they may be moving more rapidly through the middle abstracts, and thus spend less time evaluating them.

This finding also correlates highly with previous literature, in that the ranks of the document most likely to be selected also follow a similar trend, peaking in the beginning and the end of the results set.

Click Rate versus View Rate

Another interesting finding resulted from a comparison of the average time spent viewing a document and the subsequent number of clicks that abstract received. Figure 4.12 in the analyses depicted that although the time spent evaluating the first and second result were relatively equal, the first result was clicked proportionately

more. Several possible explanations are plausible for this. The first is simply that the search engine is more accurate, and Google is indeed presenting the best result first. Alternatively, users may just place a great deal of trust in Google, and even after a relevance comparison, believe that the first result is ideal, because “Google says so.”

To further address the proportion of clicks to gaze time in each abstract, Figure 5.1 was generated to depict any apparent trends. From the graph it follows that the first and 11th result – both listed at the top of the page, generate proportionately more clicks when compared with gaze time. It seems that there is much random variation for the remaining results, though the proportion for the sixth-ranked abstract is higher than the others. A potential explanation for this is that once users scroll down the page, the sixth-ranked result is then presented at the top of the page.

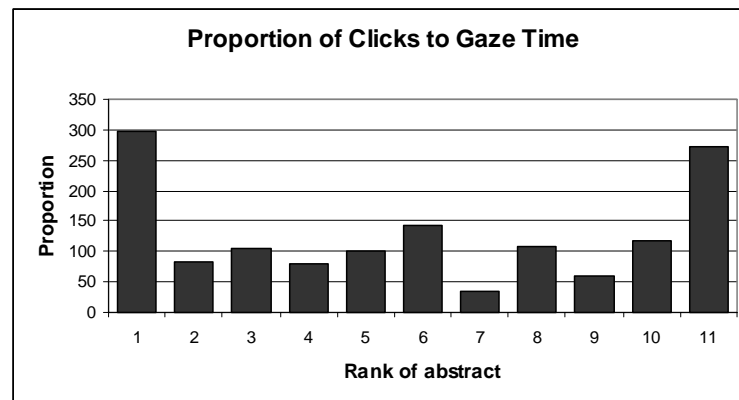


Figure 5.1 Proportion of Clicks to Gaze time, per abstract

Number of Queries per Task

When aggregated across all subjects, the number of queries formulated per task averaged 50.6. Intuitively, the more difficulty tasks required more queries, yet one query, *find the homepage of Michael Jordan, the statistician*, generated 20 more queries than the next highest task, even though it was not ranked as the most difficult query (the Cornell mansion query was most difficult). Why did this task require users

to reformulate their query so many more times? One plausible explanation is that because finding the homepage of Michael Jordan the statistician is a navigational query, users are able to develop a clearer perception of what acceptable or irrelevant abstracts could look like. If users can quickly and easily recognize this within the results-set, less time is required to evaluate each abstract, and thus, if no relevant-looking abstracts are readily perceived, users will then reformulate their query.

TASK MEASURES

Task Difficulty

Total time

It was predicted that task difficulty would influence three primary indices: the total time spent searching, the number of abstracts viewed below the selected document, and pupil dilation. Task difficulty significantly influenced the former two factors, but did not significantly influence pupil dilation. The significance of task difficulty on total time was to be expected, as when tasks are difficult, the search engine performance may not be as precise, and users may need to more thoroughly evaluate the results set. In the process of more carefully evaluating the abstracts, the searcher may also potentially learn from and incorporate the displayed information into subsequent queries. The process of selecting a document for the difficult queries will thus take more time.

Additionally, the mixed model also indicated that task type interacted with task difficulty to significantly influence total time. This relationship indicated that the easy navigational tasks took less time than the easy informational tasks. The interaction is particularly significant as it indicates that even though tasks of both types fell into the low difficulty category, the easy navigational tasks were potentially much more straightforward than the easier informational tasks. Specifically, navigational tasks

may take less time because users have a cognitive representation of what an abstract for their desired homepage should look like. These cues can be perceived much more quickly than from a more ambiguous informational query, where more key word scanning and reading within the snippet is required.

Finally, it is interesting to note that the total time taken to select a document does not vary across multiple search attempts, indicating that users may employ relatively similar strategies (at least strategies that are similarly time-consuming) whether it their first, second, or third try for finding the desired information.

Number of abstracts viewed below

H2 anticipated that the number of documents viewed below the selected abstract would differ by task difficulty. This hypothesis was supported, and indicates that users may choose to more critically evaluate alternative documents when faced with a difficult query. Based on the data from this experiment, it seems that when confronted with a difficult query, users expend more effort towards making an effective selection by making choice comparisons with an adjacent abstract. *H3* anticipated that the number of documents viewed below the selected abstract would differ by task type. This hypothesis was not supported as there were no significant differences indicating that the number of documents viewed below is greater for informational queries. Although this variable did not differ significantly by task type, task difficulty did play a role. More difficult tasks caused users to evaluate significantly more abstracts below the selected document.

