Looking for Query Terms
on Search Engine Results Pages

A thesis submitted to the College of Communication and Information of Kent State
University in partial fulfillment of the requirements for the degree of
Master of Library and Information Science

by
Alison M. Dickerhoof
August, 2014
Thesis written by
Alison M. Dickerhoof
B.A., Kent State University, 2012
M.L.I.S., Kent State University, 2014

Approved by

________________________________
Catherine L. Smith, Ph.D., Advisor

________________________________
Jeffrey W. Fruit, Interim Director, School of Library and Information Science

________________________________
AnnMarie LeBlanc, Interim Dean, College of Communication and Information
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>TABLE OF CONTENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II. REVIEW OF RELATED LITERATURE</td>
<td>2</td>
</tr>
<tr>
<td>Computer-Centric Retrieval Evaluation</td>
<td>3</td>
</tr>
<tr>
<td>Human-Centric Retrieval Evaluation</td>
<td>4</td>
</tr>
<tr>
<td>Implicit Feedback</td>
<td>5</td>
</tr>
<tr>
<td>Searchers’ Interactions with Words</td>
<td>8</td>
</tr>
<tr>
<td>III. OBJECTIVES</td>
<td>12</td>
</tr>
<tr>
<td>Research Question</td>
<td>13</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>13</td>
</tr>
<tr>
<td>IV. METHODOLOGY</td>
<td>13</td>
</tr>
<tr>
<td>Design</td>
<td>13</td>
</tr>
<tr>
<td>Data Collection</td>
<td>17</td>
</tr>
<tr>
<td>Participants</td>
<td>19</td>
</tr>
<tr>
<td>Procedure</td>
<td>20</td>
</tr>
<tr>
<td>Analysis</td>
<td>21</td>
</tr>
<tr>
<td>V. RESULTS</td>
<td>28</td>
</tr>
<tr>
<td>VI. DISCUSSION</td>
<td>33</td>
</tr>
<tr>
<td>Limitations</td>
<td>38</td>
</tr>
<tr>
<td>VII. CONCLUSION</td>
<td>40</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>41</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>A sample task</td>
<td>16</td>
</tr>
<tr>
<td>2.</td>
<td>A diagram depicting the flow of items viewed by a participant during each session. Solid arrows indicate types and directions of actions that always occurred. Dotted lines indicate optional actions. Rectangles denote items that were required to be viewed during the session and a circle indicates an item that was not required. The eye icon indicates the item for which recorded eye gaze data was later used in analysis.</td>
<td>22</td>
</tr>
<tr>
<td>3.</td>
<td>An example of the analysis spreadsheet records created to record data about SERPs that would be used for later analysis. Some of the records for participant P01 are shown here.</td>
<td>23</td>
</tr>
<tr>
<td>4.</td>
<td>A SERP screenshot with AOIs (rectangles), fixations (circles), and scan-paths (lines).</td>
<td>25</td>
</tr>
<tr>
<td>5.</td>
<td>A truncated example of Tobii Studio’s exported AOI Data that has been rearranged to show records pertaining to P01’s fixations on the author-and-book task and the corresponding SERPs. Records with the AoiNames sb_1, AOI_3, AOI_11, AOI_8, and AOI_9 represent query term fixations while records with the AoiName Content represent other fixations.</td>
<td>26</td>
</tr>
<tr>
<td>6.</td>
<td>AOI data exported for P01 after the addition of Task and QT FIXATIONS and the removal of all non-SERP and practice records.</td>
<td>28</td>
</tr>
<tr>
<td>7a.</td>
<td>Range of %QTF by participant</td>
<td>31</td>
</tr>
<tr>
<td>7b.</td>
<td>Range of %QTF by task</td>
<td>31</td>
</tr>
</tbody>
</table>
8. Lines plotting each participant’s average %AOI changes across tasks in presentation order……………………………………………………………………………………………………..32

9. Average %AOI by task………………………………………………………………………………………………………………………………………………………………………………………33
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Noun pairs used in tasks</td>
<td>17</td>
</tr>
<tr>
<td>2. Terms excluded from analysis</td>
<td>20</td>
</tr>
<tr>
<td>3a. The number of total fixations (TF), query term fixations (QTF), other fixations (OF), percent query term fixations (%QTF) and percent other fixations (%OF), relative to total fixations, by participant</td>
<td>29</td>
</tr>
<tr>
<td>3b. The number of total fixations (TF), query term fixations (QTF), other fixations (OF), percent query term fixations (%QTF) and percent other fixations (%OF), relative to total fixations, by task</td>
<td>29</td>
</tr>
</tbody>
</table>
Acknowledgments

First, I would like to thank my advisor Dr. Catherine Smith for introducing me to this subject and for her guidance and encouragement throughout this project. I would also like to thank Dr. Karen Gracy and Dr. David Robins for serving on my thesis committee and helping me perfect my work. Working with this group was a joy. Of course, I would also like to thank all of my friends and family for their continued support and words of encouragement that kept me motivated and saw me through to the end.
Chapter I

Introduction

In order to optimize online search engines, it is important to consider the humans who will be using those systems. As is stated by Saracevic (2011), “Effective designs of information retrieval systems and processes as done in computer science are dependent on the incorporation of an effective relationship with human-computer interaction, with the accent on human” (p. XXIX). There are many researchers in the field of library and information science studying human behavior as it relates to search systems, and more specifically, Web search engines. The study presented in this paper adds to this field of knowledge.

The study examined whether searchers look for their query terms when scanning a search engine results page (SERP) in order to select a Web page to fulfill a task. Understanding the focus of searchers’ attention when looking at results pages is important for the development of more effective search engines and better methods of implicit feedback. Feedback, both explicit and implicit, has been used to aid searchers with query reformulation when the initial query does not yield useful results. For search engines to be able to understand the searcher’s opinion of retrieved items, the system has to get feedback from the searcher. In early systems, the easiest way to understand the relevance of returned results from a user’s perspective was to gather explicit feedback. This method required that the searcher rate each returned result in terms of its relevance to his or her query. Implicit feedback methods were investigated and adopted after it was discovered that explicit feedback is burdensome on the searcher and that searchers are reluctant to provide it, if they will provide it at all. Implicit feedback gathers information about searchers’ behaviors and uses that information to infer the searchers’ relevance judgments of returned results. The user does not have to make any extra effort to provide implicit feedback, as
it is gathered from natural search actions already being performed. Implicit feedback is easier to
gather, but it does not always provide enough insight into the searcher’s intentions to allow the
search system to make good suggestions for query reformulation or further search. This is
especially true when search is difficult. Knowing what behaviors are naturally employed during
difficult search and how they can be used to infer relevance judgments will lead to new and
improved implicit feedback measures, which will in turn lead to systems that are more effective
for difficult search. This study is an investigation into one such specific behavior: whether
searchers visually fixate on their self-generated query terms when scanning a SERP for relevant
results to fulfill a search task.

In this paper, related research in the fields of information retrieval evaluation and user
feedback (both explicit and implicit) is reviewed. Then, a discussion of studies investigating the
use of eye-tracking to gather implicit feedback, and a brief review of the implications of
feedback for retrieval system design are presented. Next, research objectives are outlined,
including details of the research question and hypothesis. From there, the paper moves on to
describe the research methodology. This includes a discussion of data collection procedures and
analysis methods. Finally, the results of the study are presented followed by discussion,
limitations, and conclusion.

