EVALUATING SOURCES OF IMPLICIT FEEDBACK FOR WEB SEARCH

Xin Fu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2007

Approved by:

Gary Marchionini

Deborah Barreau

Diane Kelly

Barbara Wildemuth

ChengXiang Zhai

ABSTRACT

XIN FU: Evaluating Sources of Implicit Feedback for Web Search
(Under the direction of Gary Marchionini)

This dissertation investigated several important issues in using implicit feedback techniques to assist searchers with difficulties in formulating effective search strategies. The study focused on examining the relationship between types of behavioral evidence that can be captured from Web searches and searchers' interests. Web search cases which involved underspecification of information needs at the beginning and modification of search strategies during the search process were collected and reviewed by human analysts (reference librarians) who tried to infer searchers' interests from behavioral traces. Analysts' rationales for making the inferences were elicited and analyzed with the focus on understanding what evidence was used to support the inferences and how it was used. The analysis revealed the complexities and nuances in using behavioral evidence for implicit feedback and led to the proposal of an implicit feedback model for Web search that bridged previous studies on behavioral evidence and implicit feedback measures. A new level of analysis termed an analytical lens emerged from the data and provides a road map for future research on this topic. The study also put forward design recommendations for implicit

feedback systems based on the signals that analysts identified and the rules that they used in making inferences.

ACKNOWLEGEMENTS

My small accomplishments in this dissertation would not have been possible without the support of many people. I first want to thank my dissertation adviser, Gary Marchionini, for being a great mentor, an encouraging friend and a good role model. He has been able to make significant contributions to research and services while being so generous to people around him, especially his students. I will be indebted forever for his support, encouragement and guidance. My heart-felt thanks also go to other members of the dissertation committee, Deborah Barreau, Diane Kelly, Barbara Wildemuth and ChengXiang Zhai for their guidance and input throughout this research. All of them were so generous with their time whenever I requested a discussion or asked for feedback. It is my honor to work with such a wonderful dissertation committee.

I want to acknowledge the help from these individuals who helped with the study in various ways: Nick Boswell and Scott Adams for their help with the eye tracker, Carol Tobin, Lisa Norberg, Claudia Gollop, Amy Van Scoy, Ron Bergquist, Marian Fragola, Heidi Barry-Rodriguez, and Laura Sheble for their help with participant recruitment, Hai'ou Zhu, Miao Chen, and Ron Bergquist for pilot testing the study system, Miao Chen and Rachael Clemens for reviewing the inferences, and Lili Luo and Ron Brown for their help

with formatting the dissertation. I benefited from helpful discussions with Kerry Rodden, Peter Ingwersen, Loren Terveen, Frencesco Ricci, and Barry Smyth. In addition, the 13 reference librarians who participated in the second phase of the study deserve special recognition and thanks for their gracious help and collaboration.

My PhD experience would have been very different without the support and friendship from members of the SILS family and other friends at UNC. I want to thank the faculty and staff at SILS for their help and care through the years, and my PhD colleagues and other friends for being part of this wonderful period of my life. The longer I stay in Chapel Hill, the more I like this place and every one of you. I am very fortunate to have been a part of this big family.

Finally, I want to thank mom for preparing me for this trip and her unconditional love and support! This dissertation, coincidentally defended on her birthday, is dedicated to my mom.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

INTRODUCTION

The search engine is becoming increasingly important as a tool to acquire information and enable self-directed learning. Most of the current search engines rely on searchers to represent their information needs in the form of "queries", which usually consist of a few words, sometimes connected by Boolean operators. The transformation of a searcher's information need into a query is known as query formulation. It has been well documented that users often have a difficult time articulating their information needs and formulating effective search queries. For example, a Web search log analysis shows that users typically pose very short queries, usually between two and three words in length, when they search in Web search engines (Jansen, Spink, & Saracevic, 2000). As computers have become consumer products and the Internet has become a mass medium, searching the Web has become a daily activity for everyone from children to research scientists. When people demand more of Web services, such short queries typed into search boxes are not robust enough to meet all of their demands (Marchionini, 2006).

The reason why such short queries are often problematic is that they do not fully describe the information needs. An example of such a query is [two bedroom apartment],

when one is looking for a two bedroom apartment in her local area. The query is an underspecification of the information need because the geographical aspect of the information need is not expressed. Although underspecification has been observed in information seeking aided by intermediaries (Ingwersen, 1982), the presence of intermediaries alleviates the problem by identifying the problem context through dialog-like interactions with the searcher (Taylor, 1968; Ingwersen, 1982; Belkin, Seeger, and Wersig, 1983). In Web search where there is no human intermediary, searchers are on their own to learn from initial results, get a better understanding of their information problems as well as the information space, and adjust their search strategies or queries accordingly, when the initial query does not give good results. Lack of such analytical and modification skills inevitably leads to frustration and bad user experience. Therefore, how to design mechanisms to help searchers in this process becomes a challenge.

Different approaches have been discussed in the literature to address this problem, including interface to support initial formulation of query (e.g., Google Suggest[1] and White & Marchionini, 2007), interfaces (interactive query expansion via relevance feedback, e.g., Salton & Buckley, 1990 and Ruthven & Lalmas, 2003) and automatic techniques (pseudo relevance feedback, e.g., Mitra, Singhal & Buckley, 1998) to get feedback from searchers on initial results to support query reformulation, as well as collaborative search techniques

[1] http://www.google.com/webhp?complete=1&hl=en

which leverage the knowledge and experiences of multiple searchers with similar interests to improve the process of query reformulation and retrieval (e.g., Smyth et al., 2004; Freyne, Farzan, Brusilovsky, Smyth, & Coyle, 2007). Despite the demonstrated success of many of these efforts, there is an under-explored approach to further improving search engine performance and user experience, which is to capture and exploit searchers' interactions with search engines. This implicit approach to identifying searchers' interests removes the cost and the cognitive interruption to the user of providing feedback (Oard, & Kim, 2001; Kelly, & Teeven, 2003). It can also capture and utilize the feedback in a more active and timely fashion than most of the aforementioned techniques. Current commercial search engines prepare results largely based on only submitted queries. Even when relevance feedback techniques are used, modification of results does not happen until searchers click on suggested terms (i.e., make an explicit judgment on those terms). Although searchers examine results, and sometimes even navigate across pages of results, the information that they can see has been determined at the time of initial querying. Their interactions with the search system in the result examination process are largely ignored. This is a wasted opportunity for search engines to be more responsive and helpful. Instead, if search engines can consider the initial query as the explicit representation of the information needs, and in the meantime, consider any additional behavior that searchers exhibit as implicit indications of their interests, they should then try to capture searchers' interactions with search engines

and leverage these interactions to improve retrieval and results display. The dissertation explores several issues central to this approach.

Most of the previous work on implicit feedback was either on non-Web media, such as UseNet, or about the general use of the Web, rather than Web search. A few studies on Web search have been largely focused on click streams as evidence to tune search results (e.g., Joachims, Granka, Pan, & Gay, 2005; Shen, Tan, & Zhai, 2005a). In this dissertation, a wider variety of evidence and a wider range of granularity to support feedback and modification during Web search are examined. The key challenge for this approach to improving search engine performance and improving user experience is to find a set of evidence that (1) can be captured in a natural search setting, and (2) can reliably indicate users' interests and reflect their information needs. There has been some work in each aspect, but the results are not conclusive yet (c.f., Fox, Karnawat, Mydland, Dumais, & White, 2005; Joachims et al., 2007). The emphasis of this dissertation is to formally study the range of evidence that searcher behavior offers and understand how each kind of evidence can be useful and in what way. The work was conducted in two stages. The first stage, presented in Chapters 2 and 3, surveyed existing research on this subject and related topics to identify the most promising kinds of evidence. Chapter 2 focuses on the more theoretical research on the phenomenon of underspecification, while Chapter 3 summarizes empirical research that has been conducted to collect feedback from searchers and investigate the relationship between behavioral evidence and searcher interest. Key findings

from the literature review are summarized in Chapter 4 to form the baseline model of implicit feedback for Web search and introduce the research questions of this dissertation. The second stage, presented in Chapters 5 and 6, consisted of an empirical study on the problem, which involved two phases of data collection. The goal of the first phase (discussed in Chapter 5) was to recruit searchers who were more likely to suffer from underspecification problems and record their search sessions. The recordings captured different aspects of searchers' behavior and served as the stimuli for the second phase of data collection. In the second phase (discussed in Chapter 6), reference librarians (referred to as search analysts in the study) were recruited to examine the recordings of the search sessions that had been collected from the first phase and attempt to infer the interests of the searchers based on different subsets of the evidence. Their rationales for making the inferences were also elicited. The inferences and rationales were analyzed at three different levels and the results are presented also in Chapter 6. Based on the analysis, the dissertation concludes in Chapter 7 with a model of implicit feedback for Web search, contributing to the understanding of the relationship between the types of behavior that can be captured and searchers' interests. Common rules that were used by the analysts to make the inferences are also aggregated in Chapter 7 and lead to some design recommendations that can be applied in automated systems. These rules and recommendations reflect the long-term, more practical goal of this work: to develop self-contained, readily deployable techniques that can

capture searchers' actions in real time as implicit feedback and provide immediate search assistance.

CHAPTER 2

CONCEPTUAL BACKGROUND

This dissertation is not an isolated attempt to solve problems in the Web information access domain. It is only one part of an overall attempt to develop technologies that assist searchers, both on the Web and before the Web era, both novice and experienced, to better specify their information needs. This chapter introduces the conceptual background of the dissertation and places it in the larger context of facilitating end users' information seeking through formal search systems. Section 2.1 reviews discussions on what role should search systems play in specification of searchers' needs. Section 2.2 discusses the underspecification phenomenon, which the dissertation focuses on, and reviews literature on types and causes of underspecification. The last section, Section 2.3, overviews approaches to assisting searchers to specify their information needs.

2.1 The role of search systems in need specification

The activity by which humans look for information is a complicated cognitive process. It takes a wide variety of forms in different environments. It can take place through face-to-face conversation (with a friend, a domain expert, or a reference librarian), via

written communication (e.g., letters, emails), or by applying a formal search system (e.g., an OPAC system, a search engine). In any of these situations, information seeking consists of a communication between the information seeker and the information resource through some channel and sometimes via intermediary. As Web search is an example of people seeking information through machine-mediated communication, it shares some similarities with IR situations involving other formal systems. Examples of formal systems include: libraries, research firms, government agencies, electronic networks, and the growing collection of information services that make up the information industry (Marchionini, 1995).

The need specification process is an integral part of the information seeking process in which searchers transform their information needs into formalized requests, or queries. Many contemporary models on the information seeking process (Belkin, 1993; Belkin, Cool, Stein and Thiel, 1995; Ingwersen, 1992; Marchionini, 1995; Saracevic, 1996, 1997) depict the complexity in the need specification process. Complexity firstly lies in that specification takes place under the impact of a variety of user factors, such as intention, beliefs, and knowledge, and system factors, such as linguistic and pragmatic constraints. Therefore, the original information need is subject to modification during this process. Secondly, the information need itself may evolve during the information seeking process, because one's conception of the information problem is dynamic and subject to change during one's interaction with IR systems.

Information seeking models can also be examined to inform the role of IR systems in the need specification process. One perspective is to view the information seeking process as an interaction between the searcher and the information resources; then the role of the IR system is to "facilitate" or "mediate" the interaction. This "mediation" viewpoint is probably the dominant one with regard to the role of IR systems. Despite this, there have been few descriptions on how the mediation actually functions. Some noteworthy exceptions are Belkin, Seeger, and Wersig (1983), Belkin (1993) and Saracevic (1996). They suggest possible ways to design the mediation.