Task Type

Total Abstracts Viewed

The total number of abstracts viewed differed by task type, therefore supporting *H2*. *H2* postulated that users will look at fewer abstracts per results set for the navigational queries than for informational queries. This is a particularly interesting finding, as it indicates that tasks which have the potential to be found through more than one source cause the searcher to evaluate more of the available options. In looking over more documents, the searcher may be cognitively constructing internal relevance comparisons between the viewed abstracts. Conversely, when only one result is appropriate, such as when a user is engaged in a navigational query, fewer alternatives are evaluated. It is likely that the decision-making process involved with navigational queries is less complex, in that the user primarily needs to make a dichotomous decision – “yes, this looks like the right page,” or “no.” In informational queries, searchers may attempt to rank and compare the potential relevance of the displayed abstracts, as more than one document could be considered useful.

Furthermore, from the descriptive statistics, it became clear that users are most likely to evaluate two abstracts per page. Combined with the supplementary statistical analysis, one can begin to speculate plausible explanations as to why individuals primarily look at only two abstracts displayed on the results page. Based on the decision-making literature, it is quite plausible that an individual chooses to make a comparison between two documents before selecting one, most likely to ensure that their desired document is indeed the most suitable. Thus, it is possible that searchers are basing their decision on whether their desired document out-performs only one other presented abstract. Based on the ranks of the abstracts that are viewed, it appears that the two abstracts most likely to be viewed are indeed the first and second,

but it would be worthwhile to complete a follow-up test to determine whether the two documents viewed in a decision-making are indeed adjacent documents.

Finally, while the total number of viewed abstracts differed significantly across task types, the total time spent searching did not differ, potentially indicating that when performing an informational query, users may scan the results set more quickly. (Interestingly, this speculation is also supported by the present research findings, as the average fixation duration, which is discussed later in this chapter, is shorter for informational tasks.)

Number of abstracts viewed above selected document

Although no specific hypotheses related to this question, it was found that a number of factors marginally impact a users' inclination to view abstracts above their selected document. The most significant finding related to gender, in that males viewed more abstracts above their selected document than did females. This result actually replicates the studies of gender and online search that were previously discussed in the introduction. These studies indicate that there are distinct gender differences in online search, with school-age males being much more active in scanning and selecting documents than females. Furthermore, there was also a significant interaction between task difficulty and task attempt. In instances in which the search task was difficult, user behaviors primarily remained constant across all trials. However, when the search task was rated as easier to accomplish, subsequent search attempts caused users to view more abstracts per trial. It is very interesting to recognize that searchers changed their strategy in subsequent trials when the search task was less difficult. It is possible that users entered the query thinking that it would be relatively simple to accomplish, yet after realizing that it may not be as

straightforward as anticipated, they change their retrieval strategy to involve more careful selection.

Rank of selected document

Both gender and the task attempt significantly influenced the rank of the selected document. Intuitively, subsequent task attempts resulted in the selection of a lower ranked document. Furthermore, the impact of gender can also be related to the difference in the abstracts viewed above. Males tend to view more abstracts above their selected document, and perhaps this more thorough and linear evaluation of the results set causes them to select a lower ranked document. Finally, from the mixed models, it should be pointed out that while gender significantly impacts the rank of the document selected, it does not at all affect the total time involved in making a selection. This potentially indicates that males are able to scan the results set faster to have the opportunity to view and evaluate the lower-ranked abstracts.

Pupil Dilation

H5 predicted that pupil dilation would increase as an effect of task difficulty, and interestingly, it was not supported. However, pupil dilation did significantly differ by task type, with informational tasks producing greater dilations than navigational tasks. This result offers behavioral evidence in support of the assumption that the cognitive effort involved in informational tasks is significantly higher than that of a navigational task. Informational tasks likely require the searcher to more carefully read through the snippet to detect the context in which their query terms are used.

Also related to pupil dilation, *H9* predicted that pupil dilation would be lower for abstracts ranked in the middle of the results set. Again, this was supported with the three-level mixed model, indicating that pupil dilation was greatest for the first few

and last few abstracts. Again, this likely relates to the time spent in each abstract, as well as a searcher's scrolling behavior. If a user is more rapidly scrolling through the middle of the results set, their reduced attention to the middle abstracts will be reflected in changes of pupil dilation.

Fixation Duration

Fixation duration differed significantly by task type, with informational tasks resulting in shorter fixations than navigational tasks. This finding is corroborated by existing literature suggesting that fixation duration decreases as task difficulty increases. It is likely that informational tasks require more careful processing, and thus more rapid and shorter fixations are required. Furthermore, although in itself task difficulty was not significant, there was a significant interaction between task difficulty and task type on fixation duration, partially supporting *H6*. The easier informational tasks generated much shorter fixations than the easy navigational tasks, while the fixation duration for difficult tasks in both groups was very similar. Even though an informational task may be classified as "easy," it is likely that a great deal of cognitive processing and decision-making still needs to occur, proportionately more so than in "easy" navigational tasks. To further understand these differences, it will be important to generate a new model based on data that contains fixations of a smaller duration than 200 milliseconds.