Chapter II

Review of Related Literature

The first section covers the literature on the topic of computer-centric retrieval evaluation
and its shift to the use of human-centric explicit feedback as a method of judging returned item
relevance. This is followed by a natural transition into the topic of implicit feedback and its
importance in improving information retrieval on the Web without the characteristic burdens of
explicit feedback. Described are several common methods of gathering implicit feedback. Each method’s potential for improving interactive search is explained while also pointing out any shortcomings of each method, in particular, those shortcomings pertaining to difficult searches. Next, the discussion turns to the literature on the use of eye-tracking to capture searchers’ eye movements and how this data can be used as a method of implicit feedback. Because this research involves user attention to query terms in particular, studies that point out the importance of searchers’ interactions with words are also reviewed.

**Computer-Centric Retrieval Evaluation**

This review begins with an explanation of the earliest form of information retrieval system evaluation. The research into the optimization of retrieval systems began in the 1960s with the study of indexing systems and the relevance of the documents they were able to retrieve. A notable researcher in this field was Cleverdon, who studied an early system evaluation method now known as the Cranfield paradigm. Testing using this paradigm is described well by Cool and Belkin,

A ‘test collection’ is constructed, which consists of a collection of documents, a set of descriptions of information ‘needs’, and an exhaustive evaluation of the relevance of each document to each information need description. The performance of the information retrieval system is evaluated by computing the recall, the ratio of relevant documents retrieved to the number of documents in the collection that are judged relevant to a query; and the precision, the ratio of relevant documents retrieved to all of the retrieved documents. (Cool & Belkin, 2011, p. 2)

This method was used for decades by researchers in the computer science field to test and evaluate information retrieval systems. The computer-centric approach was widely used because
it was easy to apply to all systems. The method revealed that over time systems were improving on measures of both recall and precision. However, there was a problem with this approach. It utilized fabricated test queries that retrieved pre-evaluated documents from a specific corpus. Systems deemed excellent at retrieval using this system were not necessarily as highly effective when presented with real users trying to express real information needs to systems without a specific set of already-evaluated documents. Real searches do not exactly match the types of fabricated test queries used in this early type of evaluation. Also, naturally, searches often require more than one query, each of which could be separated by long periods of time. In other words, computer-centric evaluation methods, such as the method described in the Cranfield paradigm, evaluate using searches considered in isolation, which simply does not match a natural querying process. In order to improve the efficacy of search systems used by common people, the systems had to learn to interact with humans.

**Human-Centric Retrieval Evaluation**

The first promising way to gather input from users was through relevance feedback. This form of feedback has since come to be known as explicit feedback and will be referred to as such throughout this discussion. Explicit feedback requires that users of an information retrieval system provide their judgments about the relevance of retrieved documents to their information needs. In order to obtain explicit feedback, the searcher must identify which of the items retrieved from a query he or she thinks are relevant and which are not. The feedback is then used to form a modified query that attempts to return another set of results with items more like those previously marked relevant and less like those previously marked not relevant (Rocchio, 1965). To do this, the system pulls key words out of relevant documents to generate a new query or to apply higher weights to previous query terms that appear in documents marked as relevant.
(Koenemann, 1996). Again, there is a problem with this approach to improving search interactions; there are high costs to the user when providing this type of feedback mid-search, and searchers are simply unwilling to provide explicit feedback often enough for it to be useful. This makes it difficult for the system to perform any useful query modifications, if it can perform any at all (Kelly & Teevan, 2003). The effect is particularly strong when search is difficult (White, Ruthven, & Jose, 2005). Users may simply forget to provide the feedback because the search task is so demanding, but the effect is seen mostly because searchers faced with a complex or difficult task are unlikely to find any of the initial search results relevant, leaving no opportunity for useful feedback.

While explicit relevance feedback has one distinguished shortcoming, its theory also holds a lot of potential. User feedback is helpful, but the users have to be willing and able to provide it. From this dilemma the concept of implicit feedback was born. Implicit feedback requires no extra work on the part of the search system user; instead, it involves gathering information about the behaviors of the user while he or she is searching. Because of its unobtrusive nature, it can be collected in larger quantities, which makes it beneficial for research. However, implicit feedback must be interpreted, and this process can be difficult (Joachims, Granka, Pan, Hembrooke & Gay, 2005).

**Implicit Feedback**

Claypool, Le, Wased, & Brown (2001) investigated several methods of gathering implicit feedback and each method’s potential for being useful. After comparing implicit feedback data to explicit feedback ratings, the researchers discovered the best methods of gathering implicit feedback. The amount of time spent on a Web page was found to be one of the behaviors most indicative of searchers’ relevance perception, and was therefore useful to record as implicit
feedback. For example, it was supposed that the more relevant the Web page, the more time the
participant would spend on it. When comparing data gathered about time spent on Web pages
against explicit relevance ratings for those pages, it was indeed found that participants spent
more time looking at Web pages that they deemed relevant.

Click-through data, or data about which items a searcher clicked on with a computer
mouse, is widely utilized as a method of inferring user interest and gathering implicit feedback.
It must be noted that the prevalence of this method of feedback is partially due to it being widely
recorded and available, and therefore easy to utilize and study. Also, it has been found to be
effective when used with common searches. The basis for this method of gathering implicit
feedback was the thought that users did not click on links retrieved by search systems randomly,
but rather, that they used some sort of relevance judgment in order to decide which link to click
(Joachims, 2002). Click-through data did show improvements in system personalization and
ranking (Joachims, 2002; Gao, Yuan, Li, Deng & Nie, 2009), but it suffered from incomplete
click and missing click problems when the users did not click on many, if any, links (Gao et al.,
2009). In addition, this method was weakened by the position bias, which states that items listed
first on a SERP are more likely to be clicked by searchers than those items listed further down
the page. This is likely because searchers have trust that search engines, such as Google, will
always return the most relevant results at the top of the SERP (Joachims et al., 2007). Search
systems are able to overcome the setbacks associated with click-through data for common
information needs, but this feedback method is still not particularly useful for less common,
more difficult information needs and searches.

Another way to record the focus of user attention on a SERP is by using eye-tracking.
Eye-tracking is used by researchers to gather information about searchers’ attention when
scanning the SERP before making a choice of where to click (Granka, Joachims, & Gay, 2004). Also, it has been proposed that the data the eye-trackers gather about user attention on the SERP can be interpreted as implicit feedback. The reason for this approach, it has been stated, is that “people tend to pay more attention to objects they find relevant or interesting” (Puolamäki, Salojärvi, Savia, Simola, & Kaski, 2005). Recent research proposes that the information gathered through eye-tracking can be used for query modification without much, if any, user feedback. First, eye-tracking during search can be helpful to the searcher by utilizing proactive information retrieval where “the system adapts to the interests of the user” and provides suggestions for improving search (Salojärvi, Kojo, Simola, & Kaski, 2003; Hardoon, Shawe-Taylor, Ajanki, Puolamäki, & Kaski, 2007). Even more innovative, Hardoon et al. (2007) suggests terms that were fixated upon after an initial search can be extracted and used again in a new query, which may zoom in on the user’s interests even better than the user’s original query. It is clear that there may be some room for search system improvement when eye-tracking devices are used.

Since eye-tracking requires specific equipment, and that equipment is not at this time widely incorporated into the machines most commonly used for search, eye-tracking data as a source of implicit feedback for everyday searchers is not yet practical. There are researchers who have investigated the use of mouse movements to estimate visual attention, since mouse movements are much easier and less expensive to record. Rodden and Fu’s (2006) examination of using mouse movements to estimate eye movements provided inconclusive results. Later, Huang, White, and Buscher (2012) suggested that how well mouse movements can predict gaze depends on the situation. It was found that using mouse movements to approximate gaze or visual attention is useful sometimes, especially when high accuracy is not necessary.