Belkin, Seeger, and Wersig (1983) outlined 10 functions of an information provision mechanism and argued that the primary effort of such a mechanism is to understand characteristics of the user, such as where in the problem treatment process is the user located (problem state) and the user's intentions, situation, preferences and beliefs. They further suggested that the understanding is itself to be gained through dialogue-like interaction with the user. Belkin (1993) argued that interaction with texts should be the central process of IR and that only through interaction with texts can users come to understand and learn about their information needs. Therefore, the system's role as the intermediary should be played through promoting users' interaction with texts. Pennanen and Vakkari (2003) and Kelly and Fu (2006) have found empirical support for Belkin's argument. Saracevic (1996) provided a good summary of the role intermediaries play in the library setting. As both Web IR and library reference are mediated IR interactions,

Saracevic's analysis should inform us of the role a Web IR system could play to mediate the communication. Saracevic states:

> "The roles that intermediaries play can also be decomposed into levels. On the surface level, intermediaries use their mastery (knowledge and competence) about IR systems – contents, representations, metainformation, techniques, peccadilloes – not mastered by users. This is used to provide effective interaction with the system on the surface level. But on the deeper or cognitive level, **intermediaries also provide clarifying and diagnostic aspects**. They provide help in defining the problem, focusing the question, incorporating the context, and other aspects that enter into user modeling. As the interaction and search progresses they also may suggest changes in problem or question definition. All this plays a critical role in selection of search aspects on the surface level: files, terms, tactics, attributes etc. Through their professional training and experience professional intermediaries become highly skillful in user modeling (which is on a deeper level of interaction), and on translating that into the surface level of interaction with a system. (Similarly, doctors and other professionals become through experience skillful in diagnosis, which then they use in treatment.)" (pp. 7-8)

This is one of the few statements that can be found in the literature which emphasize intermediaries' role in helping define users' information problems and incorporate the contexts. At the practical end, it has also been noticed that current search engine interfaces

provide little support for the user searching process (Freund & Toms, 2002). Without the benefit of human intermediaries skilled in eliciting the information need from the user and in designing a well-formed query and search strategy, today's Web users must negotiate the information seeking process directly with the search engine. The "mediation" role of IR systems dictates the need for Web search engines to get involved in need specification by helping searchers define and redefine their problems and translate their changes in problem or question definition into effective search strategies and search terms.

Another perspective on the role of IR systems in the need specification process is represented by Belkin, Cool, Stein, and Thiel (1995), in which they argued that "supporting, and *taking advantage* of the interaction of the user with the other components of the IR system is crucial for effective IR system design" (p. 379). The notion that IR systems could and should take advantage of the interaction of the user with other components of the IR system provides theoretical foundation for the design of interactive features which are not part of the normal interaction that the user is engaged in but specifically introduced to gather user feedback as well as algorithms which actively monitor and learn from users' interaction and provide feedback on search strategies.

In sum, the role of IR systems in the information seeking process focuses on mediating the interaction between the user and the system. Two levels of mediating roles have been suggested in the literature: to support and to take advantage of the interaction. Although there is an increasingly large body of literature on empirical studies that take

advantages of interactions with the user, this notion is under emphasized from the theoretical perspective. In most information seeking models, the role of the system is still somewhat limited to passively executing queries that are submitted to the system and presenting results, which does not fully reflect the operation of state-of-art IR systems. It is high time that research in IR system's support for need specification be advanced from both practical and theoretical fronts.

2.2 Summary of studies on information need underspecification

Transforming an information need into a formal representation is not only a complex process, but also a challenging one. This section first reviews the types of challenges involved in need specification in a search process. It then focuses on one consequence that results from the challenges, the underspecification of information needs (others being wrong search tool, wrong search terms, over-specified queries, wrong syntax, etc.), and discusses the different types of underspecification, some common causes and types of searchers who are more likely to underspecify. The discussion informs the recruitment of searchers and selection of search tasks in the first phase of the empirical study.

2.2.1 Challenges in need specification

Two early articles on information need specification put forward hypotheses on how people translated their information needs into queries. These hypotheses laid the framework

for studies on challenges in the need specification process and had significant implications for design of mediated search systems.

Taylor (1968) studied the question negotiation process in the library reference interview situation. He suggests that an inquiry is not a single event, but instead a dynamic process in which the inquirer changes the question as he or she searches for a result. He pointed out that information need is a personal, psychological, sometimes inexpressible, vague and unconscious condition. He articulated four levels of information need that an individual passes through before he or she makes formal encounters with an information system or the services of an information professional. These levels are: visceral need, conscious need, formalized need, and compromised need. The visceral need is an actual, but unexpressed need for information. It is a vague sense of dissatisfaction, but it is hard to express in words. When the need becomes conscious, the inquirer forms a mental description of the need. The next level is formalized need. At this level, the inquirer comes up with a formalized statement of the need and defines boundaries of the question. The most specified level is compromised need. At this level, the inquirer recasts the question in anticipation of what the system can deliver and presents the question to the system. The queries that users submit to search engines are at this highly specified level.

This distinction of four levels of information needs has important implications for IR system design. As the queries that searchers present to IR systems reflect the compromised level of information need, or in Taylor's words "the representation of the inquirer's need

within the constraints of the system and its files" (p. 183), their models of how an IR system works may bias what they enter into the system. This means that queries submitted to the IR system can be partial or distorted representations of information needs. The skill of the intermediaries (the IR systems in this context and the reference librarian in Taylor's context) is thus to "work with the inquirer back to the formalized need, possibly even to the conscious need, and then to translate these needs into a useful search strategy" (p. 183).

Belkin's article (1980) stands as another seminal work that studied the need specification progress. He viewed information needs as originating from an anomalous state of knowledge (ASK), the realization that one lacks the knowledge to solve certain problems. He pointed out that users may have initial difficulty in specifying or even explicitly recognizing what is wrong, and especially in recognizing and specifying what is necessary to make things better. Instead of classifying information needs into specifiable and not specifiable, he placed them in a continuum of specifiability, from needs which are precisely specifiable or nearly so (i.e., when the user knows exactly what is necessary to satisfy the need) to needs which can not be specified or can be specified only very vaguely (i.e., when the user is conscious of a need but does not know what information would be appropriate to satisfy it). His perhaps more important contribution was the contemplation on factors which account for non-specifiability of information needs. He broke down the query formulation process into two steps. First, the user needs to pass the cognitive spectrum of information needs to understand what the problem is and what will be needed to solve the problem.

Then the user needs to pass the linguistic spectrum to express the need as a formal request. Difficulties in either step can result in ill-specified queries: users may not clearly realize their information needs, or they may not have the capability of expressing their needs appropriately in the system's terms.

Belkin's work reinforces Taylor's hypothesis that queries (the compromised need) can be partial or distorted representations of the actual information need (the visceral need or the conscious need) and attributed the non-specifiability of information need to cognitive and linguistic reasons. The non-specifiability of information need explains why the sorts of IR systems based on the "best match" model are inappropriate. It also points to the need for an IR system to not only process the submitted query, but also design mechanisms to help the searcher overcome difficulties in understanding and expressing the need.

Another noteworthy work on the need specification process was done by Freund & Toms (2002), in which they discussed need specification in the context of the query process. They suggested that queries are shaped by complex factors, such as situation, topic, and participants' motivation. These factors act as filters that operate within the cognitive space of the searchers between their information needs and the actual queries they submit (Figure 2.1).

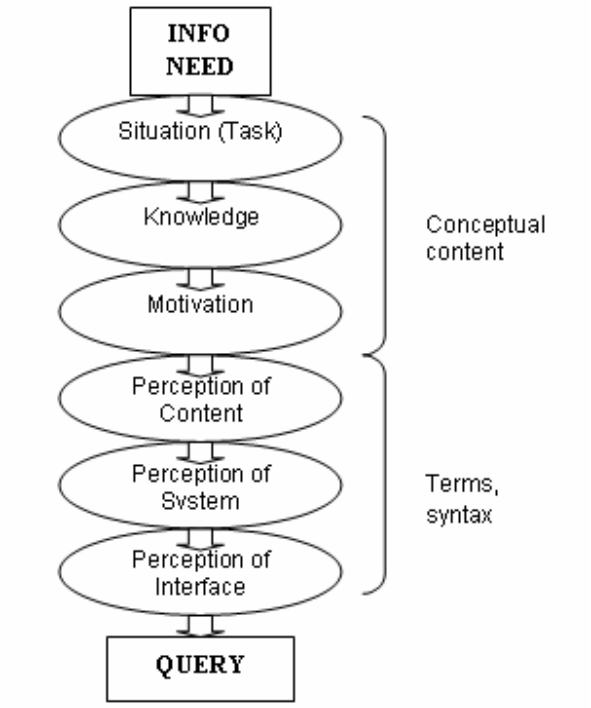


Figure 2.1. Query filters proposed by Freund and Toms (2002, p. 74)

This work, together with Belkin's work, suggests the range of factors that have an influence on the specification of information needs. As will be examined next, these factors can become challenges to searchers and directly or indirectly cause the underspecification of information needs.

2.2.2 Causes for underspecification

Current search interfaces are mostly designed to support analytical search strategies (Marchionini, 1995), which assume that searchers have well defined information needs and require searchers to communicate the needs in terms of queries. Unfortunately, this also means that searchers are less likely to succeed if they lack the knowledge about the field or if the search task requires browsing and exploration (White, Kules, Drucker & Schraefel,

2006). If searchers are in these situations, but use current search engines to look for information, it is very likely that they can not specify their information needs exactly. Realizing this constraint, searchers sometimes choose to apply a mixed strategy: they specify some parts of their information needs and use analytical search to acquaint themselves with the domain, the vocabulary and the resources before they formalize their search strategy, or to find the website that contains the information they may need and browse from there. Consider this example: a college student has no prior experience in programming, but is interested in taking a programming course to start learning. She has no idea which course at her university will be suitable for her, but she knows that programming courses are normally offered in the computer science department. So, she decides to start the information seeking by searching for the homepage of the computer science department at her university. The query she uses is something like [computer science department ABC university]. Just looking at this query will naturally lead to the conclusion that it is an underspecification of her needs, since concepts like "programming" and "beginner course" are missing from the query. However, from the searcher's point of view, this query serves her intermediate purpose well. From the department homepage, she can go on to the course offering page and get a list of courses and their descriptions. She can then get some idea on what is available and decide to either search or browse to get detailed information about a particular course that she would like to take.

The possibility of underspecification as a strategy, as applied in the above example, adds an additional level of complexity when studying why people underspecify. This also means that it is unreliable, if not impossible, to study the query alone out of its context. When research is designed to study the underspecification phenomenon, the researcher needs to have a good understanding of subjects' search tasks (or in the cases of assigned tasks, subjects need to have a good understanding of what is required) and plans to collect qualitative data on their search strategies and motivations for each move, instead of simply relying on literal analysis of queries in isolation.

Among the causes for underspecification not as part of the search strategy, the first is searchers' cognitive limitations. Geisler (2003) suggests that among the four levels of information needs described by Taylor (1968), it is only at the comprised level that one can express the problem in the form of a query required by a search system; the other three represent varying degrees of awareness of the problem, none of which is developed enough to enable the searcher to enter an effective query to a traditional information retrieval system. Due to the cognitive limitations, searchers have difficulties in understanding the information problem and figuring out what is required to resolve the problem.