SUBJECT VARIABLES

Satisfaction with search engine

Understandably, there was a negative correlation between searcher satisfaction and task difficulty. Interestingly, there were four queries for which the satisfaction level varied greatly, while the difficulty did not. The difference in satisfaction levels for

these queries is likely to be attributed to searcher expectations regarding the task. If the user begins his or her search with the impression that their search goal will be accomplished relatively easily, they are likely to be much less satisfied with the search engine if the information is difficult to find. If user enters the search with lower expectations of success, they are likely to become much more satisfied than if a query of the same difficulty took more effort to find.

Expertise

The present search tasks presented little variability in expertise between the tasks, and expertise was therefore excluded from the linear mixed model analysis. Therefore, *H10* was not addressed. However, it should be pointed out that the majority of the subjects rated their expertise as rather low on the majority of the queries, with the exception of the query relating to current events (the date of the NY primary), and the Greyhound bus query (likely due to the fact that college students often need to take the bus!) Some simple correlations were conducted to indicate that several of the “easy” queries were those in which users considered themselves to be more expert. Furthermore, search satisfaction negatively correlated with difficulty.

Gender Correlation with Field of Study

Because significant gender effects emerged for two of the test models, it was important to ensure that there were no distinct demographic differences to influence the findings. Thus, based on the survey data that was collected after each subjects’ search session, several multivariate analyses of variance were performed to test for differences in search behavior by gender. This analysis showed that there were no significant differences between males and females with respect to the types of content

most frequently searched for online, attitudes towards online search engines, and the name of the search engine most frequently used.

Furthermore, of all the search tasks, only one – “What is the name of the actor who starred as the main character in the original Time Machine movie?” – revealed significant gender differences in favor of the male searchers. There were no significant differences with respect to expertise, satisfaction, and difficulty for the nine other tasks. Male searchers had significantly higher levels of expertise and satisfaction for this query, and also found it to be less difficult than did females.

Table 5.1 Significance test for Gender on the Time Machine movie query

Rankings for Time Machine actor query, by gender			
	<i>Df</i>	<i>F</i>	<i>Sig.</i>
<i>Expertise</i>	1	4.396	.044
<i>Difficulty</i>	1	5.935	.021
<i>Satisfaction</i>	1	6.992	.013

Interestingly, gender was also significant when analyzing one other factor – undergraduate major. The majority of the subjects fell into two distinct majors – Information Science and Communication, and the remaining subjects fell into other social science majors (Human Ecology and Psychology) and other science/engineering majors (Computer Science, Biology, and Economics). There was a significant effect between gender and the type of undergraduate major, with females being significantly more likely to major in Communication or the Social Sciences. To further test for experimental effects due to undergraduate major, all majors were classified into three groups: Social Science, Information and Computer Science, and Other (including Music, Economics, and Biology), and it proved to have a significant effect ($F = 5.91$, p

= .007). It would be worthwhile to include undergraduate major in models for further analysis to determine whether a students' course of study impacts their ability to perform in online search tasks.

Table 5.2 Gender differences within undergraduate major

<i>Department ID</i>	<i>N</i>	<i>Mean</i>	<i>Std. Error</i>
<i>Communication</i>	15	.733	.114
<i>Information Science</i>	11	.182	.133
<i>Other</i>	8	.250	.156

CHAPTER 6

CONCLUSIONS

This chapter will draw from the analysis and discussion to generate significant conclusions that can be applied to contexts related to information search.

Information Visualization

The results from this study indicate that our viewing is indeed strongly influenced by the display of the retrieved results. Understanding how searchers attempt to find information online can ultimately lead to insights in the design and structure of online search interfaces and systems. For instance, because the difficulty and type of search task performed (and potentially the level of expertise a searcher brings to the context) all induce significantly different user viewing behaviors, interfaces can be structured to exploit or accommodate these behaviors. Having eye-tracked individuals as they search for information, the results from this study provide an overview of what users' natural viewing behaviors are likely to be. Ultimately, designers should consider these findings so that future interfaces can potentially exploit these natural behaviors to maximize their usability and effectiveness.

A number of researchers have already explored and manipulated the display of the retrieved results in an information retrieval interface (Shneiderman, 2000; Pirolli, et al, 2001; Veerasamy & Belkin, 1996). The majority of these visualization tools are built on the premise of offering more data to the user, in the form of unique data clusters or a hierarchically distributed tree structure. Shneiderman (2000) has advocated a two-dimensional visualization display which arranges the data points along two meaningful axes, such as publication date or year as the x-axis, and categorical data,

such as journal name, as the y-axis. He has also included “sliders” and other tools that enable the user to interact with and mildly change the display of information.

Additionally, Pirolli, Card, and Van DerWege (2001) have developed what they call the “Hyperbolic Tree Browser” to show a graphical tree structure of the relationship between documents and their categories.