When evaluating Web search engines, one must not consider retrieval and query
formulation as the only components of the system’s overall usefulness. Research has also focused on the visual elements of a search engine’s interface and how those elements relate to search engine design. Halverson and Hornof (2004) described the lack of research into search engine design and visual search. They indicated the need for an investigation into which visual elements best capture user attention. The identification of such visual elements can then be used to create more functional search engine design. Eye-tracking systems and studies like this one can be useful in determining areas of visual attention and in providing information that can be used to improve the visual design of search engines.

So far, the evolution of evaluation and feedback has been covered from the original computer-centric focus on retrieval system performance as defined by unnatural test queries and created corpora into relevance feedback provided by humans, first explicitly, then implicitly. It is at this point that the focus of the literature review will shift to discuss studies that focus on the searcher’s interaction with words, as studies have shown that they can be a good source of implicit feedback.

**Searchers’ Interactions with Words**

Even before the ubiquity of Web search engines seen today, research investigating the use of user-generated terms for query modification can be found. Golovchinsky, Price, and Schilit (1999) explored the use of document annotations and mark-ups as a potential source of implicit feedback. The study used standard explicit relevance feedback as a control and implicit feedback as the experimental variable. The researchers hypothesized that the terms extracted from user-generated annotations would be better at capturing the aspects of documents that participants felt were relevant than would traditional explicit relevance judgments on each entire document, and that query expansion using feedback from the implicit method would outperform
query expansion using feedback from the explicit method. The participants were told to use certain queries, which were predetermined to retrieve a set of documents containing both relevant and non-relevant items. Participants were then asked to read those retrieved documents and annotate them electronically. The words used in the annotations were extracted by XLibris and used again as feedback for query expansion. The participants were also required to judge each document’s relevance to their query, and that relevance judgment was used as a source of explicit feedback. Query expansion was performed using feedback from both methods. The queries that were formed using keywords extracted from the electronic annotations produced better results in terms of recall and precision than did queries formed from explicit relevance feedback, suggesting that user-generated terms are useful for providing query modification.

User interaction with words displayed on a SERP has also been shown to be a useful form of implicit feedback. The 2012 study by White and Buscher used selected text (that is, text that was highlighted by mouse clicks) as a source of data for feedback before re-ranking the results page. This was done using the words that appeared in result snippets (lines of text extracted from a Web page situated under the link, used as an abstract) containing the selected text. The study was preliminary, and the authors suggested that recording what users did after selecting the text might yield even more accurate feedback. Although this method did not utilize user-generated terms for feedback, it reinforces the notion that human interaction with words on SERPs is useful for improving relevance ranking in terms of what the user finds relevant to an information need instead of simply what the search engine determines is most relevant.

There is evidence to suggest that bolded query terms in the titles of Web pages direct users’ visual attention. When a person submits a query to a search engine, the results appear as titles, with URLs underneath them and snippets underneath that. Yue, Patel, and Roehrig (2010)
found that users were more likely to click on links (titles) with many bolded query terms in them than they were to click on links which did not contain a high number of bolded query terms. While the study by Yue et al. (2010) found that effects from position bias were much stronger than those effects resulting from query bolding, the query bolding effect was still significant. Interestingly, there were no significant effects found for clicks on links that had a high number of bolded query terms in the snippet (as opposed to the title link). This study did not use eye-tracking.

Yue et al. (2010) supported previous findings of Clarke, Agichtein, Dumais, and White (2007), who were interested in how the features of the captions (title, snippet, and URL) presented on SERPs influence how users click on results. The study found, among other things, that it is important for query terms to be present in each result’s caption. Specifically, the total absence of query terms in the caption negatively affects the number of times a searcher will click on a result. The findings led the researchers to develop several guidelines, the first of which provides evidence of the importance of query terms appearing on the SERP in the caption for each result. The first guideline states, “Whenever possible all of the query terms should appear in the caption, reflecting their relationship to the associated page” (p. 141). This is because, the researchers suggest, it may be difficult for a searcher to determine the relevance of a caption that does not include the submitted query terms. This supports the notion that searchers do look for their query terms on the SERP and that the presence of those terms is important for the searcher in determining which results will take them to useful sources.

When evaluating the results of a search study, Toms et al. (2008) unintentionally identified an interesting phenomenon involving how the participants of the study generated their query terms. The researchers were interested in the effects of different tasks on search behavior.
Three types of information goals and two types of task structures were used to create 12 tasks, and the differences between search behaviors for each task were recorded and analyzed to determine the number of overlapping query terms in each task. An overlapping query term is a word that appears in both a participant’s query and in the task description provided by the researcher. Also, the analysis described how many terms were used in each query. It was discovered that the participants’ queries used words from the task description more often than original words. On average, over all task types, 1.55 query terms overlapped with keywords found in the task description. The number of terms per query, on average over all task types, was 2.52 keywords. The difference between the two numbers is only 0.97, which represents the number of terms per query which were not attributable to task description words. The numbers show that the study’s participants obtained more than half of their query terms from the words provided in the task description. This behavior, if true for all Web search studies with written search tasks, is problematic for researchers interested in replicating situations in which it is not easy for searchers to think of the terms they will use in a query. Because this study aimed to replicate difficult search, its tasks were designed to reduce this term overlap. The task design is discussed in more detail subsequently.

The words on SERPs, both displayed by the search engine and generated by the user, have been shown to be useful as implicit feedback. User-generated query terms are more useful for query modification than are searchers’ simple judgments of the relevance of retrieved items. Similarly, the words users choose to interact with, by highlighting them with mouse clicks, for instance, are useful in changing the way retrieved results are ranked in terms of relevance to the user’s query. Also, bolded query terms on the SERP guide the attention of searchers. However, it is not known whether searchers scanning a SERP are looking at the words they used in their
query. This study addresses this question. Its results are important in understanding more about the behaviors people employ when determining which link to click, as well as how they choose words for a new query when a prior query does not return all or any of the items necessary to meet their information need.

Chapter III

Objectives

Previous studies have found that understanding the eye movement behavior of searchers while looking at results lists can be a useful source of data for implicit feedback, which helps improve search, both in terms of inferring relevance judgments and presenting useful query modification suggestions (Granka, Joachims, & Gay, 2004; Puolamäki, Salojärvi, Savia, Simola, & Kaski, 2005; Salojärvi, Kojo, Simola, & Kaski, 2003; Hardoon, Shawe-Taylor, Ajanki, Puolamäki, & Kaski, 2007; Hardoon et al., 2007). Other research suggests that the presence of self-generated query terms on the SERP direct user attention and are important for making decisions about which links to click (Yue, Patel, and Roehrig, 2010; Clarke, Agichtein, Dumais, & White, 2007). The studies suggest that searchers look at their query terms when scanning a SERP because the words are useful in determining the relevance of each link and therefore the adequacy of the searcher’s query. This study seeks to describe the natural behavior of people using a Web search engine, so that we may better understand how users interact with query terms they use. More specifically, the research question addressed is: When scanning the results page of an online search engine, do users fixate visually on their submitted query term(s) more than they fixate on any other visual element.

Because the findings of Toms et al. (2008) suggest that participants in search studies use the search task text when creating queries, this study used tasks that required participants come
up with their own query terms. Therefore, the findings of this study allowed us to describe how
users interact with self-generated query terms.

Research Question

When using an online search engine, do users fixate visually on their submitted query
term(s) more than any other visual element when scanning the results page for a useful Web
page?

Hypothesis

When looking for a useful Web page on a SERP, participants will fixate on their query
terms more than any other visual element.

Chapter IV

Methodology

In this section, the experimental design is outlined, and then the methods used to collect
data are described. Next are descriptions of participants and the session procedure. Details of
analysis are then provided, followed by results, discussion, limitations, suggestions for future
work and a conclusion.