There have been studies on what type of people and under what conditions people are more likely to have difficulties understanding their information problems. The general observation is that searchers are more likely to have cognitive difficulties when they search in a new area. Belkin (1980) gave two examples of this type of situation. One example is a

researcher entering a new field or problem area who needs to know how his or her knowledge relates to the new problem. The other example is a person entering a new social structure, such as a new city, country, or job, who needs to know how to get on in the new situation. In both cases, a problem is recognized, and it is recognized that information might be necessary to resolve the problem, but precisely because of the inquirer's lack of knowledge about the problem area, it is impossible to specify what would resolve it. Incomplete queries and queries missing some aspects of the information needs are usually consequences of cognitive difficulties.

The second cause of underspecification is searchers' linguistic limitations. Even when information problems are well defined and searchers understand what is required to resolve the problems, they may have difficulties in presenting the problem to search systems.

Linguistic difficulties can be broken down into several levels, some of which can result in the underspecification problem. At the first level is the natural challenge to express one's thoughts in language. This is not only observed in IR interactions, but also in other types of human-human communication and human-computer interaction. The closest example comes from studies of patron-intermediary dialogue during reference interviews in libraries. Ingwersen (1982) studied the search procedures in the library and found a tendency for library patrons to simplify even well defined information needs when expressing search requests to librarians. He noted that "user need seem often to be presented

as a label which may create ambiguity problems" (p. 165). This "label effect" strips the information need of its context, which the librarian must try to identify. It is easy to imagine that difficulties in expressing information needs exacerbate in Web search where end users conduct searches directly without the help of intermediaries.

The second level of difficulty results from the differences between the searcher's vocabulary and the author's (domain specific) vocabulary as well as the organization of the languages (syntax). It is well known that people often use different words to describe the same things (Furnas, Landauer, Gomez & Dumais, 1987). Bennett (1972) described the problem that users often have with communicating their information needs to systems, because users are forced to communicate using the system's vocabulary and not their own. Lacking the knowledge about system vocabulary and syntax, users may express information needs in a way that is not supported by the system. Bilal (2000) observed middle school students' searching behaviors and noted that they made all kinds of errors, including using natural language queries in a system that did not support natural language (Yahooligan!) and using vocabulary that was either too broad or too specific. Toms and Bartlett (2001) noted that expression of information needs to an IR system requires not only the selection of appropriate words, but also knowledge of system specific attributes, e.g., truncation and phrase specification. Wang and Pouchard (1997) analyzed queries submitted to a university Website and found that users had difficulties with the syntax and semantics of different search engines; more than 30% of the searches resulted in zero-hit outcomes. They

concluded that improvement of search engines (automatic error correction and context-sensitive help) can eliminate some of the errors. Freund and Toms (2002) found that term selection, rather than syntax, was the main issue in query formulation and that participants of their study did not seem to have clear strategies for this process. Rather, it was based on experimentation, and was significantly influenced by their perceptions of what resources were out there in the Web space.

Vocabulary problems are most likely to cause ambiguous queries. If a searcher submits a query like [New York pizza], it is very likely that she is not aware of different interpretations of the query that can be made by the system, so she does not take the effort to further qualify the query (e.g., New York pizza restaurant, or New York style pizza). If a searcher is familiar with the system vocabulary space and aware of the ambiguity inherent in the language, she can choose to disambiguate the query before submission. Examples of such queries are [Java programming], [Java coffee] and [Java island Indonesia].

Related to the vocabulary problem is the phenomenon of misspelling. Freund and Toms (2002) found that about 10% of the queries in a search study using Google contained some type of errors and these queries formed an important type of cases for query reformulation. They further noted that several of the longest chains of reformulated queries were the result of misspellings, despite the fact that Google prompted the participants with the correct spellings. They suggested that "one difficulty is that many participants did not notice the Google prompt, and another is that some of those who saw it did not believe that

the suggestion was correct" (p. 80). Schaefer, Jordan, Klas and Fuhr (2005) corroborated this argument. They found that even during known item searches, users still need an average of four to five queries to find the information they were looking for. They noted that one major cause of the need for repeated querying is faulty queries and that most of the "errors" fall into the category of misspellings or typographical errors. However, misspelling normally would not lead to underspecified queries; they are simply erroneous queries.

Given the prevalence of linguistic problems, it is natural to wonder who are more likely to have such problems. Schaefer et al. (2005) note that searchers are more likely to have vocabulary problems when they are involved in topical searches (as opposed to known item searches). The vocabulary problem will add to the uncertainty that users already have due to their information problems. Low level errors, like spelling mistakes or inadequate Boolean logics, can be explained as an expression of uncertainty or fear when starting a search. In addition, some studies show that experienced searchers develop and use queries more effectively than non-experts (Lazonder, Biemans & Woepreis, 2000; Lucas & Topi, 2002). In some other studies, domain knowledge is shown to be also very important. Hölscher and Strube (2000) compared the search habits of novice and expert Web searchers and found that successful Web searches relied on a combination of experience and domain knowledge. In some cases, novice searchers with good domain knowledge compensated for their lack of query formatting skills with greater verbal creativity. Marchionini (1989) also found that system expertise is of less importance to information seeking than is domain

expertise. In sum, these studies suggest that searchers with complicated, less defined information needs (as in topical searches), less experiences and less domain knowledge are more likely to encounter language problems.

If searchers' initial queries do not retrieve results that are relevant to their needs, then they are often faced with the problem of trying to figure out how to reformulate their queries. This can be particularly problematic when searchers are searching in areas about which they have little or low familiarity (Kuhlthau, 1993; Vakkari, 2000). This is also a challenge to novice searchers and it often takes them more efforts to recognize the problem and resolve it. Freund and Toms (2002) describe the complexity of the process of iterative query construction and reformulation. In this process, searchers have to draw upon their own knowledge and skills, information in Web resources, search results, and system feedback to negotiate a path through the search. These are the qualities and skills that novice searchers normally do not possess. The author's own experience of running and observing search studies also suggests that (self reported) novice searchers are less responsive to system feedback and less skillful in adjusting their search strategy or modifying search terms based on result inspections. They are also more likely to miss system prompts, such as Google's suggestions for correct spellings.

To summarize, this subsection analyzed the possible causes of underspecification. Searchers sometimes choose to underspecify their information needs as they break down the original information need into intermediate steps, but otherwise, underspecification may

mainly be caused by searchers' cognitive and linguistic limitations. In general, people are more likely to underspecify when they search in a new field and have a complicated information need involving multiple aspects. Moreover, novice searchers are more likely to suffer from underspecification problems than experienced searchers due to their lack of knowledge about the system vocabulary and/or system syntax and experiences in reformulating queries based on system feedback.

2.3 Assisting searchers with the underspecification problem

All the work reviewed in the last section pointed to the need for search intermediaries to help searchers overcome the difficulties in formulating effective queries. Three major approaches to providing such help are summarized below: designing search supportive interfaces, providing search recommendations via social search techniques, and collecting feedback from the searchers.

There has been a significant body of research that aims to improve the current interfaces to better support searchers' articulation of information needs and formulation of queries. One thread of efforts provides help to searchers when they formulate initial queries. Examples of these efforts include the Google Suggest function and a recent study by White and Marchionini (2007) both of which provide query expansion options when search engine users type their queries. Secondly, when results are presented, the interface needs to help the searcher form a mental model of the result set and better understand what items are

available and how they are organized. This involves research on search results visualization (Eick, Steffen, & Sumner, 1992; Shneiderman & Plaisant, 2004; Tanin et al., 2000; Lin, 1997; Furnas, 1981), overview/preview (Greene, Marchionini, Plaisant & Shneiderman, 2000; Geisler, 2003) and surrogation (Boekelheide et al., 2006).

A second approach to providing query formulation help is through collaborative searching techniques (Smyth et al., 2004; Freyne et al., 2007). These techniques recommend queries that have been used by past searchers with similar interests and are believed to be similar to the current query (presumably because they are about the same topic) or automatically expand the current searcher's query with terms from previous, similar queries or terms from documents retrieved in response to these queries.

The third and the most related to the dissertation approach to helping searchers overcome difficulties in formulating queries is to collect feedback from searchers and use search algorithms at the backend to incorporate the feedback to improve retrieval. Ingwersen (1996) pointed out that users often know additional information about their information needs beyond what they typically communicate to information systems. So, when the search is done through a search intermediary, such as a reference librarian, she asks questions to understand the user's information need and users her knowledge about the system to formulate appropriate queries. When it comes to the Web search where searches are done by end users, search engines should take the similar role and collect feedback from the searchers to better understand their needs. Moreover, since research has demonstrated

that searchers' information needs evolve during the search process (Saracevic, 1996; Bates, 1989), it is important for search systems to get feedback from the searcher constantly through the search.

While specific techniques to collect and use the feedback will be reviewed in the next chapter, it should be mentioned that user feedback can be collected to achieve different goals and the specific goals that a study aims to achieve have significant impact on what types of feedback should be collected and which methods are appropriate. Marchionini and Mu (2003) suggested three types of goals to conduct user studies: needs assessment studies to understand problem contexts and inform design, usability tests to assess specific design decisions, and studies of user behavior that use novel interfaces as stimuli. Atterer, Wnuk, and Schmidt (2006) argued that the main application of tracking users' behaviors has been usability tests of websites, but with a tracking approach that is flexible enough, it is also possible to use the tracking for constant evaluation of live websites, profiling users and monitoring their interactions with websites. Methods of collecting feedback should be selected based upon the goals. For example, methods that are suitable for usability tests are not necessarily appropriate for needs assessment because when needs assessment studies are carried out, the system has usually not been designed yet, so some usability testing methods such as eye-tracking are not applicable. The focus of this dissertation is to capture and exploit searchers' interactions with the search system to infer searchers' interests and

improve retrieval. Therefore, when methods to collect searcher feedback are reviewed in Section 3.3, emphasis is put on methods that capture searchers' behaviors.

CHAPTER 3

RELATED WORK

There are two ways that feedback can be collected from searchers. One way is to explicitly ask the searcher. This is often done in Cranfield-style evaluations of information retrieval systems, and has been quite useful in developing and tuning information retrieval algorithms. The other way, implicit feedback, observes searchers' behavior and infers their interests from their interactions with the system. The focus of this dissertation is the implicit approach to collecting feedback from searchers, so after a brief review of literature on explicit feedback in Section 3.1, a more detailed review will be provided on implicit feedback literature in Section 3.2. Finally, as an important aspect of implicit feedback studies, methods to capture the feedback from searchers (in this case, their behaviors) are summarized in Section 3.3.

3.1 Research on explicit feedback

Relevance feedback is a classic IR technique that supports the iterative development of a search query using examples of relevant information (Salton & Buckley, 1990). It is used after an initial set of documents (or, in the Web environment, Web pages) have been

retrieved. The predominant viewpoint was that by providing searchers with terms used to index documents, they would be equipped with a more appropriate vocabulary with which to formulate queries; all searchers needed to do was to select the most appropriate terms from the display. In its simplest form, searchers are presented with and requested to examine the top ranked documents and identify which of these documents are relevant. Keywords from these selected documents are then extracted and added to the searcher's query or used to re-weight existing query terms. Since searchers are involved in the process to make explicit judgments on the relevance of documents, the technique is sometimes called "explicit relevance feedback".