The primary intention of researchers when developing these visualizations is to “amplify human cognition,” and to “reduce the cost structure of information” (Pirolli, et al, 2001; Card, et al, 1999). However, there is little evidence to suggest that in an information search context, the user benefits by viewing a high-level overview of all available information. In the context of daily Web search, the goal of providing a high-level overview seems rather lofty and unnecessary. No more than 50% of Web queries are classified as informational, and it is for informational queries that these visualizations are most effective. For simple navigational queries, when there is only one correct answer, there is little need for users to visualize a structural overview of all irrelevant abstracts. However, if overview displays are desired, in these instances, a better alternative would be to provide a link structure, highlighting which sites link to and link from a selected abstract.

Furthermore, Veerasamy and Belkin (1996) evaluated users with both an enhanced visualization interface and a traditional retrieval interface, and found that the advanced visualizations produced significantly different user behaviors for only one task. This is just one reported instance to indicate that users do not always appreciate and benefit from the advanced visualizations in a search and retrieval interface.

In the case of developing advanced visualizations for information retrieval, it may be more helpful to step away from the perspective of technologically-centered design (ie – *we have the ability to make cool and interactive visualizations, so let's do*

it!), and instead think about online information search in the context of user-centered design or even activity centered design (Gay & Hembrooke, 2004).

In evaluating the visualizations that have currently been created, it does not seem that designers have really addressed the users' key tasks and purposes. They seem to assume that the user first wants a hierarchical overview of the available information, and secondly that they will take the time to critically evaluate it and use it to inform their decisions. However, based on this search experiment, users click on documents in less than five seconds, and it is likely that they appreciate the fast and efficient nature of an online search engine, particularly its display of comprehensible results.

Thus, in this case it may be helpful to look at online search in the context of activity-centered design – ie, what is the activity that the user will be performing, and what is the best way to accommodate these needs? For instance, when individuals are searching for information, they generally exhibit the same preferences and desires: the need for fast and useful information. The results from this study indicate that users do not like to scroll or spend much time evaluating their results set, and because of this, it is unlikely that users are not going to appreciate the advanced options and information overview that a visualization display will offer.

However, this is not to say that alternate non-linear displays are ineffective; this is only to say that they are not appropriate for all search tasks. Enabling a searcher to learn from the presentation of displayed results would be an obvious advantage in informational tasks, particularly in specialized digital library environments. Potential effective displays can show a cluster diagram of search results, informing the user of the relationships within and between information in the results set.

Interpreting relevance from multiple clicks

One of the potential problems in the interpretation of clickthrough behavior is to determine satisfaction-related relevancy measures for both the selected document and the entire results set. This is a particularly relevant question in cases when a searcher clicks on multiple links in the answer set. For instance, a searcher may click on the top ranked result, then return to the results page, and click on abstracts 3, 5, and 6. Has this searcher clicked on many of the retrieved results because they are all relevant and useful for his or her task? Or, is the searcher resorting to multiple documents to find the desired information due to his or her dissatisfaction with the first selection?

Both of these scenarios are indeed plausible, but the answer cannot be generated through clickthrough data alone. From the eye-tracking data, it appears that users rather reformulate their query than scroll through multiple results, and furthermore, in 38% of the cases, a user did not select an abstract at all.³

In the case of this experiment, the users continued to search until they either found the answer to the task or the allotted time per query had expired. Thus, in this experiment, all of the repeat clicks per page were due to a users' attempt to find a better result – *not* for the purpose of clicking on and finding additional relevant information. However, based on the low likelihood of multiple clicks per page, this experiment seems to indicate that if a user does not find what they need after one click, they are more likely to reformulate their query for a new results set. Therefore, in instances in which clickthrough data exhibit multiple clicks, it may be more likely that the results are all relevant to the user.

³ These cases are representative of instances in which a searcher: 1) reformulated the query without selecting a document; 2) mistyped a term and clicked on Google's spelling correction (thus not selecting an abstract); and 3), the search session was terminated by the experimenter due to time.

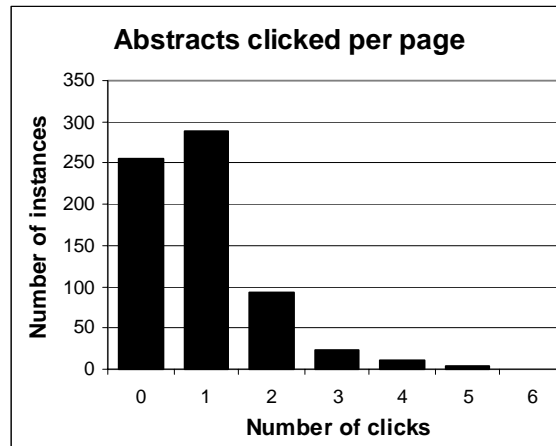


Figure 6.2 Number of clicks per page

General Trends

Popularity of specialized, vertical portals

One of the interesting issues discovered when going through the experimental process was the degree to which users relied on specialized, vertical portals to access their desired information. For instance, when searchers were told to “find the actor who starred in the original Time Machine movie,” many of them asked if they could just use <http://www.imdb.com> (the Internet Movie Database) instead of searching through Google. When reminded that the only requirement was that the Google interface was used, the users then typed “imdb” into the search engine, only to click on the top-ranked result – the Internet Movie Database – a vast source of information about movies and actors.