Design

The study utilized Tobii Studio and a Tobii T60 eye-tracker to collect and observe eye
gaze data for people searching on the Web in a laboratory setting. Participants were assigned
tasks that required them to self-generate query terms to be used during search. The SERPs were
then analyzed to reveal whether participants fixated on their self-generated query terms when
looking at the SERP to find a Web page to fulfill the task. The specific aspects of the design are
detailed below.

In this study, participant eye movements were recorded during Web search. These were
analyzed in order to understand whether people visually fixate on their self-generated query terms when looking for a Web page on a SERP. Fixation was defined by Joachims et al. (2005) as, “A spatially stable gaze lasting for approximately 200-300 milliseconds, during which the visual attention is directed to a specific area of the visual display” (p. 156). Eye-movement data indicates where a searcher is looking; however, in order to make the tracking data useful, the location of each fixation must be coded using *areas of interest* (AOIs), which are discussed next.

The Tobii Studio eye-tracking system contains a tool that allows researchers to draw areas of interest (AOIs) on images captured during eye-tracking, such as on a screenshot of a SERP. An AOI is defined by Poole and Ball (2006) as “an analysis method used in eye-tracking. Researchers define areas of interest over certain parts of a display or interface under evaluation, and analyze only the eye movements that fall within such areas” (p. 10). AOIs are typically defined after data has been gathered, and this is done by manually drawing rectangular boxes around those areas. Manually drawing AOIs around query terms on SERPs has been done before to analyze eye-tracking data from a search study (Balatsoukas & Ruthven, 2012). Once AOIs are defined, Tobii Studio’s analysis functions can then determine automatically many things about the fixation(s) that landed within the boundaries of those areas, including whether a fixation occurred there. The process of defining what constitutes a query term and the boundaries of AOIs are discussed below.

Participants were presented with tasks to initiate and motivate Web search. Because the type of task assigned can affect the time spent on SERPs, it was important to assign the appropriate type of search task to motivate searchers to spend some time looking at SERPs. There are three types of commonly observed natural search tasks, and one of them allows for more complex searching. Broder (2002) identified types of natural search as transactional,
informational, and navigational. The intent of a transactional information retrieval task is “to perform some Web-mediated activity”, such as shopping; the intent of an informational information retrieval task is “to acquire some information assumed to be present on one or more Web pages”; the intent of a navigational information retrieval task is “to reach a particular site” (p. 5). Both the transactional and navigational tasks focus on finding one Web page and using it to complete the task. Only the informational task allows for a more complex search task with a high likelihood of more than one query. Because of this, the tasks used in this study were designed to have informational goals.

Research by Toms et al. (2008) suggests that participants in Web search investigations often copy their query terms from the tasks they are given. Giving participants the query terms to use in search does not mimic natural, difficult search; therefore, the tasks used in this study have a unique design. Participants were given two keywords to evoke thoughts of a broader concept, but were prohibited from using those two keywords during search. This ensured that the words used to search were self-generated by the searcher. It also more closely resembled a natural setting in which query terms are difficult to generate, which makes the findings of this research applicable to difficult search. A detailed description of the construction of search tasks follows.

Each task was created using two well-known nouns, (which were related to each other by a broader concept), and a set of directions. Each task directed participants to find a Web page that provided information about a relationship between the two nouns, and to search without using those nouns as query terms. The nouns were chosen so that both related to a single topic that could be referred to by a different word or words. Specifically, the first noun in each task named something that could be used or could itself act upon the second noun. That action could
be described using one or many words, and those words were the intended query terms. The nouns provided in the sample task (see Fig. 1) aimed to evoke thoughts of an action such as *herding*. A list of all twelve noun-pairs used in the study can be found in Table 1. In an initial design of the study, the use of query suggestions was also going to be prohibited because it would allow participants to submit queries containing words the participant did not generate. Because it was determined that prohibiting this action would create an unnatural search environment, this limitation was later removed, and participants were permitted to click on queries provided by Google through its instant query suggestion service. There was no data collected on the frequency of query suggestion use, but the researcher’s observation was that its use was minimal.

Tobii Studio has a system for structuring studies; researchers create *projects* containing *tests*. Each test has its own *timeline of media (stimuli)*. For this study, only one project containing one test was used. Three types of media were used – *instructions, images, and Web*. Instructions to participants were typed directly into Tobii Studio by the researcher. Task statements were created externally as .jpg image files, and then uploaded to Tobii Studio and

---

**Find a website that provides information about a relationship between the terms DOG and SHEEP**

You may use any strategy to search for the website except you may NOT use the provided terms in your query. You may use any other words that come to mind. You may reformulate your query as many times as necessary. When you have reached a website that provides satisfactory information, press F10 on the keyboard.

---

Figure 1. A sample task.
Table 1. Noun pairs used in tasks

<table>
<thead>
<tr>
<th>Task Noun Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>author and book</td>
</tr>
<tr>
<td>brush and cat</td>
</tr>
<tr>
<td>citizen and ballot</td>
</tr>
<tr>
<td>dog and sheep (practice)</td>
</tr>
<tr>
<td>eye and page</td>
</tr>
<tr>
<td>father and child</td>
</tr>
<tr>
<td>maid and house</td>
</tr>
<tr>
<td>match and wood</td>
</tr>
<tr>
<td>car and road</td>
</tr>
<tr>
<td>needle and thread</td>
</tr>
<tr>
<td>oven and food</td>
</tr>
<tr>
<td>tractor and seed</td>
</tr>
</tbody>
</table>

placed on the timeline as images. Access to the Web (termed Web stimuli) was placed between task images to allow participants to search online using Internet Explorer for access to google.com. An instruction slide directing participants to take a brief break was placed on the timeline after the sixth task image. To control for any unintentional interaction effects between tasks, the order of task presentation was controlled automatically by Tobii Studio using a Latin square design. The practice task image, though, was always presented first. To discourage any purposeful fixations on or visual avoidance of the query terms on the SERP, the researcher did not mention the purpose of the study during instructions.

Data Collection

Data collection involved running a Windows 7 PC to host Tobii Studio version 2.1.13 which was connected via a TCP/IP connection to the Tobii T60 Eye-tracker. During the execution of each stimulus on the timeline, Tobii Studio gathered eye gaze data and the participants’ view of the eye-tracker screen (screenshots). Only the eye gaze data gathered while participants were viewing Google SERPs was later used for analysis. All of the default settings were used in Tobii Studio for eye-gaze data collection, which included using the Tobii Fixation
Filter. The Tobii Fixation Filter is used to define and generate fixation data, using a model that analyzes the path of eye movements to determine when the eyes have stopped moving and have fixated on something. The model uses the position of the gaze and elapsed time in combination to separate fixations from other eye movements. In order to determine which points do and do not belong to the same fixation, the Tobii Fixation Filter, “detects quick changes in the gaze-point-signal using sliding averaging” and then determines which points belong to the same fixation using a threshold value in pixels (Tobii Studio 1.X User Manual, 2008). The default fixation radius (threshold) is 35 pixels. These default settings have been used in at least one other LIS study of Web search (Kules & Xie, 2011). All browsing history, including temporary Internet files, cookies and Web page data, history, download history, form data, and passwords in Internet Explorer were deleted between participants so that one participant’s browsing history did not influence the search results of later sessions.