In addition to asking searchers to judge the relevance of documents, some relevance feedback techniques work at the passage or term/phrase level. Passage level relevance feedback is similar to that at the document level, except that potentially relevant document snippets, instead of the entire document, are displayed for feedback. Those passages are either extracted from the documents by some algorithms (e.g., Shen & Zhai, 2004) or selected out of the documents by the user (c.f., Harper, Koychev, Sun & Pirie, 2004). Term/phrase level relevance feedback presents the users with certain (ideally the most discriminative) terms or phrases extracted from potentially relevant documents and adds those terms or phrases selected by users to the query. Compared with document or passage level relevance feedback, the term/phrase level feedback reduces the noise introduced by irrelevant terms, but has the disadvantage of losing the context in which terms/phrases

appear. Without appropriate context, it might be difficult for users to understand how terms are used, why terms are suggested, and how such terms might be used to improve retrieval. Previous research does not provide a clear idea about how term context will affect user behavior and retrieval. Joho, Coverson, Sanderson and Beaulieu (2002) presented users with two types of displays for query expansion, list and menu hierarchy. They found no significant differences in retrieval performance across display types, although subjects selected about 4 more terms on average from the menu hierarchy. Subjects in this study further stated that they believed that the menu hierarchies gave them a better idea of the contents of retrieved documents. Kelly and Fu (2006) also compared the effectiveness of presenting relevance feedback terms in isolation versus in sentence context, but the results were not conclusive.

There are also some studies (e.g., AbdulJaleel et al., 2003) which use clustering techniques to generate documents or document snippets for relevance feedback. Instead of determining the documents merely by their rankings at the initial result set, it applies a clustering algorithm to the retrieved documents and obtains clusters. Then, the centroid document or the highest ranking document in the cluster (or part of it, such as passages, terms or phrases) is used to represent the cluster and displayed for relevance feedback.

Explicit relevance feedback techniques have their limitations. In particular, empirical studies have led to the general finding that relevance feedback features are not used. For example, participants in a series of studies by Belkin et al. (2001) rarely used

relevance feedback features and often commented on the quality of terms suggested by the system. Belkin et al. speculated that users may not have used relevance feedback features in these experiments because they were involved in complex information-seeking tasks in a novel environment, and may not have had additional cognitive resources available for learning and experimenting with features. Further evidence has shown that users often have a difficult time selecting the best terms for query expansion, if they are willing to select them at all. In many cases, users do not understand why certain terms have been suggested and in many other cases, the terms which the system suggests are not necessarily the best. For instance, in a study of simulated interactive query expansion, Ruthven (2003) demonstrated that users are less likely than systems to select effective terms for query expansion. Ruthven found some potential benefit of term relevance feedback if the best terms were used in query expansion, but went on to note that users are unlikely to select these terms because of problems with current relevance feedback interfaces. In a Web-based study, Anick (2003) found that users made use of a term suggestion feature to expand and refine their queries. However, this did not result in improvements in retrieval performance.

Despite so, there seems to be a recent trend in increasing use of relevance feedback techniques on the Web. For example, Google and Yahoo display "related terms" at the bottom or top of results list pages for certain queries (e.g., [Michael Jackson] for Google

and [web hosting] for Yahoo in Figure 3.1. As these features are still being evaluated, it is unclear how these suggested queries are selected and ranked[1]. Given the implementation of Google Suggest discussed previously, it is reasonable to contemplate that the term suggestion features in search engines are more likely to be based on statistical information of term occurrence captured in the query log, rather than analysis of web pages that are retrieved (as in the case of document based relevance feedback).

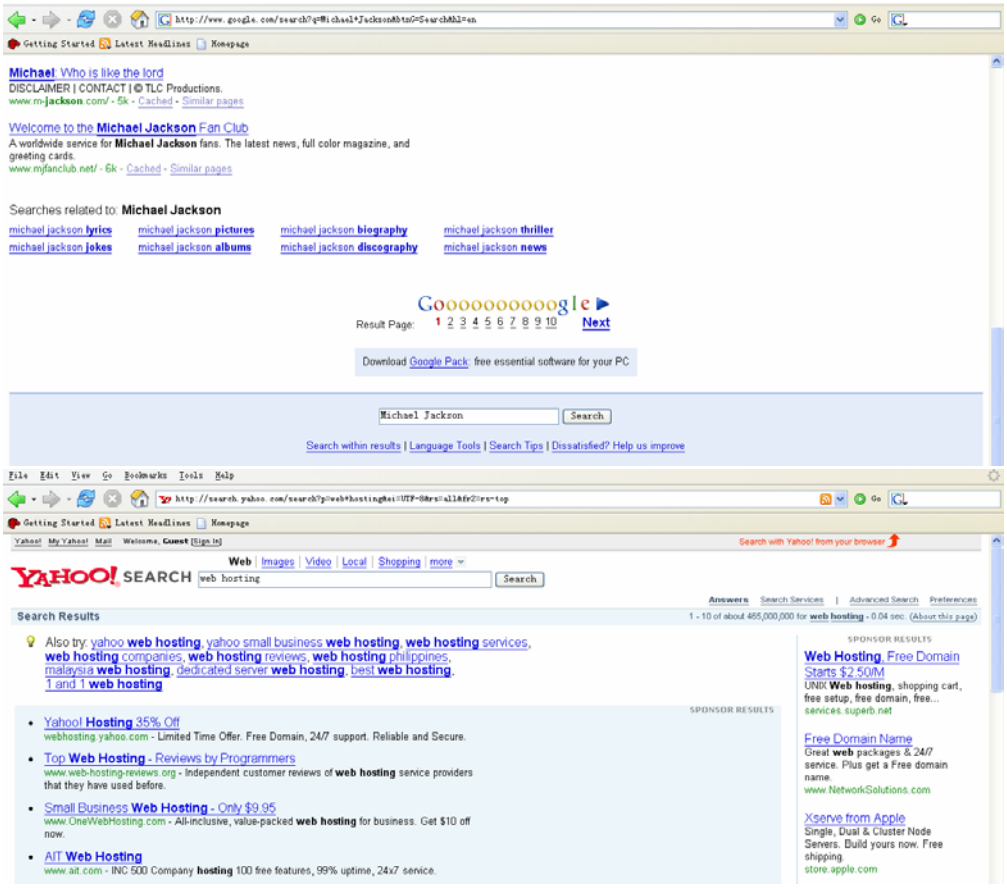


Figure 3.1. Examples of Google (above) and Yahoo's related term suggestion features

[1] http://searchengineland.com/070115-173039.php

In the meantime, some variants of the relevance feedback technique have been widely implemented in Web search engines and were reported to be useful (e.g., Rappoport, 2003). For example, Google performs spell check for queries entered by users and suggests correct spelling if misspellings are suspected. Although it is designed as an error prevention mechanism, it can also be viewed as a variant of the relevance feedback technique because it offers potentially related (in this sense, correct) query terms and allows users to give feedback on whether the suggested spelling is really what they intended to use.

3.2 Research on implicit feedback

Instead of relying on searchers to make explicit judgments on the relevance of documents or terms, implicit feedback techniques unobtrusively watch searchers' natural interactions with the system and obtain information about their interests from the behaviors. The primary advantage to using implicit techniques is that such techniques remove the cost and the cognitive interruption to the searcher of providing feedback (Nichols, 1997; Oard, & Kim, 2001; Kelly, & Teeven, 2003). They have been described as a promising approach to identifying user preference and improving retrieval performance and user experience (Kelly, & Teeven, 2003; Fox et al., 2005).

The main goal of this section is to survey the literature that discusses studies on observable behaviors as potential indicators of searchers' interests. Two types of studies are reviewed: studies that built frameworks to classify behaviors that can be used for implicit

feedback and studies that examined the relationship between behaviors and searchers' interests empirically.

3.2.1 Classifications of observable behaviors as implicit feedback

Nichols (1997) provided the first classification of observable behaviors as implicit feedback (Kelly, 2005). He suggested 13 types of implicit rating information, including purchase (price), assess, repeated use (number), save / print, delete, refer, reply (time), mark, examine / read (time), consider (time), glimpse, associate, and query. This classification, although very coarse, served as the foundation for subsequent works that developed frameworks of observable behaviors, as reported in Oard and Kim (2001), Claypool, Le, Waseda, and Brown (2001) and Kelly and Teeven (2003).

Claypool et al. (2001) discussed both explicit and implicit interest indicators. They classified interest indicators into seven categories: explicit, marking, manipulation, navigation, external, repetition, and negative interest indicators. The corresponding behaviors are summarized in Table 3.1.

Table 3.1. Summary of interest indicators discussed in Claypool et al. (2001)

Type of interest indicators	Behaviors
explicit interest indicators	rate on a scale
marking interest indicators	bookmark, delete bookmark, save, email or print
manipulation interest indicators	cut and paste, open new browser, search within page, or scroll
navigation interest indicators	click links
external interest indicators	heart-rate, perspiration, temperature, emotions and eye-movement
repetition interest indicators	spend time, lots of scrolling, revisits
negative interest indicators	absence of above

Oard and Kim (2001) presented a framework for modeling the content of information objects based on observation of how users interact with those objects in the course of information seeking and use. They categorized potentially observable user behaviors in two dimensions. One was the type of behavior. In this dimension, four categories were identified: examination, retention, reference, and annotation. The other was the minimum scope at which each behavior can be observed. The levels were segment (portion of an object, e.g., a paragraph or a screen), object (complete object, e.g., a document), or class (collection of objects, e.g., multiple documents in a folder). Note that the "minimum scope" indicates the smallest unit normally associated with the behavior, so it is possible that behaviors may have analogues at larger scopes (e.g., viewing an entire document instead of viewing a screen), but not normally at smaller scopes (e.g., purchasing a paragraph instead of purchasing the entire document). Kelly and Teeven (2003) directly

built on and extended Oard and Kim's (2001) classification, adding the "create" type of behavior and a few more behaviors. The extended classification is displayed in Table 3.2.

Table 3.2. Classification of behaviors that can be used for implicit feedback (from Kelly and Teeven, 2001, p. 19)

	Minimum Scope			
Behavior Category		**Segment**	**Object**	**Class**
	Examine	View Listen Scroll Find Query	Select	Browse
	Retain	Print	Bookmark Save Delete Purchase Email	Subscribe
	Reference	Copy-and-paste Quote	Forward Reply Link Cite	
	Annotate	Mark up	Rate Publish	Organize
	Create	Type Edit	Author	

When applied to the Web, records of the "select" behavior are often called "click streams", which can be captured via server-side logging or video logging. As the minimum scope for "view" is a portion of a document, it can only be captured by eye-tracking, and arguably inferred from mouse movements. As Oard and Kim (2001) pointed out, "observing behavior at a scale below that of a complete object might provide more precise evidence of

the user's intentions than object-scale observations alone, but at the cost of a somewhat more complex data collection effort" (p. 41). "Scroll", "find" and "query" can take a searcher to a segment of a document. They can be detected at the client side and recorded via logging. From the searcher's perspective, examination behaviors involve no cost except for time. Therefore, time is often used as a measurement of interest when examination behaviors are studied.

Cost also plays an important role in determining the extent to which retention behaviors reflect searchers' real interests. "Bookmark", "save", "email" (to oneself) and "subscribe" only cost computer resources, so they are less strong indicators of searchers' interests than "print", which involves more expensive resources (ink and paper, in addition to the printer), and "purchase", which directly involves spending money, thus offers "extremely strong evidence of the value ascribed to an object" (Oard & Kim, 2001, p. 41). However, "print" can also serve other purposes and do not necessarily reflect searchers' interests. As Oard and Kim (2001) noted, people sometimes print documents merely to facilitate examination because paper still has many advantages over electronic displays. "Delete" is a reliable indicator of relative interests, when retention is a default condition, as in email systems, but it does not seem to be relevant to a searcher.

When searchers exhibit the "reference" and "annotation" types of behaviors, they relate the information to their tasks or add to the value of an information object. These behaviors, though involving little material resources, require additional cognitive resources.