In these two instances, users felt that the specialized portals would be either a more efficient and/or comfortable means to discover their desired information. Although individuals tend to use search engines for general-purpose information, this indicates that for specific bits of information, users may be more comfortable using a specialized portal that is familiar to them. This behavior also raises an interesting issue regarding the degree to which Internet users are loyal to a particular site and/or

search engine. As shown through the post-questionnaire measures, if individuals are unable to find the information they need, they are very unlikely to change their search engine, but much more likely to change their query. Although this seems very intuitive, it further emphasizes that individuals are relying on very few sources, or search “brands,” to find information. It seems to highlight that once a search engine is perceived as credible and comprehensive, users will continually rely on that one site.

Impact on Advertising

Because Google does not accept payment from advertisers for higher-ranked search results, they have begun an advertising system called “AdWords,” which presents paid advertisements as text boxes to the right of the search results, and also above all of the search results in the form of highlighted tags. Although a script removed all of these ads from the page during the experiment, the conclusions generated from this study can still be used to develop insight about future advertising content.

The results from this research indicate that searchers make a document selection within the first few seconds of page viewing, and also that searchers most often click on the first promising link that they view. These results indicate that it is very unlikely that a searcher will attend to advertising on the opposite side of the page, especially if the task is not particularly difficult.

Differences between Informational and Navigational Tasks

Informational and navigational tasks each generated their own unique patterns of eye movements, specifically noted through differences in fixation duration. Existing eye-tracking literature shows that task difficulty will influence eye movements in reading tasks and visual search tasks in distinct ways. More specifically, fixation duration increases with task difficulty in reading tasks, while it decreases with

difficulty in visual search tasks. The results from this study indicate that navigational and informational queries each embody characteristics of a visual search and reading task, respectively. The data from this experiment showed that the average fixation duration in navigational tasks decreased with task difficulty, while the fixation duration for informational tasks increased with task difficulty. Essentially, this could mean that individuals are more prone to document scanning during the navigational queries, yet are more prone to careful reading when performing informational queries. This baseline knowledge can then be used to better inform the presentation of results based on query input. For instance, if the search system is able to distinguish between navigational and informational queries, the interface can be designed to better accommodate careful reading of snippets when users enter an informational query. Conversely, the results set can more explicitly facilitate scanning of the titles and URLs if the user entered a navigational query.

CHAPTER 7

FUTURE RESEARCH

Relevance Judgments

As discussed previously in this thesis, click-through data, though very useful and readily available, should be interpreted within the context of user behaviors before it becomes a widely used and reliable indicator of search performance.

We are presently beginning an analysis that will combine manual relevance judgments with the eye tracking data. One part of the analysis will address aspects related to the retrieval performance of the search engine, while others will relate relevance judgments to eye movements. First, volunteers were recruited and were informed of the overall search tasks. Each volunteer was presented with approximately 100 different queries that users formulated in response to the ten search tasks. Manual coders then viewed the abstracts that were displayed for each query and rank-ordered them according to most and least relevant to the search task. Approximately one-third of the queries were judged by two coders to ensure that the inter-coder reliability was acceptable.

Based on the manual relevance judgments, we can evaluate several issues. First, we can simply use a rank-order correlation coefficient to determine whether the manual rankings are equal to Google rankings. We may find that manual judgments generate rankings significantly different than Google on some queries, depending on the task or difficulty; alternatively, we may find that the manual judgments are not significantly different than those generate by the Google search engine.

However, the most interesting issues to address with the manual relevance judgments emerge when incorporating them into the context of the eye tracking data. For instance, based on the eye tracking data, it seems that users skip over some

abstracts, or spend proportionately less time in certain abstracts. We can then compare this information to the relevance judgments to determine if the abstracts skipped over are indeed less relevant than the surrounding abstracts. Users may perceive certain key words in their periphery that could provide them with a general impression of relevance for certain abstracts. Furthermore, it appears that even though the first and second- ranked results are viewed nearly equally, the first link gets clicked on much more often. The manual relevance judgments can then help us to determine if the first link really is indeed that much more relevant than the second.

Manual Relevance Judgments and Pupil Dilation

Another key issue that can be accomplished with the relevance judgments is to determine whether pupil dilation differs based on document relevancy. Eye tracking literature indicates that pupil dilation is larger when individuals are interested or aroused in the content. If certain abstracts are more relevant than others, it would be interesting to see pupil dilation increases with greater abstract relevance.

Search Success

Furthermore, another interesting concept for analysis is to address the of search success as a dependent variable. For this analysis, the number of experimental tasks that were correctly answered could be analyzed with pupil dilation, task difficulty, task type, and fixation duration. These results could potentially provide a very insightful indication of which ocular indices can best predict search success.