After all sessions were completed, additional data was collected by coding the fixations on each SERP using AOIs (defined above). Tobii Studio contains a Visualization tab in which researchers can analyze the eye gaze data for pertinent media. All media screenshots appeared in a media list arranged alphabetically. The Web pages were collected here as distinct screenshots and were named automatically by URL. Web page images representing SERPs were grouped together automatically in the media list because they were all named with URLs beginning https://www.google.com/search. A new Web page screenshot was generated whenever a participant clicked a link in the Web browser. So, if a participant was looking at a SERP, then clicked on a link to view a Web page, then hit the back button in the browser to go back to look at the same SERP again, that SERP was represented by two separate items in the media list. The two images looked identical, but the eye gaze data for each was different because it was gathered
at two different points in the participant’s search session. Web page screenshots were also generated for any other sites viewed by participants, but only those corresponding to Google SERPs are used in further analysis. Any further mention of Web page screenshots will refer only to those representing SERPs and will be referred to as SERP screenshots.

On each instance of each included SERP in the media list, the researcher manually drew AOIs around each query term as it appeared in each image; however, not every word in a query was identified as an AOI. Only major concept words were identified as AOIs, and all other words were excluded. For the purposes of this study, major concept words were words important to understanding the main ideas of the query. A comprehensive list of excluded words can be found in Table 2. Typically, these words were articles, conjunctions, prepositions, and pronouns. For example, when prepping the query “how to and voting and democracy” for analysis, the words how, to, and and would have been excluded from AOI drawings and therefore from all further analysis. Only the query terms identified as AOIs were included in any further query term counts, so the number of query terms in the above example is not six but two (voting and democracy).

After AOIs were drawn, Tobii Studio compared the eye gaze data gathered during the experiment with the AOIs drawn post hoc to determine when fixations occurred within AOIs. AOI Data was exported as a text file from Tobii Studio and used in analysis. A file was created for each participant, where each record in the file indicated a fixation. Fixations on AOIs were identified by distinct AoiNames. A more detailed description of these files and their use during analysis is described in the analysis section.

Participants

Participants were recruited using CROPs, a pool of undergraduates who are required to
Of the 21 appointments available, 20 were scheduled and 18 participants completed the study. Many participants took less than the allotted 45 minutes to complete the study. Each participant earned course research credit. Every participant had either normal or corrected-to-normal vision. Based on education level and major areas of study indicated on a demographic questionnaire, it can be inferred that no participants had experience in any expert search training. All participants were between the ages of 18 and 24.

**Procedure**

Each session was scheduled to last 45 minutes and participants were permitted to complete up to 12 tasks (one of which was practice) during the session. Approximately the first 5 minutes of each session was used to verify the participant’s identity and to have each participant fill out a demographic questionnaire and sign both an attendance sheet and a consent form. Brief
verbal instructions were provided and then the participant was then seated and his or her position relative to the eye-tracker was fine-tuned. Calibration for the eye-tracking system was performed next, followed by the display of additional brief instructions on the Tobii eye-tracker screen. Participants were permitted to ask questions before, during, and after the practice task, although some questions were asked during the experiment. Each participant moved through the items presented at his or her own pace with no time limit for any particular item. Participants used the space bar on the keyboard to move through instructions and task images. To end a Web search session and view the next task, participants pressed F10. A diagram depicting the flow of items viewed during each session can be found in Figure 2. Each task was displayed in print on the table next to the computer for reference during search. When the break instructions were reached, participants were allowed to pause for as long as they wished. The self-paced nature of this set-up helped to eliminate some sources of stress and allowed each individual participant to search at a natural pace. In order to discourage rushing, participants were not notified of the maximum number of search tasks available. After the final search of a session was performed, participants were presented with an instruction item indicating that the session was complete. At that time, each person was thanked and given a copy of the consent form. Every participant whose eye gaze data is included in the subsequent analysis was able to complete all tasks during the session time. None of these participants had to be stopped early, although there were other participants, not included in the final analysis, whose searches would have lasted longer than the allotted session time, and those participants were stopped when the session was over.

**Analysis**

The following process was used to analyze the collected data. First, each SERP, represented by a distinct screenshot in Tobii Studio’s Visualization media list, was examined
individually. After starting with the first SERP on the media list and coding AOIs on approximately twenty-five of them, it was determined that coding all SERP screenshots from all eighteen subjects was not feasible in the time allotted for analysis. As the original plan called for analysis of a minimum of ten subjects, this became the goal. The ten subjects were selected through a rational process as follows: First, four of the eighteen participants were excluded because their demographic questionnaire responses indicated that English was not their native language. Also, a fifth participant was excluded because Tobii Studio crashed several times during that particular session. Finally, ten of the thirteen remaining participants were selected at random for analysis. AOI coding then proceeded.

Because the SERP screenshots were not arranged in a useful order for our purposes, steps had to be taken to collect and arrange the available data in a useful way. Before analysis began, a separate Excel analysis spreadsheet was created. Included in the analysis spreadsheet was one record for each SERP that appeared as a screenshot in Tobii Studio’s media list, including screenshots for participants excluded from analysis and those representing practice tasks. An
Figure 3. An example of the analysis spreadsheet records created to record data about SERPs that would be used for later analysis. Some of the records for participant P01 are shown here.

Identifier number was given to each record in the SERP column representing that screenshot’s position in the media list. The associated participant ID (P_ID) was also recorded. Having these identifiers for each SERP screenshot in the media list aided in collecting data during the AOI coding process (described below). As analysis proceeded, more information was added to each record of the spreadsheet. Figure 3 depicts the analysis spreadsheet.

AOIs were drawn on one SERP screenshot at a time. All SERP screenshots for one participant were coded with AOIs before moving on to draw AOIs on screenshots for the next participant. The portion of the media list pertaining to SERP screenshots was scrolled through in its entirety for each participant in order to find every related SERP. The AOI drawing and data collection process for each SERP is described next.

First, the participant’s query as it appeared in the SERP screenshot (with excluded words) was copied into the spreadsheet (See the Notes column in Figure 3). Next, the number of query terms, not including excluded words (see exclusion rule in Data Collection) was recorded in Number QTs in the spreadsheet (see Figure 3). Then, using the computer’s mouse, AOIs were drawn manually by the researcher around all included query terms. A rectangular shape fitting snugly around each word was used for all AOIs. AOIs were drawn in the following order: Starting at the top of the screenshot image, an AOI was drawn around included query terms
appearing in the search box. Then, an AOI was drawn around every instance of each query term as it appeared anywhere in the SERP text, including advertisement text. Each AOI was labeled automatically with the convention \textit{AOI\_n} where \textit{n} equals the sequential position of that AOI in drawing order. Query terms appearing in the search box were manually renamed with the prefix \textit{SB\_} for \textit{search box} instead of \textit{AOI\_}. Exact word matches were included along with any other form of each query term, including singular, plural, verb, noun, and adjectival forms. In the example query presented in the discussion of data collection methods (“how to and voting and democracy”), the words vote, votes, voting, voter, democracy, democracies, democratic, democratically, democrat, and democrats would have all been included. AOIs were not drawn around abbreviations, compound words, or partial words, such as in a link preview containing \textit{democ}. Words that appeared in images and not directly in the SERP text were not included. Care was taken to ensure that each AOI covered one whole word while excluding any part of any other word, but in some instances this was not possible and some overlap occurred. In many such instances, query terms were clustered together so tightly that their corresponding AOIs overlapped. Other times, the AOI box was slightly too large and could not be reduced, so it overlapped with a punctuation mark, or a small part of an adjacent non-query term word. Examples of manually drawn AOIs can be seen in Figure 4 as rectangles around query terms.