Kelly and Teeven (2003) treated "email" as a retention behavior; however, it can be argued that although emailing to oneself is a retention method, just like "bookmark" and "print", emailing a Web page to other people constitutes a recommendation or endorsement[1]. So, "email" can also be classified under "reference", in a similar way as "forward". Since all "reference" and "annotation" behaviors involve assessing the values of the information objects, and/or some kind of internalization, they can also be good indicators of searchers' interests.

It should also be noted that both Oard and Kim (2001) and Kelly and Teeven (2003) studied implicit feedback provided by people's general information behaviors when they use a variety of computer applications, such as word processing software, email clients, and Web browsers. Some types of behaviors in their classifications do not directly apply to the discussion here which focuses on the more specific Web search behaviors, typically through using a search engine in a Web browser. For example, when discussing the "link" behavior, Oard and Kim (2001) stated that "hypertext links from one Web page to another and bibliographic citations in academic papers also create links from a portion of an object (characterized, perhaps, by some neighborhood around the link itself) to an entire object" (p. 42). From this perspective, the "link" behavior is beyond the scope of this dissertation, except in very special situations such as someone building a website, searching and finding

[1] In a later article, Kelly (2005) refers to Kelly and Teeven (2003) and notes that "… email describes the behavior where one finds a useful document and emails it to oneself" (p. 172), but this was not made clear in the original article.

a good webpage related to its content, and deciding to create a link to it. For the same reason, "cite" is not pertinent, either. Behaviors categorized as "reference" and "annotate" in Table 3.2 were largely not searchers' behaviors, but those of Web page creators. Lastly, Kelly (2005) pointed out that "rate" is typically used as explicit feedback, so it will not be included in the discussion on implicit feedback hereafter.

The last type of behaviors, "create", was not included in Oard and Kim (2001)'s classification, but added by Kelly and Teeven (2003). However, Kelly and Teeven (2003) did not give specific examples of "type", "edit" and "author". From their statement that

> "The 'Create' behavior category describes those behaviors the user engages in when creating original information. An example of a 'Create' behavior is the writing of a paper." (p. 19),

it seems these behaviors are not typically performed by searchers either. Neither are they relevant to the ultimate goal of predicting searchers' interests. So, they are beyond the scope of this dissertation.

In addition to the observation that Oard and Kim's (2001) and Kelly and Teevan's (2003) classifications included explicit feedback behaviors and behaviors not directly related to information seeking, another challenge in applying those classifications to the discussion here is that the scope of the implicit feedback discussed in their works was primarily focused on content (Jansen & McNeese, 2005). This makes it hard to categorize some behaviors related to the search interface, no matter what types of contents are

displayed at the interface. Examples of such behaviors include eye movements and mouse movements while one is viewing a Web page.

As both Oard and Kim (2001) and Kelly and Teeven (2003) acknowledged, the classifications of behaviors are not exhaustive. It requires no stretch of imagination to think of some other searcher behaviors which are not included in the classifications, but may be used to infer searchers' interests on the content. For example, if a searcher finds a Web page in a foreign language that she does not understand, but based on some clues (such as an image, or an abstract written in her native language) decides to pursue a translation of the page, this behavior is a very strong indicator of her interest on the page, given the cost and effort involved. There are also other dimensions which have been suggested in the literature to characterize behavioral evidence of interests. For example, Shen, Tan, and Zhai (2005b) made the distinction between long term (e.g., query history) and short term (e.g., immediately viewed documents) contexts. Kelly and Teeven (2003) suggested the distinction between evidence based on individual's behavior and those on the group level. A few other articles (e.g., Claypool et al., 2000; Jansen & McNeese, 2005), although not specifically focused on developing classifications of observable behaviors, included some discussions on this topic.

3.2.2 Empirical studies on relationships between searchers' behaviors and interests

For implicit feedback techniques to work, three issues need to be addressed: choosing techniques to capture behaviors, establishing the reliability of these behaviors as

evidence of searchers' interests, and designing algorithms to exploit the evidence. Among them, the fundamental question is to identify what observable behaviors mean, especially their relationships with searchers' interests (Kelly & Teeven, 2003). In this subsection, empirical studies that aimed to answer this question are reviewed.

Table 3.3 summarizes implicit feedback studies from two dimensions: the media on which the behaviors take place and the tasks that motivate the behaviors. Implicit feedback studies were conducted before the Web era. Many early studies were conducted on media such as UseNet (e.g., Stevens, 1993; Morita & Shinoda, 1994). It is not clear if findings from these non-Web studies apply to the Web because the Web represents a multimedia environment with free authorship and a variety of information, from serious "stuff" to entertainment information, advertisement, or even spam, which did not exist in the traditional online media. It is reasonable to expect that people's behaviors differ when they seek information on the Web versus on other types of media. Task may have a significant impact on people's behaviors, too. Kelly and Teeven (2003) pointed out that "implicit feedback is often difficult to measure and interpret, and should be understood within the larger context of the user's goals and the system's functionalities" (p. 25). When the goal is to search the Web for information, it is reasonable to expect that people may exhibit different types of behavior or the same type of behavior should be interpreted differently from, for example, when they browse the Web to read the news or do online shopping. Therefore, when studies are reviewed here, attention is paid not only to the types of

behaviors that were examined, but also to the context in which these behaviors were captured. Understanding the context will help us interpret the findings more accurately.

Table 3.3. Implicit feedback studies classified based on media and tasks

	Non-search	Search
Non-Web	Golovchinsky et al. (1999): reading, annotating and judging relevance of documents using pen tablet Konstan et al. (1997): using UseNet reader software for natural tasks Morita & Shinoda (1994): reading articles from newsgroups Salojarvi et al. (2003): judging the relevance of newspaper articles based on titles Stevens (1993): reading UseNet news using study software	
Web, but with modified interface or added agent		Joachims (2002): meta search engine Strive White et al. (2002b): special interface, Alta Vista backend Jung et al. (2007): document search system SERF Lv et al. (2006): search agent PAIR Shen et al. (2005c): Web search interface running on TREC data White et al. (2002a): generic interface connected to Google Zhang & Soe (2001): interface agent of homegrown search system WAIR

Table 3.3 Implicit feedback studies classified based on media and tasks (continued)

	Non-search	Search
Web	Atterer et al. (2006): one search, one setting up an online calendar Claypool et al. (2001): unstructured Web browsing Cooper & Chen (2001): library catalog search Goecks & Shavlik (2000): Web browsing by one of the authors (150 pages on machine learning, 50 pages on other topics) Hijikata (2004): free browsing of subject selected websites Kelly & Belkin (2004): general Web use; task was a study variable Kim et al. (2000): finding sources for a research paper Maglio et al. (2000): attentive system installed on computer desktop Puolamaki et al. (2005): judging relevance of articles based on titles Rafter & Smyth (2001): online job search Zhang & Callan (2005): reading news in a customized browser 1 hour per day for 4 weeks	Agichtein et al. (2006) Fox et al. (2005) Joachims et al. (2005) Rodden & Fu (2007)

For non-search studies, the table lists the tasks that participants were doing while their behaviors were captured. As the table shows, very few studies have been specifically focused on Web search while more studies of implicit feedback on the Web either did not pay attention to task, or were conducted during general browsing activities or during searching in special systems. As the focus of this dissertation is on Web search, implicit

feedback studies on Web search are examined below in detail while only some key papers in other categories are reviewed. Also, some of the relevant papers are reviewed in Section 3.3 in the context of techniques to capture behaviors as implicit feedback.

Agichtein, Brill, Dumais, and Ragno (2006) used both server side and client side logging to capture searchers' natural interactions with a commercial search engine in a 21 day period from three aspects: query features, browsing features and clickthrough features. Query features include query length, fraction of shared words between query and title, summary, URL, and domain, and the overlap between two adjacent queries. Browsing features are used to characterize interactions with pages beyond the results page, such as dwell time and number of clicks to reach the page from the query. Clickthrough features include result ranking, click frequency and whether there is a click on the next or the previous result. They demonstrated that these behaviors could be used to build user behavior models that can more accurately predict users' preferences of search results. Although searchers' behaviors were captured in the natural search environment, the evaluation of search results relevance was done by judges.

Fox et al. (2005) is another example of studies conducted in the natural environment. They recruited 146 Microsoft employees to participate in the study over a 6-week period. Participants used a customized browser (IE add-in) for their normal activities while the mouse and keyboard use was recorded when participants searched MSN or Google. Based on mouse and keyboard activities, result-level implicit measures (time, scrolling, clicking,

result position, number of visits, exit type, page characteristics, bookmarking and printing) and session-level implicit measures (query count, results visit, end action, and average of some result-level measures) were computed. Participants were also asked to provide explicit feedback after each result visit and after each search session. Their goal was to construct predictive Bayesian models that predict the relationships between implicit measures and explicit judgments of satisfaction at both the page and session levels. They found that clickthrough was the most important individual variable but that predictive accuracy could be improved by using additional variables, notably time spent on the result page and how a searcher exited a result or ended a search session. Furthermore, it is worth mentioning that they used several measures to characterize different aspects of some behaviors. For example, when they considered scrolling, they did not only consider if a user scrolled down the page, but also scrolling count, average seconds between scroll, total scroll time, and maximum scroll.

Joachims et al. (2005) presented an empirical evaluation of interpreting clickthrough evidence. By performing eye tracking studies and correlating predictions of their strategies with explicit ratings, they demonstrated that it is possible to accurately interpret clickthrough events in a controlled, laboratory setting. Their evaluation methodology also required the availability of explicit judgments, but unlike most other studies, they collected such data from external judges by asking them to weakly order search results based on how promising they looked. They chose this relative assessment method because “it was

demonstrated that humans can make such relative decisions more reliably than absolute judgments for many tasks" (p. 156).

Rodden and Fu (2007) analyzed mouse movements on Google search results pages. They conducted a study where 32 participants were asked to complete a range of tasks using Google, and tracked both their eye movements and mouse movements. They found that within a single visit to the result page, there was a high degree of overlap between the sets of page regions covered by the mouse and eye. Mouse movements sometimes closely tracked eye movements in terms of distance, region, or sequence, but certainly not all of the time. Mouse and eye were generally closer in the Y direction than in the X direction. They also found that mouse movements showed potential as a way to estimate which results page elements the user had considered before deciding where to click, e.g., by noting which regions were covered by the mouse during the visit, or measuring the vertical distance traversed. They suggested that mouse movements have some potential as a method of determining whether the user has noticed the answer to their question on the results page itself, but are unlikely to tell us much about which aspects of the surrogate the user is taking into account when making a decision. Finally, they pointed out that it is very hard to automatically identify useful behavior patterns from mouse data alone and that it is hard to classify users since each one seemed to exhibit all of the patterns (keeping the mouse still, using the mouse as reading-aid, and using the mouse to mark an interesting result) to varying degrees.

Claypool et al. (2001) designed a customized browser called the Curious Browser to collect the behaviors of 75 students who were instructed to use the browser for 20-30 minutes of unstructured browsing in a lab environment. Subjects were asked to provide explicit ratings of the pages upon exit and those ratings were used to evaluate the implicit measures, including the time spent on page, the time spent moving the mouse, the number of mouse clicks, and the time spent scrolling. The findings suggested that the time spent on a page, the total amount of scrolling on a page (with keyboard or mouse), and the combination of time and scrolling have a strong positive relationship with searchers' interests.