Perceived Trust and Credibility in Google

Another question to consider is how blindly users rely on Google to provide relevant search results. For instance, if Google presented links with low relevancy at

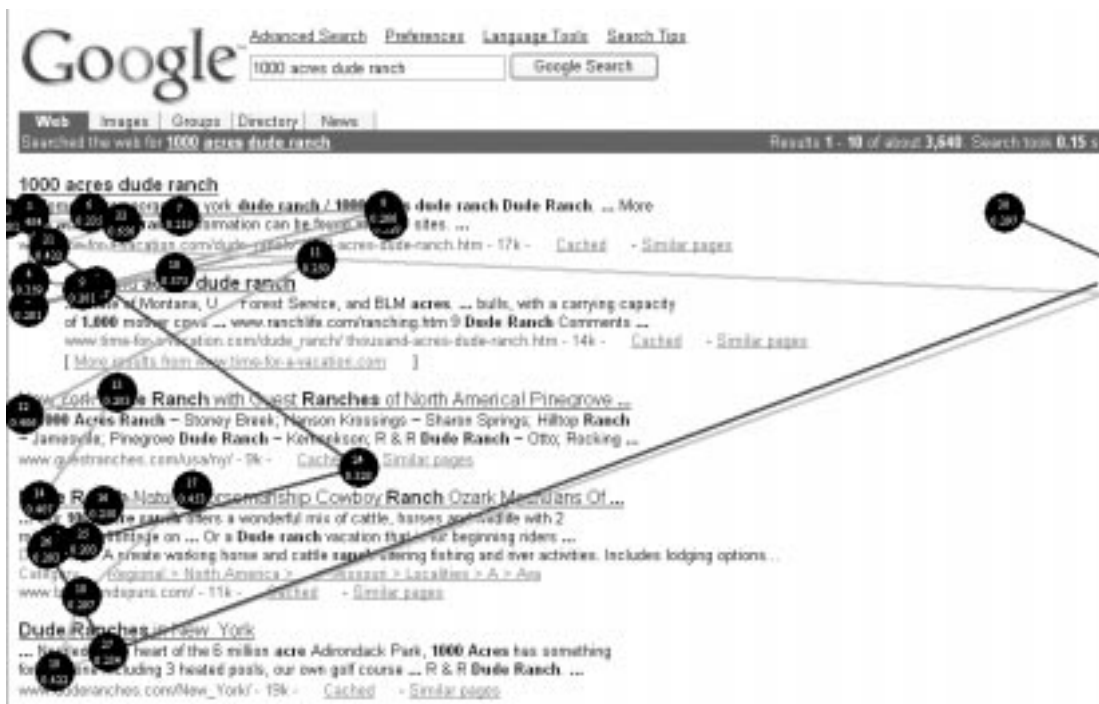
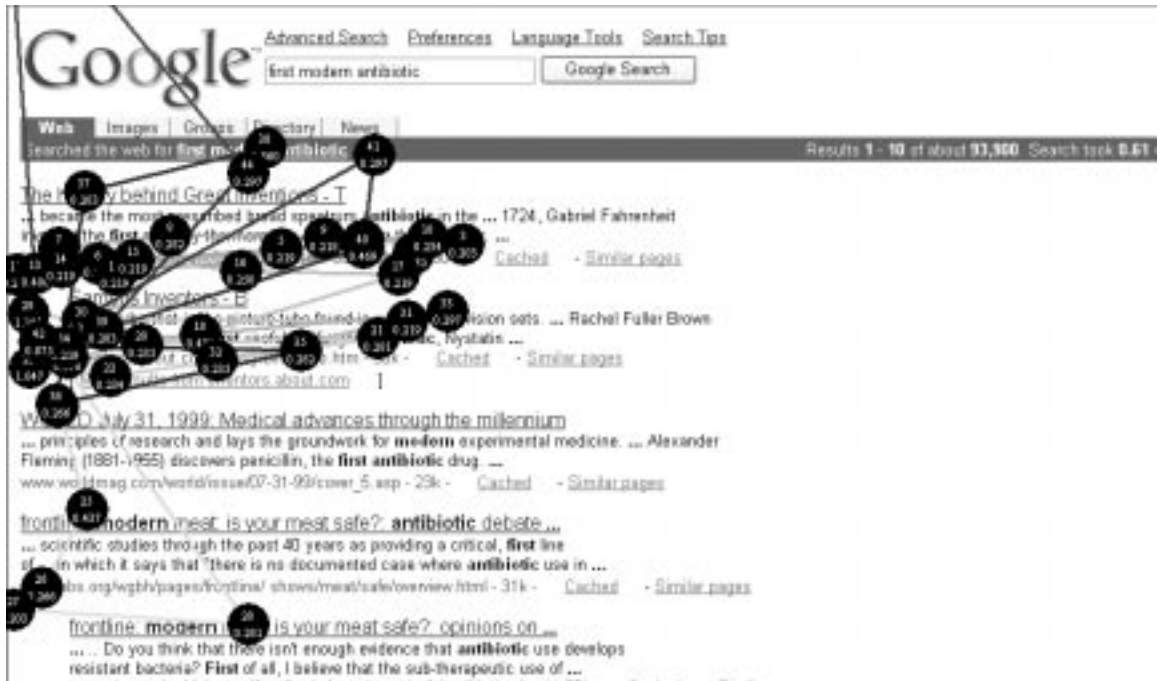
the top of results set, would users still trust that these lower-ranked documents were still very relevant, simply because the perceived Google ranking was very high? A script could be implemented to invert the results set, with the 10th ranked result instead being presented as the first. Research questions could then address whether searchers still select a document in such a short amount of time, whether the same number of abstracts are viewed, as well as many other of the same measures that were addressed in this present analysis.