After all AOIs were drawn on all SERP screenshots corresponding to one particular participant, that participant’s AOI fixation data was exported as a text file from Tobii Studio and opened with Excel. This exported data provided one record for each fixation anywhere on each SERP, including fixations on AOIs. Sometimes, an AOI was fixated on more than once. In that case, each fixation on the AOI was recorded separately, in its own row. Each record contained a field for \textit{timestamp} (a record of the time, in milliseconds, that a fixation was recorded), \textit{fixation}
Figure 4. A SERP screenshot with AOIs (rectangles), fixations (circles), and scan-paths (lines).
duration (the length of time, in milliseconds, that the fixation lasted), AoiNames (the AOI labels automatically generated during AOI drawing (AOI_n) or renamed to indicate a query term in the search box (SB_n)), and StimuliName (the name of each item as it appeared in the media list, with URLs naming SERP screenshots). There was also a column labeled AoiID, which is supposed to label each AOI sequentially as it was drawn on each SERP, but in the exported data all values in this column were zero for all participants, so this column was ignored. See Figure 5 for an example of Tobii Studio’s exported AOI Data.

Data properties:

<table>
<thead>
<tr>
<th>Recording date</th>
<th>Recording time</th>
<th>Study</th>
<th>Subject</th>
<th>Recording</th>
<th>Screen resolution</th>
<th>Coordinate unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/21/2014</td>
<td>09:01:47:572</td>
<td>Search Tasks</td>
<td>P01</td>
<td>Rec 01</td>
<td>1280 x 1024</td>
<td>Pixels</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>FixationDuration</th>
<th>AoiIds</th>
<th>AoiNames</th>
<th>StimuliName</th>
</tr>
</thead>
<tbody>
<tr>
<td>387869</td>
<td>666</td>
<td>0</td>
<td>Content</td>
<td>authorbook.jpg</td>
</tr>
<tr>
<td>388535</td>
<td>500</td>
<td>0</td>
<td>Content</td>
<td>authorbook.jpg</td>
</tr>
<tr>
<td>389151</td>
<td>67</td>
<td>0</td>
<td>Content</td>
<td>authorbook.jpg</td>
</tr>
<tr>
<td>465164</td>
<td>233</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>465397</td>
<td>466</td>
<td>0</td>
<td>sb_1</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>469794</td>
<td>83</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>469878</td>
<td>1266</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>511265</td>
<td>199</td>
<td>0</td>
<td>AOI_3</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>511681</td>
<td>200</td>
<td>0</td>
<td>AOI_11</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>511881</td>
<td>150</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>512031</td>
<td>200</td>
<td>0</td>
<td>AOI_8</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>512231</td>
<td>200</td>
<td>0</td>
<td>AOI_8</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>512431</td>
<td>183</td>
<td>0</td>
<td>AOI_9</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>512614</td>
<td>150</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>512764</td>
<td>300</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>513063</td>
<td>433</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>513496</td>
<td>200</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>513696</td>
<td>216</td>
<td>0</td>
<td>Content</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
<tr>
<td>513913</td>
<td>200</td>
<td>0</td>
<td>sb_1</td>
<td><a href="https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2">https://www.google.com/search?hl=en&amp;source=hp&amp;q=reading&amp;gbv=2</a></td>
</tr>
</tbody>
</table>

Figure 5. A truncated example of Tobii Studio’s exported AOI Data that has been rearranged to show records pertaining to P01’s fixations on the author-and-book task and the corresponding SERPs. Records with the AoiNames sb_1, AOI_3, AOI_11, AOI_8, and AOI_9 represent query term fixations while records with the AoiName Content represent other fixations.
In order to prepare the data for statistical analysis, two fields were added to the exported AOI data sheet, *Task* and *QT FIXATIONS*. To label each record with its corresponding task, the records were arranged in timestamp order. This allows the use of the task image names (listed under StimuliName) to determine which media records corresponded with each task. For example, all records occurring after those labeled *authorbook.jpg*, but before records corresponding to the next task image, were coded *author and book* in the Task column. *QT FIXATIONS* were determined using an IF statement which looked at the AoiNames. Those labeled Content were coded as non-query other fixations (OF) and anything else was labeled a query term fixation (QTF). See Figure 6 for an example of the AOI data with the added fields.

Finally, all records containing fixation information for any media item other than SERP screenshots and all records pertaining to practice tasks were removed.

Using the values of QT FIXATIONS, the number of QTFs and OFs were counted. The totals of QTF and OF were summed to determine total fixations (TF) for each participant. Finally, query term fixations per total fixations (QTF/TF) and other fixations per total fixations (OF/TF) were calculated for each participant. This process was repeated with each sheet of exported AOI Data, one for each participant (10). Then, all records from each participant’s modified AOI Data sheet were combined to create one spreadsheet with all AOI Data records for all participants. This data was then rearranged to display QTF and OF by task. The same process as was followed for each participant was followed to obtain the QTF and OF counts for each task. Then, QTF/TF and OF/TF were calculated for each task. Tables 3a and 3b present the counts and calculation values both by participant and by task. These are discussed in the results section.

For the last part of analysis, the number of AOIs on each SERP screenshot, as recorded in
Figure 6. AOI data exported for P01 after the addition of Task and QT FIXATIONS and the removal of all non-SERP and practice records.

The results show that 13.5% of all fixations, on average, across all SERPs, were on self-generated query terms. This means that 86.5% of fixations were on words other than query terms, or on other non-word parts of the SERP. Clearly, participants fixated on their query terms far less than they fixated on other parts of the SERP. Tables 3a and 3b list the number of QTF and OF fixations and their corresponding percent values relative to the TF for each participant.
Table 3a. The number of total fixations (TF), query term fixations (QTF), other fixations (OF), percent query term fixations (%QTF) and percent other fixations (%OF), relative to total fixations, by participant.

<table>
<thead>
<tr>
<th>Participant</th>
<th>TF</th>
<th>QTF</th>
<th>OF</th>
<th>%QTF</th>
<th>%OF</th>
</tr>
</thead>
<tbody>
<tr>
<td>P01</td>
<td>1642</td>
<td>307</td>
<td>1335</td>
<td>18.7</td>
<td>81.3</td>
</tr>
<tr>
<td>P04</td>
<td>1586</td>
<td>224</td>
<td>1362</td>
<td>14.1</td>
<td>85.9</td>
</tr>
<tr>
<td>P05</td>
<td>966</td>
<td>137</td>
<td>829</td>
<td>14.2</td>
<td>85.8</td>
</tr>
<tr>
<td>P08</td>
<td>1304</td>
<td>153</td>
<td>1151</td>
<td>11.7</td>
<td>88.3</td>
</tr>
<tr>
<td>P09</td>
<td>721</td>
<td>81</td>
<td>640</td>
<td>11.2</td>
<td>88.8</td>
</tr>
<tr>
<td>P12</td>
<td>950</td>
<td>132</td>
<td>818</td>
<td>13.9</td>
<td>86.1</td>
</tr>
<tr>
<td>P14</td>
<td>413</td>
<td>34</td>
<td>379</td>
<td>8.2</td>
<td>91.8</td>
</tr>
<tr>
<td>P15</td>
<td>1230</td>
<td>119</td>
<td>1111</td>
<td>9.7</td>
<td>90.3</td>
</tr>
<tr>
<td>P17</td>
<td>287</td>
<td>37</td>
<td>250</td>
<td>12.9</td>
<td>87.1</td>
</tr>
<tr>
<td>P18</td>
<td>748</td>
<td>104</td>
<td>644</td>
<td>13.9</td>
<td>86.1</td>
</tr>
<tr>
<td>AVERAGE</td>
<td></td>
<td></td>
<td></td>
<td>13.5</td>
<td>86.5</td>
</tr>
</tbody>
</table>

Table 3b. The number of total fixations (TF), query term fixations (QTF), other fixations (OF), percent query term fixations (%QTF) and percent other fixations (%OF), relative to total fixations, by task.