Cooper and Chen (2001) studied the behaviors of searchers of a Web-based library catalog using server-side logs. They considered a search session as relevant to the searcher if any of four types of actions were performed: saving, printing, mailing, or downloading a citation. They then used five categories of variables, session variables, search variables, display variables, error variables, and help variables to predict the binary relevance of a session. Most of these variables were specific to catalog search, such as "the number of different databases used during a session", "the number of different indexes used during a session" and "the number of author searches in a session", while some also applied to other types of search situations, including the general Web search, such as "the length of a session in seconds", "the number of searches performed during a session", and "the total number of items retrieved in a session". A number of "deprived variables" based on sums, averages,

and proportions of the observed "base variables" were also included in the prediction model. As a result, for a population of 905,970 sessions, of which 17.85% of the sessions were relevant, their methodology predicted that about 11% of the sessions were relevant.

Jung et al. (2007) used a proxy server to record searchers' queries and result selection behavior, as well as searchers' binary relevance judgments. Searchers had a chance to provide these judgments at each page that they visited. Explicit ratings were also obtained by using external assessors. Both sources of explicit ratings were compared to three subsets of click data: documents reached directly from the search results, the document last requested by users before initiating a new search or leaving the system, and documents reached by following a link from a page other than the search results page. Results suggest that the last visited document category of click data has the highest percentage of explicit positive ratings, followed by the clicks from the search results list, and then clicks beyond the search results list.

Zhang and Soe (2001) built a Web-based personalized information filtering system called WAIR (Web Agents for Information Retrieval) which could monitor searchers' browsing behaviors. Their experiment considered four sources of implicit feedback: reading time, bookmarking, scrolling and following up links in filtered documents. They found that bookmarking reflects user's interest most strongly, but following up hyperlinks and scrolling were not strong indicators for relevance of documents. They also found that the participants spent more time reading relevant documents than irrelevant ones, but large

reading time (10 or more seconds) was occasionally spent on neutral and irrelevant documents.

Several observations should be highlighted from the review of these studies. First, although a wide range of behaviors have been studied empirically, time and link selection (clickthrough) are by far the most frequently studied implicit measures of users' interests. The observation is also made by Kelly (2005), who suggested that many researchers elected to study these two behaviors because they are "seemingly easy to monitor and gather and are available for every object with which the user interacts" (p. 173). This suggests an important angle to look at the different types of behaviors: their frequencies. Kelly notes that the confidence one has in inferring the user's interests based on a behavior is related to the number of opportunities that one has to observe the behavior. The more frequent a behavior occurs, the weaker it is as an indicator of the user's real interest. On the other hand, however, if a behavior is so rare that it can hardly be observed in a normal use setting, it has limited use in inferring the user's interests either. For example, Goecks and Shavlik (2000) noted that "although bookmarking a page is likely the action most highly correlated with user interest in a page, it is too rare an event for it to be of much use" (p. 130). Kim et al. (2000) noted that only two cases of printing behavior were available from the data they collected, so "no meaningful interpretation on the data collected could be made with only two cases". They further suggested that "the low frequency of the printing behavior might

have resulted from a disparity of goals among the subjects", pointing to the importance of considering the study context, which will be discussed later in this section.

Similar conclusions can be drawn for most of the reference and annotation types of behaviors, which exist in very small quantities under normal Web search settings. Nonetheless, it is worth noting that interface design and human-system interaction style has a potential impact on what types of implicit feedback are available for observation and use (Kelly, 2005). Kelly cited the example of CiteSeer which was an automatic generator of scientific literature databases. CiteSeer displayed document citations to the user, who could then view the full text, rate the document, view citations made to the document and view the bibliography of the document. Kelly noted that this type of interaction style provided more opportunities to collect implicit and explicit feedback than one which only allowed the user to query and examine search results. Such an observation is also supported by recent development in the social networking community. Two successful examples are Amazon (which uses purchase as implicit feedback) and Del.icio.us (which explores social bookmarking).

It is interesting to note that Goecks and Shavlik (2000) used the total amount of page activity to infer the Web browser's interests on a page. Their system labels a page as a positive instance of the user's interests if the user performs a large number of actions on the page. The actions they consider are link selection, scrolling, and mouse activity.

Secondly, another factor that influences the reliability of behavioral sources of implicit feedback is the user's "deliberateness" (Kelly, 2005) in engaging in a behavior. The more resources (cognitive, time, material, or financial) a behavior requires, the more likely that the behavior is deliberate. For example, in a study of bookmarking behavior, Rucker and Polanco (1997) argues that "in contrast to a click, which can be inadvertently done and rarely takes much effort or investment, bookmarks are the result of a very intentional act, something which (especially if the bookmark is placed in a folder) takes some degree of thought and effort, making them a less 'noisy' input for inference" (p.73).

Considering the frequency and deliberateness of behaviors, the types of behaviors towards the bottom (e.g., reference or annotate behaviors) in Table 3.2 are stronger indicators of interests than examination behaviors since the former occur less frequently and more deliberately. In one study (Cooper and Chen, 2001), four types of retention behaviors, saving, printing, mailing, and downloading a citation, were even used as "relevance indicator variables", an equivalent to the "ground truth" feedback that were explicitly provided by the users in most studies. In contrast, less deliberate behaviors, such as examination behaviors have found be affected by the context and highly individually dependent. For example, Fox (2003) found that printing and bookmarking were highly indicative of Web document satisfaction, but dwell time was highly individually dependent. Zhang and Seo (2001) found that bookmarking reflected the user's interest most strongly, while following-up the hyperlinks and scrolling were less strong indicators for relevance of

documents. In terms of reading time, Zhang and Seo (2001) found that although the documents on which users spent longer time to read were more likely to be rated as "relevant", there is some ambiguity on the difference between "long time" and "short time" around 10 seconds.

Thirdly, more research needs to be conducted to understand what observable behaviors mean and how they change with respect to contextual factors (c.f., Kelly, & Teeven, 2003). Despite the general observation that information seeking behavior is affected by task in a variety of ways (Vakkari, 2003), it is argued that research on implicit feedback has paid little or no attention to task (Kelly & Belkin, 2004). It can be noted from Table 3.3 that different studies are conducted in very different settings, observing different types of participants doing different types of tasks, and most studies have only investigated a single task. For instance, although examining the same behavior, viewing, with the same focus on viewing time, Morita and Shinoda (1994), Rafter and Smyth (2001), Kim et al. (2000) and White, Ruthven, and Jose (2002b) observed the behavior when subjects were involved in very different tasks: reading Usenet news, reading online job descriptions, reading academic journal articles, and reading search result summaries in a document summarization system. It remains to be determined whether their findings reflect behavior of Web searchers in general (Hsieh-Yee, 2001), or other contexts involving reading.

There are many examples which can be cited to demonstrate that contextual factors have significant impact on how behaviors should be interpreted. For instance, time has been

demonstrated to be a reliable indicator of users' interests in some online reading environment, such as UseNet (Morita & Shinoda, 1994; Konstan et al., 1997). Konstan et al. (1997) even found that "predictions based on time spent reading are nearly as accurate as predictions based on explicit numerical ratings" and that "the relationship between time and rating holds true without regard for the length of the article" (p.84). However, in a TREC interactive search study, Kelly and Belkin (2001) found that the length of time that a searcher spent viewing a document was not significantly related to the user's subsequent relevance judgment. In the Web information seeking context, Kelly and Belkin (2004) found that reading time varied between subjects and tasks, which made it difficult to interpret.

Another widely examined behavior, link selection, has similar problems. Although clicking search results is generally regarded as a positive indicator of the searcher's interests (Joachims, 2002), it is possible for searchers involved in fact-retrieval type of tasks (e.g., the definition of a word, or a stock quote) to find the answer simply by reading the snippet or from a special section (usually the top) of the page. In those cases, no click would occur even though the search results are very relevant to the searcher.

These observations suggest that while it is important to understand what measures can be accurate predictors of relevance, it is also important to understand what mediating factors, perhaps not immediately visible from information-seeking behavior, can influence the effectiveness of implicit feedback. Such factors may include individual characteristics

(e.g., search experience of the user and the stage in the search), task complexity, topic, document collection and search environment (Kelly & Belkin, 2001; Kelly & Teeven, 2003; White, Ruthven, & Jose, 2005; White & Kelly, 2006).

Fourthly, in terms of evaluation methodology, almost all the studies found evaluate the reliability of observable searcher behaviors as implicit interest indicators by somehow comparing them against explicit ratings. This approach is based on the assumption that explicit ratings give more accurate information on what a searcher finds interesting and useful. If a behavioral measure is found to correlate well with explicit ratings, it can potentially be used in lieu of or in conjunction with the explicit feedback. Two notable exceptions were Cooper and Chen (2001) and Rafter and Smyth (2001). As mentioned above, when studying searches of online library catalog, Cooper and Chen (2001) used four types of retention behaviors, saving, printing, mailing, and downloading a citation, as "relevance indicator variables" and use them to evaluate other implicit indicators of interests. When studying users' behaviors on a job search website, Rafter and Smyth (2001) assumed that the action of a user applying for a particular job online is a reliable indicator of her interest in that job; therefore, they evaluated the two implicit behavioral measures (time spent reading a job description and the number of times a user revisits the description) based on how well they correlated with job application. Although retention of a citation and job application in these two cases seem to be good indicators of users' interests, it is rare that such behavioral indicators are available in the general Web search context. Therefore, a

more viable alternative is to turn to the searchers and ask them to provide explicit ratings on search results. Then, implicit measures of interests can be compared to the explicit ratings. If a behavioral measure is found to correlate well with the explicit ratings, it can potentially be used in lieu of or in conjunction with the explicit feedback.

A second general approach to establishing the reliability of implicit feedback is through demonstrating the contribution of implicit feedback measures to the improvement of retrieval performance. This approach is often used in studies on personalization systems (e.g., Teeven, Dumais & Horvitz, 2005; Shen et al., 2005a). In these studies, certain behaviors were used as implicit feedback to build user profiles or customize search results. Then, performances of the systems that used implicit feedback were compared to baseline systems to test the utility of implicit feedback. A potential problem with this utility oriented approach is that the reliability of implicit measures is mingled with the effectiveness of the algorithm that builds on top of them so that it is not clear whether a lack of improvement in retrieval performance should be attributed to the lack of association between the type of behavior and searchers' interests or to the ineffectiveness of the algorithm that implements the implicit feedback.

Fifthly, although the usefulness of the behaviors is a major factor that determines what are studied, in practice, another factor has also been discussed extensively in the literature: limitations of observation techniques. Technical feasibility has a direct impact on which behaviors can possibly be made available for study and how they can be interpreted.

For example, Goecks and Shavlik (2000) noted that "technology limitations currently prevent the agent from obtaining an accurate measure of the amount of scrolling by a user" (p. 130). Kelly (2005) pointed out that although obtaining implicit feedback about a segment will presumably provide more precise information about the user's interests, there is less research across the minimum scope categories of segment and class because for most systems, the unit with which the user most often interacts is the object, which makes it more expensive to observe behaviors at other scopes. A good example is the cost of capturing which segments of a page one looks at using eye-tracking techniques versus only capturing which pages one looks at using automatic logging techniques. Kelly (2005) also suggested that "there is less research investigating the behaviors of retain, reference, annotate and create since it is often necessary to collect this information from the client, rather than the server, which usually requires specialized software and permission from users" (p. 173).

In addition, a common limitation of observation techniques is that they do not capture the intention of the behaviors. This makes it problematic to use to infer the searcher's interests. For example, Rucker and Polanco (1997) found that users tended to bookmark for wildly different reasons, ranging from genuine interest to a transient need to return to a page. Some behavior may even be unintentional, which introduces "noise" to the analysis. For example, Kelly and Teeven (2003) noted that the amount of time that an object is displayed does not necessarily correspond to the amount of time that the object is

examined, yet display time is traditionally treated as an equivalent to reading time, introducing potential inaccuracy.