APPENDIX A

MOST POPULAR SEARCH QUERIES AND THEIR FREQUENCY

20: michael jordan statistician
11: thousand acres dude ranch
11: one thousand acres dude ranch
9: 1000 acres dude ranch
7: time machine movie
7: carnegie mellon university graduate housing
6: imdb
6: greyhound bus
6: first modern antibiotic
6: emeril lagasse
6: carnegie mellon graduate housing
5: one thousand acres ranch
5: greyhound
5: graduate housing carnegie mellon
5: emeril
4: tallest mountain in new york
4: michael jordan statistics
4: greyhound route map
4: ezra cornell's house
4: ezra cornell house
4: emeril chef
3: tallest mountain, new york
3: new york tallest mountain
3: new york mountains
3: new york democratic primaries
3: mr. cornell
3: michael jordan statistician homepage
3: first antibiotic
3: emril chef
3: emeril cooking
3: dude ranch warrensburg
3: cornell
3: carnegie mellon campus life
3: carnegie mellon

APPENDIX B
SAMPLE EYE TRACKING OUTPUT



REFERENCES

- Antes, J.R. The time course of picture viewing. *Journal of Experimental Psychology*, 103, (1974), 62-70.
- Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc: Boston, MA.
- Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3-10.
- Card, S., Mackinlay, J., Schneiderman, B. 1999. *Information Visualization: Using vision to think*. San Francisco: Morgan-Kaufmann.
- Duchowski, A. T. A Breadth-First Survey of Eye Tracking Applications. *Behavior Research Methods, Instruments, & Computers (BRMIC)*, 34(4), November 2002, pp.455-470.
- Drucker, H., Shahrany, B., and Gibbon, D. (2002). Support vector machines: relevance feedback and information retrieval. *Information Processing and Management*, 38, 3, 305-451.
- Feusner, M. Eye-tracking visualizations. Retrieved April 1, 2004 from Cornell University, Human-Computer Interaction Group Web Site: <http://www.hci.cornell.edu/eyetracking/visualizations.php>
- Foltz, M., and Davis, R. 2001. Query by attention: Visually Searchable Information Maps. *Proceedings of Fifth International Conference on Information Visualization, London*.
- Gay, G. & Hembrooke, H. 2004. *Activity-Centered Design: An Ecological approach to designing smart tools and usable systems*. MIT Press: USA.
- Granka, L., Hembrooke, H., Gay, G., & Feunser, M. 2003. *Correlates of visual salience and disconnect: an eye-tracking evaluation*. Retrieved April 1, 2004, from Cornell University, Human-Computer Interaction Group Web Site: <http://www.hci.cornell.edu/eyetracking/EyeTrackingsalience.pdf>.

- Hearst, M. A. (1999). User Interfaces and Visualization. In Baeza-Yates, R. & Ribeiro-Neto, B. (Eds). *Modern Information Retrieval*. Addison-Wesley: Boston, MA.
- Hearst, M. A. (2000). Next generation web search: Setting our sites. *In IEEE Data Engineering Bulletin*.
- Hembrooke, H., Granka, L., Gay, G., Liddy, L. The Effects of Expertise and Feedback on Query Formation and Expansion: Developing a Typology of Search Term Strategy Use. Accepted with revisions in, *Journal of the American Society for Information Science and Technology (JASIST)*.
- Henderson, J. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7, 11, 498-504.
- Hess, E. and Polt, J. 1960. Pupil size as related to interest value of visual stimuli. *Science*, 132, 3423, 349-350.
- Hess, E. and Polt, J. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science*, 143, 1190-1192.
- Hsieh-Yee, I. 2001. Research on Web Search Behavior. *Library and Information Science Research* 23, 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*, 44(3): 161-174.
- 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: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2), 207-227.
- Jansen, B.J., Spink, A., Bateman, J., Saracevic. T. (1998). Real life Information Retrieval: A study of user queries on the Web. *SIGIR FORUM*, 32 (1), 5-17.