<table>
<thead>
<tr>
<th>Task</th>
<th>TF</th>
<th>QTF</th>
<th>OF</th>
<th>%QTF</th>
<th>%OF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author and Book</td>
<td>799</td>
<td>155</td>
<td>644</td>
<td>19.4</td>
<td>80.6</td>
</tr>
<tr>
<td>Brush and Cat</td>
<td>1012</td>
<td>127</td>
<td>885</td>
<td>12.5</td>
<td>87.5</td>
</tr>
<tr>
<td>Car and Road</td>
<td>1101</td>
<td>137</td>
<td>964</td>
<td>12.4</td>
<td>87.6</td>
</tr>
<tr>
<td>Citizen and Ballot</td>
<td>1239</td>
<td>137</td>
<td>1102</td>
<td>11.1</td>
<td>88.9</td>
</tr>
<tr>
<td>Eye and Page</td>
<td>765</td>
<td>59</td>
<td>706</td>
<td>7.7</td>
<td>92.3</td>
</tr>
<tr>
<td>Father and Child</td>
<td>932</td>
<td>160</td>
<td>772</td>
<td>17.2</td>
<td>82.8</td>
</tr>
<tr>
<td>Maid and House</td>
<td>772</td>
<td>104</td>
<td>668</td>
<td>13.5</td>
<td>86.5</td>
</tr>
<tr>
<td>Match and Wood</td>
<td>531</td>
<td>49</td>
<td>482</td>
<td>9.2</td>
<td>90.8</td>
</tr>
<tr>
<td>Needle and Thread</td>
<td>466</td>
<td>56</td>
<td>410</td>
<td>12.0</td>
<td>88.0</td>
</tr>
<tr>
<td>Oven and Food</td>
<td>1043</td>
<td>154</td>
<td>889</td>
<td>14.8</td>
<td>85.2</td>
</tr>
<tr>
<td>Tractor and Seed</td>
<td>1187</td>
<td>190</td>
<td>997</td>
<td>16.0</td>
<td>84.0</td>
</tr>
<tr>
<td>AVERAGE</td>
<td></td>
<td></td>
<td></td>
<td>13.5</td>
<td>86.5</td>
</tr>
</tbody>
</table>

and each task, respectively.

A Friedman test was used to investigate differences in %QTF between tasks and between participants. Significant differences were found between tasks ($\chi^2 (2, N = 10) = .026, p < .05$)
and between participants ($\chi^2 (2, N = 10) = .009, p < .05$). Post hoc tests revealed no significant pairwise differences for tasks. The post hoc test between participants found significant differences only in the extremes, as discussed next.

The variability in both measures is noteworthy. There is a greater than 10% difference between the highest and lowest %QTF values across both tasks and participants. As depicted in Figures 7a and 7b, differences are found in the extremes. Participant P01 fixated on the largest percent of query terms, with a %QTF of 18.7. The smallest %QTF was 8.2 and it belonged to P14 (See Figure 7a). The author-and-book task had the highest %QTF, with 19.40% of its fixations landing on query terms. Only 7.71% of fixations were on query terms for the task eye-and-page (see Figure 7b).

The results show that on average, participants fixated on only 4.3% of displayed query terms on each SERP. Additionally, using Friedman’s test, no significant differences were found between %AOIs for either tasks ($\chi^2 (2, N = 10) = .402, p < .05$) or participants ($\chi^2 (2, N = 10) = .483, p < .05$). In observing the variability in AOI% it was noted that some participants had very high and very low %AOI on different tasks. In order to see whether there was any systematic change from the beginning to the end of the session, % AOI was plotted by presentation order (see Figure 8). It can be seen that there was no correlation between task presentation order and %AOI for most participants. P01 and P18 are exceptions. P01 experienced a general trend of decreasing %AOI throughout the session, with a small upward trend seen in the last two tasks. P18 also experienced a general downward trend in %AOI with a spike upwards on the last task. Other participants’ %AOI values alternated between extremes for each task. Differences in average %AOI for tasks were also observed. Some tasks had a higher associated %AOI value than others. The author-and-book task had an average 7% AOI fixation rate whereas the
Figure 7a. Range of %QTF by participant

Figure 7b. Range of %QTF by task
Figure 8. Lines plotting each participant’s average %AOI changes across tasks in presentation order
citizen-and-ballot, eye-and-page, maid-and-house, and needle-and-thread tasks each had a 3% AOI fixation rate (see Figure 9). This suggests that participants were more interested in finding their own query terms for the author-and-book task than they were while searching for the citizen-and-ballot, eye-and-page, maid-and-house and needle-and-thread tasks.

The Tobii Fixation Filter used in this study to determine fixations does not use a definite millisecond threshold when determining fixations. Other eye-tracking Web search studies have used a 100ms fixation duration threshold when examining eye gaze data (Cutrell & Guan, 2007; Kules & Xie, 2011). An investigation into the millisecond duration of fixations used in this study revealed that two percent had a duration less than 100 ms. Of that two percent, eleven percent were QTFs and eighty-nine percent were OFs. These findings indicate that the inclusion of fixations with durations less than 100 ms did not affect the overall results of this study.
Chapter VI

Discussion

This study sought to discover whether searchers look at (fixate on) the query terms they submit to a search system. Eye gaze behavior, as recorded by an eye-tracker is one behavior exhibited during search that may be useful for interpreting a searcher’s relevance judgments and creating good query modification suggestions.

The results here are interesting in that they do not support previous research findings. It was found that query term fixations accounted for, on average, a small portion of overall fixations (13.5%) on SERPs across tasks and participants. This is not consistent with the findings of other studies which suggest that query terms in result captions exert a large influence over visual attention (Yue, Patel, and Roehrig, 2010; Clarke, Agichtein, Dumais, & White, 2007).

There may be several reasons why the results are not wholly consistent with the results of prior research; two such reasons are discussed in detail next.
First, the scope of this study did not include investigating the where fixations on SERPs landed if not on query terms. Only the difference between fixations on query terms and fixations on other parts of the SERP was analyzed for each SERP. However, breaking down the entire SERP on a word level to determine the types of terms fixated upon would provide more insight into the much higher instance of fixations on parts of the SERP other than query terms. Further research will investigate the other parts of the SERP fixated on and whether the rate of fixations on query terms is relatively high or low compared to the rates of fixations on other, well-defined, word-level, parts of the SERP. For instance, further investigation may reveal that participants fixated mostly on bold words displayed on the SERPs, whether the bolded words were query terms or not. Google currently bolds terms on SERPs that are related to the query terms and not just exact matches (ex: For the query *feline grooming*, Google would bold instances of the word *cat* on the SERP). In the past, Google SERPs bolded instances of the exact query terms and their stems only; this bolding was later expanded in about 2010 to include words that Google determines to be related to the words in the query (Official Google Blog, 2010). Studies suggest that bolded terms on SERPs direct visual attention (Yue, Patel, and Roehrig, 2010; Clarke, Agichtein, Dumais, & White, 2007). Since the bolded terms on Google SERPs are no longer only query terms, it may be that participants’ visual attention was largely influenced by these bolded non-query terms, explaining the relatively high proportion of fixations on words other than query terms. Future research will investigate the effects term bolding has on directing visual attention as well as the probability that a bolded word on a SERP is a non-query term. Finally, excluding non-essential words used in each query from the analysis (such as articles, prepositions, and conjunctions) may have affected the results. If no query terms were excluded
from analysis, the percent query term fixations reported in this study could only be higher; however, there is no data available on how much those values would change.

Findings of Eickhoff, Teevan, White, and Dumais (2014) suggest that after an initial query searchers use words present on SERPs to reformulate their queries. Results of the study showed that 25% of the participants’ newly added terms could be found on previously viewed SERPs. Since the query is being modified, it can be assumed that the information need was not completely satisfied (or was not satisfied at all) by the results present on the prior SERPs.