3.3 Survey of methods to capture searcher behavior

The aim of this section is to survey the methods that have been used to capture searcher behavior and can potentially be applied in this study. By discussing the advantages and limitations of each method, evidence-based choices for data collection methods for the first phase of the study were made.

Different methods are used to collect data on different aspects of user behavior at different levels, which range from micro (mechanical) level *actions* such as eye movements, mouse movements, mouse clicks, scrolling, and key strokes, to macro (algorithmic) level *activities*, such as selecting menu items, following links, filling forms, and pressing buttons, all associated with *mental activities*. A macro level activity can consist of one or more micro level actions (e.g., filling a form entry involves multiple key strokes and clicking on the "submit" button) and/or over a certain object (e.g., mouse click on a link or a button). At the global level, the totality of activities related to a certain task or during a certain period of time (e.g., during a laboratory study) forms a *session*.

From the data collection point of view, data is collected either through direct observation (the researcher watches the user's actions and takes notes), or some other forms of recording such as log and video (Preece et al., 1994). Direct observation can be useful to

gain a general understanding of the use of the system, but it is obtrusive (users may be constantly aware of their performance being monitored, which can alter their behavior) and too crude to capture users' interactions with the system in detail, so this method will not be further pursued here. Among indirect observation techniques, eye-tracking techniques can be used to capture eye movements, client-side logging software can be used to capture mouse movements, transaction logs can capture other mechanical levels actions (mouse clicks, scrolling and key strokes) and algorithmic level activities (following links, filling forms and pressing buttons), and video taping methods can capture the context of the user activity and users' behaviors in continuity. Finally, users' mental activities during the search process can not be directly observed by the researcher; therefore, verbal protocols are used to elicit descriptions of what users are thinking about while they carry out search tasks.

3.3.1 Logging

Logging is an intentionally fuzzy name given to encompass a set of techniques which automatically record a user's actions. It involves having the computer automatically collect statistics about the detailed use of the system. An important distinction of different logging techniques is where the transaction log is generated and stored. There are two general approaches. The first is the server-side approach in which a Web server records and stores the interactions between a user (actually a browser on a particular computer) and the server in a log file on the server. This approach is mainly used to capture link clicks and the information that users submit via HTML forms, such as query terms and relevance

judgments. The information stored typically includes the client computer's IP address, access time, among other fields (Spink & Jansen, 2005). If studies are conducted on third party search engines (in which researchers do not have access to the server-side logs) or if certain interventions are needed before search results are displayed, researchers can set up proxy servers to transparently intervene and capture the interactions. Hyperlinks on the pages that are presented to the users do not lead directly to the suggested page, but point to a proxy server. When the user clicks a link, the request is recorded by the proxy server, before the server fetches the page, (optionally) performs the intervention, and displays it to the user.

The log file that is generated in this server-side approach is usually referred to as the "transaction log". Jansen (2006) defined a transaction log for Web searching as "an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine" (p. 408). Likewise, "transaction log analysis" on a search system has been defined as "the use of data collected in a transaction log to investigate particular research questions concerning interactions among Web users, the Web search engine, or the Web content during searching episodes " (Jansen, 2006, p. 409). Peters (1993) provided a historical review of this technique, while Sandore (1993) reviews methods of applying the results of transaction log analysis.

In the Web search environment, transaction log analysis is conducted at multiple levels (Spink & Jansen, 2005; Jansen, 2006). One is to focus on the queries submitted to a particular search engine or several search engines of interest. By mining a large set of queries (e.g., all queries over a six-month period), researchers can study query patterns, trends, time fluctuations, topical features and so forth. Examples of such work include Beitzel, Jensen, Chowdhury, Grossman, and Frieder (2004), Wang, Berry and Yang (2003), and Jansen, Spink, and Saracevic (2000), as well as Google's Zeitgeist[1].

Another approach to using transaction log data in search engines is to examine "clickthrough data" or "click streams" which indicate which results are clicked in response to which query as well as the ranking of results. Joachims (2002) described clickthrough data more formally as triplets (q, r, c) consisting of the query q, the ranking r presented to the user, and the set c of links the user clicked on. The assumption for analyzing clickthrough data is that the results that are clicked are more relevant to the query than those that are ignored. Therefore, search engine developers use click streams to determine the quality of results and tune their algorithms: the larger proportion of highly ranked results are clicked, the better is the search and ranking algorithm. For example, Joachims (2002) used clickthrough data as training data to learn a retrieval function, which is then used to adapt the algorithm of a meta-search engine (Striver) to a particular group of users.

---

[1] http://www.google.com/press/zeitgeist.html

Jansen and Pooch (2001) pointed out that transaction log analysis is "the most reasonable and non-intrusive means of collecting user-searching information from a large number of users" (p. 236). Despite so, there is limited information that server-side logging can capture. Essentially, it can only capture information that the user submits through the browser or the links that the user clicks, but not the actions between submissions or clicks. A lot of valuable information is therefore missed, which includes the order the user fills in different fields on the form, the links that the user considers (looks at or hovers over) but does not click on, the places on the screen that are designed to be not clickable but the user attempts to click, the use of the "back" button and other behaviors such as printing and bookmarking. Hargittai (2002) pointed out that data sets collected with server-side logging can not capture the use of the "back" button on browsers, which comprises up to 30% of people's browsing activities (Tauscher & Greenberg, 1997) and may be considered a part of one's level of search sophistication. Choo, Detlor, and Turnbull (1999) pointed out that server-side logs or proxy server logs do not capture Web access from the browser's local cache, which typically provides most of the Web pages requested via the Back and Forward buttons in Web browsers; neither do they log actions such as bookmarking, printing a Web page, or finding terms in an open page.

To address these limitations, client-side logging is increasingly used in research and commercial search systems to collect finer-grain information on users' behavior. In the academic research environment where researchers have more control over users' work

environment, client-side logging is often performed through customized browser or dedicated logging software installed on study computers; whereas in a more practical or commercial environment, a toolbar is often used to implement the logging.

In a study to understand how an online information system could automatically predict which Web documents users prefer by monitoring their online behaviors with documents, Kelly (2004) provided each of 7 subjects with a laptop and asked them to use the laptop over a 14-week period. Subjects were informed that all of the activities that they performed while using the laptop, including online searching, email and word processing, would be logged. Logging was done in two ways. One was to use the commercial client-side logging software "WinWhatWhere Investigator" that was installed on the laptops. The other was to direct all online activities performed on the laptops through a custom built proxy logger. The logging software unobtrusively monitored and recorded subjects' interactions with all applications including the operating system, web browsers and word processors. For search activities, information such as the browser used, URLs and page titles (for all visited Web pages), start, finish and elapsed times, queries, and raw keystrokes made at a document, were recorded. Also recorded by the logging software were three types of retention behaviors of interest to the researcher: printing, saving, and bookmarking. The function of the proxy logger was simply to save a local copy of each page request made by subjects.

Claypool et al. (2001) created a customized Web browser (Curious Browser) to record the online behavior of 75 students, who were instructed to use the browser for 20 to 30 minutes of unstructured browsing, and obtain their explicit relevance ratings of web pages. The behaviors that were examined include mouse movement, mouse clicks, scrolling, and time on page.

Choo, Detlor, and Turnbull (1999) studied how managers and IT specialists use the Web in a natural setting. Researchers installed a piece of client-side logging software (WebTracker) to record participants' Web-use activities during two-week periods. The application ran transparently whenever the participant's Web browser was being used. It recorded all URL calls and requests, as well as most browser menu selections, including "Open URL or File", "Reload", "Back", "Forward", "Add to Bookmarks", "Go to Bookmark, "Print", and "Stop". The log was stored on the participants' hard disks and collected at the end of the study.

As reviewed by Jansen (2006), there are also other commercial or academic software available to generate client-side logging, with varying functionality. Examples include Morae 1.1[1] by TechSmith and Wrapper[2].

---

[1] http://www.techsmith.com/products/morae/default.asp

[2] http://ist.psu.edu/faculty_pages/jjansen/academic/wrapper.htm

Besides being used by commercial search engines to compute advertising charges, logging has been used to obtain statistics and patterns of system use (Hert and Marchionini, 1998; Jansen, Spink, and Saracevic, 2000) and identify usability problems (Nielsen, 1993). With regard to this dissertation, the most relevant use of logging is to infer users' interests by monitoring their behavior. The work by Claypool et al. (2001), Kelly (2004), and Shen and Zhai (2005) fall into this category. It is worth noting that all these studies used client-side logging, as the focus was on the behavior and interests of individual users, rather than the collective behavior of the user population. In addition, Shen and Zhai (2005) argued that client-side logging has two remarkable advantages over the server-side approach, for the purpose of what they call "personalization": the protection of privacy and the reduction of the server load.

The literature suggests that using logging to study users' interactions with search systems has the following advantages (Peters, 1993; Nielsen, 1993; Jansen & Pooch, 2001; Spink & Jansen, 2005; Shen & Zhai, 2005; Jansen, 2006). Firstly, logging the users' actual use of the system is particularly useful because it shows how users perform their actual work. Since the data is collected unobtrusively while real users are using real systems on the Web to search for information that they really want to pursue, the log data should most closely represent the unaltered behavior of users. Secondly, logging (especially server-side logging) provides a method of automatically collecting data from a large number of users working under different circumstances. Thirdly, the data can be collected fairly

inexpensively. It does not require the researcher to be present. The costs are basically the software and storage.

The limitations of using logging techniques have also been discussed in literature. Firstly, logging a user's use of a system raises privacy concerns (Volokh, 2000; Jansen, 2006; Wang, Hawk, & Tenopir, 2000). This should first be addressed by informing users when interaction logging is performed and allowing users to disable the log if they so desire. Additionally, efforts should be made so that only summary statistics are being collected and results will only be reported in a form where individual users can not be identified. However, there have also been arguments that in cases where data are derived from larger segments of the online population, no information is available about specific users, and thus it is impossible to make any claims about how attributes of users may be related to their online behavior (Hargittai, 2002).

A second limitation of logging techniques is that they do not record the reasons for the search, the searcher motivations, or other qualitative aspects of use. Neither do they provide reasons why certain behavior happens. Therefore, it is advocated that logging techniques should be used in conjunction with other methods that capture users' information needs, comments and reactions while using a system, and their satisfaction with the system (Peters, 1993; Griffiths, Hartley, & Willson, 2002; Spink & Jansen, 2005). Preece et al. (1994) note that researchers often combine video, audio and keypress or interaction logging so that they can relate revealing data about body language (posture, smiles, scowls and so

on) and comments or more detailed audio protocols with records of the actual human-computer interaction. Although this may sound ideal, Preece et al. (1994) further note that this approach has two drawbacks. It can be expensive to buy or build synchronized equipment. The volume of data collected can also be daunting to analyze.

Thirdly, information contained in the logs can be inaccurate or hard to interpret for various reasons. For example, an IP address or cookie is typically used to identify users from a transaction log; however, as more than one person may use a computer, the IP address is an imprecise representation of the user. Session identification can also be troublesome. For example, Catledge and Pitkow (1995) had to delineate session boundaries artificially from captured log files (all events that occurred over 25.5 minutes apart were delineated as a new session) because they relied on client-side caching of search activities and some users had left their machines running for long periods without any interaction.