- Joachims, T. 2002. Optimizing search engines using clickthrough data. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, 132-142.
- Josephson, S., and Holmes, M.E. 2002. Visual Attention to Repeated Internet Images: Testing the Scanpath Theory on the World Wide Web. *Proceedings of Eye Tracking Research & Applications: Symposium 2002, ACM SIGCHI*, 43-51.
- Just, M.A. & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329-354.
- Kirsh, D. (1995). The Intelligent use of space. *Artificial Intelligence*, 73, (1-2), 31-68.
- Lankford, C. (2000). Gazetracker: Software Designed to Facilitate Eye Movement Analysis. *Proceedings of the Eye Tracking Research & Applications Symposium: ACM SIGCHI*, 43-51.
- Large, A., Beheshti, J., Rahman, T. (2002). Gender differences in collaborative web searching behavior: an elementary school study. *Information Processing and Management* 38 (3), 427-443.
- Lazonder, A., Bieiemans, H., & Wopereis, I. (2002). Differences between novice and experienced users in searching for information on the World Wide Web. *Journal of the American Society for Information Science*, 51, 576-581.
- Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute.
- Loftus, G.R. and Mackworth, N.H. 1978. Cognitive Determinants of Fixation Location During Picture Viewing. *Journal of Experimental Psychology: Human Perception and Performance* 4, 565-572.
- Lohse, G. and Rosen, D. Signaling quality and credibility in yellow page advertising: the influence of color and graphics on choice. *Journal of Advertising*, 30, 2 (2001), 73-85.
- Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge University Press.

- McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*. John Wiley and Sons: New York.
- Meyers-Levy, J., and Maheswaran, D. 1991. Exploring Differences in Males' and Females' Processing Strategy. *Journal of Consumer Research*, 18, 63-70.
- Nakayama, M., Takahashi, K., and Shimizu, Y. (2002). The Act of Task Difficulty and Eye-movement Frequency for the 'Oculo-motor indices'. *Proceedings of Eye Tracking Research & Applications: Symposium 2000, ACM SIGCHI*, 43-51.
- Nordlie, R. (1999). User Revealment – A comparison of initial queries and ensuing question development in online searching and human reference interaction. *Proceedings of SIGIR*, 11-18.
- Pan, B., Hembrooke, H., Gay, G., Granka, L., Feusner, M., Newman, J. 2004. The determinants of Web page viewing behavior: an eye-tracking study. *Proceedings of the Eye Tracking Research and Applications: Symposium 2004, San Antonio*, 147-154.
- Pelz, J.B., Canosa, R., and Babcock, J. (2000). Extended Tasks Elicit Complex Eye Movement Patterns. *Proceedings of the Eye Tracking Research and Applications: Symposium 2000, ACM SIGCHI*.
- Pirolli, P., Card, S., Van Der Wege, M. (2001). Visual Information Foraging in a Focus + Context Visualization. *Proceedings of CHI 2001, Seattle, Washington*.
- Rayner, K. (1998). Eye movements in reading and information processing: Twenty years of research. *Psychological Bulletin*, 124: 372-422.
- Roy, M., Taylor, R., Chi, M. (2003). Searching for information on-line and off-line: Gender differences among middle school students. *Journal of Educational Computing Research*, 29 (2), 229-252.
- Russo, E. J. & LeClerc, F. (1994). An Eye-Fixation Analysis of Choice Processes for Consumer Nondurables. *Journal of Consumer Research*, 21, 2, 274-290.

- Salogarvi, J. Kojo, I., Jaana, S., & Kaski, S. (2003). Can relevance be inferred from eye movements in information retrieval? *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03), Hibikino, Kitakyushu, Japan*, 261-266.
- Schneiderman, B., Byrd, D. & Croft, W.B. (1997). Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*, 1.
- Schneiderman, B., Feldman, D., Rose, A., Ferre Grau, X. (2000). Visualizing Digital Library Search Results with Categorical and Hierarchical Axes. *Proceedings of the fifth ACM conference on Digital libraries*, 57-66.
- Sherman, C. (2002, August 29). Why search engines fail. *Search Engine Watch*. Retrieved May 15, 2004 from <http://searchenginewatch.com/searchday/article.php/2160661>
- Silverstein, C., Henzinger, M., Marais, J., Miricz, M. (1998). Analysis of a very large AltaVista query log. Technical Report, Hewlett Packard Laboratories, Number SRC-TN 1998-014, Oct. 19.
- Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing and Management*, 38, 401-426.
- SPSS. Linear mixed-effects modeling in SPSS. Technical report.
- Spurgeon, D., and Duke, S. PROC MIXED: Key concepts for appropriate mixed-model analysis.
- Veerassamy, A., Belkin, B. (1996). Evaluation of a tool for visualization of information retrieval results. *Proceedings of the Annual ACM conference on Research and Development in Information Retrieval, SIGIR*, 85-92.
- Viviani, P. 1990. Chapter 8. In Kowler, E. (Ed.) *Eye Movements and Their Role in Visual and Cognitive Processes*. Amsterdam: Elsevier Science.
- Wolfe, J. Moving towards some solutions in the enduring controversy of visual search. (2003). *Trends in Cognitive Science*, 7, 2, 70-76.

Wooding, D. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, 34 (4) (2002), 518-52.

Yarbus, A.L. 1967. *Eye Movements and Vision*. Plenum Press: New York.