Although the Eickhoff et al. (2014) did not employ the use of eye-tracking, knowing that terms in subsequent queries can be found on previously viewed SERPs suggests that when a query is unsuccessful, searchers nonetheless spend time looking at its SERP for new words that may help them satisfy the information need with a new query. If searchers do look at unsuccessful SERPs for new terms to use in query modification, this behavior may explain the high percentage of fixations on other areas of SERPs reported in this study.

Further investigation of this study’s results to determine the locations of the visual fixations seen on SERPs before query modification will reveal whether participants fixated on words that were used in subsequent queries for the same task. If this analysis reveals that participants did exhibit this behavior, it would support the notion that eye-tracking can be useful to search systems aiming to improve service when the searcher is having difficulty expressing the information need. Eye-trackers could allow search systems to extract automatically fixated-on non-query terms from unsuccessful SERPs to be used in query modification with the goal of providing a more relevant and useful set of results.

Currently, major Web search engines, like Google, have good methods of retrieving and ranking search results and suggesting query modifications for everyday users when the
information need is common; however, but when a person’s information need is more difficult or unique, or the person is having a hard time expressing their need in terms the search system recognizes, the search engines do not perform as well and it is more difficult for searchers to find meaningful results (Aula, Khan & Guan, 2010). It is for these types of searchers and information needs that search systems need to develop better methods of understanding information needs to retrieve relevant results and provide good query reformulation suggestions. In these instances, it would be helpful for the search system to be able to make inferences about the intentions of the searcher in order to both detect unsuccessful searches and remedy them through query suggestions, query reformulation, or result re-ranking. The eye movements of searchers scanning SERPs for relevant results may reveal when the searcher is having difficulty and, when the searcher does not find relevant results, the focus of the searcher’s visual attention may provide insight into which words may be useful in query modification to meet the information need.

Although in this study query term fixations accounted for an average 13.5% of all fixations on SERPs, across both participants and tasks there was variability in the percent of query terms fixated. This indicates variability in the way people look at SERPs and a possible variability due to characteristics of tasks. These differences will be discussed in more detail next.

Aula, Majaranta, and Räihä (2005) found that searchers more accustomed to using Web search engines were more likely than those less accustomed to employ an economic result evaluation strategy, looking only at the first few results before clicking to open the Web page. Those with less experience using Web search engines were more likely to use an exhaustive evaluation approach, and looked at many results before choosing where to click or when to reformulate. In terms of the Aula et al. (2005) study, most or all participants of the study presented in this paper would have likely been considered experienced Web searchers because
the assumption can be made that the majority of undergraduate freshman possess great familiarity and experience with using Web search engines. Further analysis of the results of this study is necessary to determine what kind of search result evaluation method was used by its participants. It may make sense to evaluate the results of this and other search studies with the intention of aiding search for someone who is experienced at using an online search engine. This is useful because it is suggested that users of these services become accustomed, over time, to expect relevant results to appear at the top of the results list, making in-depth evaluation of the results unnecessary.

Another interesting approach would be to discover whether there is any difference in the number of query term fixations between novice and expert searchers (in terms of formal search training). None of the participants of this study could be considered expert searchers, so there was no distinction to be made between each participant and his or her level of search expertise.

The tasks used in this study did not evoke the same number of total fixations across all SERPs, which may be an indication that some tasks were more difficult than others. Looking at the total fixations on all SERPs for each task (see Tables 3a and 3b), it can be seen that citizen-and-ballot task evoked 1239 total fixations across all participants, while searchers performing the needle-and-thread task fixated only 466 times in total. There are several ways to interpret this as an indication of the difficulty of the tasks. First, it may be assumed that the citizen-and-ballot task was more difficult than the needle-and-thread task because participants spent more time looking at SERPs for the citizen-and-ballot task. However, the exact opposite assumption could be drawn as well. Perhaps searching for the needle-and-thread task was difficult, and participants spent a lot of time reading Web pages to find the desired information, whereas finding a suitable Web page from just the snippets presented on the SERP was easier when searching for the
citizen-and-ballot task. It may not be feasible to infer difficulty from the number and types of fixations alone. Further analysis is needed in order to determine indications of the most and least difficult tasks used in this study.

Eye-tracking could be useful in aiding search systems if a pattern of eye gaze behavior is found to correspond well with searchers’ relevance judgments. For example, if people spend less time looking at their query terms on SERPs when they judge the results to be less relevant, eye-tracking systems might detect this and display useful query suggestions by detecting which words the searcher is reading. Using eye-tracking may also allow systems to change dynamically their presentation of results by re-ranking results based on estimated relevance judgments or by changing the bolding of terms in the results based on the words fixated by the searcher. Because prior research suggests that the presence of query terms in results greatly influences visual attention, automatic result reranking suggested above could redirect the searcher’s attention when search is difficult and the searcher is not finding any relevant results. In short, using eye-tracking could be used to help search systems provide better results and suggestions when the searcher is having a hard time expressing their intended information need.

Limitations

Due to the limitations of this study, its results may not be generalizable. First, participants were drawn from a potentially homogenous pool of first year university students, all of whom were close in age. Results may be different for a different population. For example, it has been discovered, through the use of eye-tracking that older adults visually attend to different parts of Web pages than do younger adults (Romano, 2010). Additionally, while most people take a breadth-first approach to evaluating search results by looking only at the first few results (rarely more than three), some searchers evaluate a results list with a depth-first approach by looking all
the way down a results page before clicking (Klöckner, Wirschum, & Jameson, 2004). There were not enough participants involved in this study to guarantee a representative sample of the different types of searchers. Also, because of time constraints, analysis for only 10 participants was performed. Patterns of scanning and fixation may be detectable in a larger sample.

Although careful attention was paid to strict adherence to consistent methods, only the author, who was aware of the purpose of the study, performed all data collection and analysis. For this reason, following the AOI drawing process, the author’s advisor reviewed the AOI coding on every SERP. The analysis revealed a 1% average error rate in drawing the AOIs. Most of the errors involved missing a query term (not drawing an AOI around it), but in a very small number of cases, AOIs were drawn around non-query terms. An automated method of detecting and coding words in SERPS, such as the processed used by Cutrell and Guan (2007), would allow for deeper investigation of the types of words searchers scan on results pages.

The author observed that participants were focused on finding instances of the nouns in the task description, instead of simply finding a Web page that contained information on the broad topic defined by those nouns. Further analysis of the data gathered from this study could reveal whether this was the case. It is recommended that future research which intends to discover natural search behaviors take a different approach to crafting the tasks. In future work it may be useful to provide the concept word (“herding” for the dog and sheep example) and then ask participants to find information on that concept without using that word.

Finally, it must be noted that while evaluating the results of eye-tracking studies is useful for discovering behaviors and patterns, those results do not provide any insight into the motivation behind the behaviors and patterns. Future research is needed in order to understand
the intentions of the searchers and their motivations for the eye movement behaviors they employed.

Chapter VII

Conclusion

The goal of this research was to answer the question, “Do people fixate on their own query terms on a search engine results page more than anything else?” Because previous search studies indicated that participants take their search terms directly from the search task, this work took a novel approach to creating tasks that required that participants come up with search terms on their own. Tobii Studio along with the Tobii T60 eye-tracker was used to gather eye gaze data. When the total number of fixations on each SERP was compared to the number of fixations on query terms, results showed that the participants did not fixate on their query terms more than any other part of the SERP. Quite the opposite is true, with more than eighty percent of fixations falling outside of the query term AOIs. Understanding where people look on a SERP is important to understanding how people use Web search engines. Understanding the way people use search engines allows for the kind of search engine optimization necessary to make these systems effective for common searchers with difficult or uncommon information needs.
References