Another source of inaccuracy in using the data generated by logging is due to the dynamic nature of the Web. It is commonly known that search engine results are constantly changing. The same search engine may very well return different results for the same query as time evolves. Logging, however, can only capture the status of the user-system interaction at particular time points. Changes to the system may make it hard for researchers to replicate searches or compare across systems (Griffiths et al., 2002).

Fourthly, the logged data may not be complete. As mentioned before, server-side logging can not capture interactions between clicks or submissions. Another source of

inaccuracy is caching. When a user accesses the page of results from a search engine using the "back" button of a browser, this navigation accesses the results page via the cache on the client machine, so the server will not record this action. Special procedures have to be taken into account for such incompleteness.

Finally, some authors (e.g., Kurth, 1993; Wang et al., 2000) noted that the volume of data generated by logs can cause difficulties for analysis.

3.3.2 Eye-tracking

Eye-tracking was first used in the 1800s to study eye movements during a reading process. Through direct observations at first and eye tracking equipment later, people realized that reading does not involve a smooth sweeping of the eyes along the text, as previously assumed; instead the eyes make short stops, called fixations, and intermediate quick saccades[1]. Since non-intrusive eye trackers were invented, eye-tracking has been used increasingly as a tool to study the cognitive processes of humans performing a wide variety of tasks involving a user interface. The technique, most extensively used in experimental psychology, is based on Just and Carpenter's (1980) strong eye-mind hypothesis which states that there is no appreciable lag between what is fixated and what is processed. That is, we can infer what users think about by monitoring which word or object they look at, and for exactly as long as the recorded fixation. However, it is easy to notice that one can attend

[1] http://en.wikipedia.org/wiki/Eye_tracking

to something different than what one is looking at. This phenomenon, called covert attention (Posner, 1980), presents a challenge to the eye-mind hypothesis. When covert attention happens during an eye-tracking study, the resulting scan path and fixation patterns would often show not where the subject's attention has been, but only where the eye has been looking. With regard to this discrepancy, the current consensus is that (visual) attention is about 100 and 250 milliseconds ahead of the eye, but as soon as (visual) attention moves to a new position, the eyes will want to follow (Hoffman, 1998; Deubel & Schneider, 1996).

Rayner (1998) and Duchowski (2002) reviewed the development of the eye-tracking technique and its applications in different areas. The most pertinent application to this dissertation is to use eye-tracking in human computer interaction studies, especially to study searchers' interactions with Web search engines. Researchers typically analyze eye movements in terms of *fixations* (a spatially stable gaze lasting for approximately 200-300 milliseconds, during which visual attention is directed to a specific area of the visual display), *saccades* (rapid movements between fixations), *pupil dilation* (typically used as a measure to gauge an individual's interest or arousal in the content they are viewing), and *scan paths* (visualization of eye fixations on a page in order) (Rayner, 1998). Common analysis metrics include fixation or gaze durations, saccadic velocities, saccadic amplitudes, and various transition-based parameters between fixations and regions of interest. When eye-tracking is done over a group of people, aggregate images of their fixations can be

generated. The images, often called “heatmaps”, give a vivid representation of which region draws the attention of the group.

Eye-tracking can be used in human computer interaction studies of different types, including Web usability inspection (e.g., Duchowski, 2002; Goldberg & Kotval, 1999; Schroeder, 1998) and comparison of design options for a prototype system, or comparison of a prototype web site with a competitor site (e.g., Goldberg, Stimson, Lewnstein, Scott, and Wichansky, 2002; Rele & Duchowski, 2005). More recently, eye movements have also been used as a source of implicit feedback for information retrieval (c.f., Puolämaki et al., 2005). It is argued that gaze is by far one of the most important nonverbal signs of human attention, so searchers’ eye movement data can be a more reliable source of implicit feedback than self reported subjective data generated by methods such as focus group and verbal protocol (Schiessl, Duda, Thoelke, & Fischer, 2003; Salojärvi et al., 2003). Eye-tracking also has the distinctive advantage of providing insights into searchers’ behavior between clicks and allowing inferences of their interests on search results, especially on those results that they do not click on. Such information can either be used for post-trial, off-line improvement of search algorithms based on the aggregate browsing patterns of a group of searchers, or be used real-time to allow systems to respond to or interact with a particular searcher based on the observed eye movements (Duchowski, 2002). Additionally, eye tracking not only allows researchers to gather qualitative data, but also produces gaze plots and other quantitative data about eye movements.

Maglio and colleagues (Maglio, Barrett, Campbell, & Selker, 2000; Maglio & Campbell, 2003) designed a prototype attentive agent application (Simple User Interest Tracker, Suitor) that monitored eye movements while the searcher viewed web pages in order to determine whether the searcher was reading or just browsing. If reading is detected, the document is defined relevant, and more information on the topic is sought and displayed. However, they did not verify the feasibility of the application empirically.

Salojärvi et al. (2003) studied the relationship between eye movements and relevance judgments. They note that pupil dilation increases while viewing relevant abstracts. That is, a larger diameter typically signifies higher interest in the content matter. This suggests that pupil dilation can be an important indicator of users' interests on a search result. They also find that relevance of document titles can be more reliably predicted by eye fixations than specific words. However, this study only used three subjects, so it is not clear if the result can be generalized to a larger user population.

Puolamäki et al. (2005) designed a controlled experiment to study the potential of combining eye movements and collaborative filtering to predict the relevance of scientific articles. Only three subjects participated in the eye-tracking part of the experiment, with their gaze directions measured at a sampling rate of 50 Hz while they performed an artificial task of scanning 80 pages, each containing 6 titles of scientific articles, and choosing the two most interesting titles. Although the results suggested that the prediction accuracy with eye movements or with eye movements combined with collaborative filtering was

significantly better than by chance, the findings are subject to the same limitation in generalizability as those of Salojävi et al. (2003). In general, it is safe to conclude that the reliability of inferring relevance implicitly from eye movements is inconclusive so far.

In addition to using eye movements as implicit feedback, a few other studies used eye-tracking data to examine the reliability of other sources of implicit feedback. Granka, Joachims and Gay (2004) used eye-tracking to better understand how searchers browsed the presented search result abstracts and how they selected links for further exploration. They pointed out that better understanding of searcher behavior is valuable for improved interface design, as well as for more accurate interpretations of implicit feedback (e.g., clickthrough) for machine learning. Joachims et al. (2005) used eye-tracking to study the searchers' decision making process before they clicked on search results and evaluated the reliability of clicks as indicators of relevance. They found that clicks were informative but biased. They suggested that it is more appropriate to interpret clicks as relative preferences, rather than absolute relevance judgments.

Despite the desirable features, the eye-tracking technique also has its drawbacks. First of all, eye-tacking data provides excellent low-level traces of human behavior but does not stand alone in explaining how or why people use interfaces (Jacob, 1991; Sibert & Jacob, 2000). Specific cognitive processes can not be inferred directly from a fixation on a particular object in a scene. For instance, a fixation on a face in a picture may be indicative of recognition, liking, dislike, or puzzlement. Eye-tracking is therefore often coupled with

other methods, such as verbal protocols (concurrent or retrospective). Penzo (2005) argued that the combination of eye-tracking and think-aloud methods provide a broad overview of the problems a user encounters in a user interface while performing a task because the think-aloud protocol collects qualitative data such as a user's mood through tone of voice and facial expressions, while eye-tracking gathers and records quantitative data such as pupil diameter, fixation coordinates, and fixation length.

Secondly, in most cases, eye-tracking studies require special devices (eye trackers) so that it can only be carried out in the usability lab. On the one hand, this is not the searcher's typical search environment, which may have an impact on their behavior. On the other hand, the existence of eye-tracking devices (the eye tracker, and sometimes screen and voice recorders), the calibration procedure, as well as the requirement that subjects remain seated still (for eye-trackers to capture the data) during the study session may make subjects more or less feel conscious about the study and not behave as they would when they search at home or work. Associated with this is the high cost of user studies involving eye-tracking, including the monetary cost incurred by the need to bring subjects to the usability lab and the time cost due to the infeasibility of conducting eye-tracking studies in a "batch" mode. In addition, the cost and availability of the eye-tracker itself also limits the application of the technology (Li, Babcock, Parkhurst, 2006).

Finally, people have reported difficulties in analyzing eye-tracking data. Granka and Rodden (2006) pointed out that, in general, existing eye-tracking software lacks specialized

features for analysis of studies where web pages are used as the stimuli, e.g., dealing with repeat visits to the same page, or page content that changes dynamically. Schiessl et al. (2003) also noted the immense analysis time of data generated by eye-tracking studies.

3.3.3 Mouse tracking

Mouse tracking can be regarded as a special case of client-side logging. As mentioned earlier, a mouse click is a proven indicator of a user's interest in a web search result. While extremely valuable, clicks do not tell the whole story of the user's interaction with the search results page. For example, since a user's selection of a particular search result is based on the surrogate shown on the results page, it would be useful to have a better idea of which aspects of the surrogate users are paying attention to when making each decision about where to click. Such detailed information can not be captured by the transaction log. Also, in some cases it may be possible for the user to find the answer to a fact-finding question simply by reading the snippet, and many search engines now choose to present relevant information on the page directly, e.g., the definition of a word, or a stock quote. In both of these cases, no click would occur even though the user may have satisfied their information need. In such situations, techniques which can record more subtle signals are needed.

Eye tracking can provide insights into users' behavior at a fine level, but as described above, eye tracking equipment is expensive and can only be used for studies where the user is physically present in front of the eye tracker. In contrast, the coordinates

of mouse movements on a web page can be collected accurately and easily, in a way that is transparent to the user. This means that it can be used in studies involving a number of participants working simultaneously, or remotely, greatly increasing the volume and variety of data available. Therefore, it is natural to consider mouse tracking as a potential alternative to the more expensive eye-tracking technique.

Given this notion, a central question in mouse tracking is how closely mouse movements reflect eye movements. If mouse movements follow eye movements closely, then mouse tracking techniques can be used in lieu of eye tracking for all purposes that eye tracking is used for, such as usability inspection, prototype comparison, and capturing behavior as evidence of implicit relevance feedback. Kantor et al. (2000) discovered that users tended to follow the mouse pointer by the eye while browsing Web pages and suggested that they exhibited such behavior because they had to click links that they were interested in with the mouse. In a small study with 5 participants, Chen, Anderson and Sohn (2001) looked at the relationship between eye movements and mouse movements on a set of general web pages. They divided each web page up into logical regions, and found that there was a high correlation between the total times that the eye and mouse stayed in each region, per page. Mean distance between eye gaze and mouse pointer was 290 pixels, and this dropped to about 90 pixels in situations where the user was moving the mouse within or to a "meaningful" region of the page (i.e., one that had actual page content in it). If the user moved their mouse over a region, there was an 84% chance that they also looked at it.

When the user made a sudden mouse movement towards or within a particular region, the user was looking at that region in more than 70% of cases.

Cognitive modeling researchers have studied eye-mouse coordination during tasks that involve locating and selecting a given target item from graphical user interface menus of various lengths. As well as the target item, the menus contain distracters whose degree of closeness to the target (and relevance to the task) can be manipulated in experiments – more relevant distracters tend to cause users to hesitate more and recheck items before making a selection. Studies (e.g., Byrne, Anderson, Douglass, & Matessa, 1999; Cox & Silva, 2006) have found that users exhibit a number of different mouse movement behaviors:

- the user's mouse remains still, either in the initial location or in a neutral area off to the side, until the target item has been located with the eyes;
- the mouse movements track the eye movements, usually lagging slightly;
- the mouse is used as a marker to keep track of the most promising item found so far, while the user continues to consider further items with their eyes.

Cox and Silva (2006) manually classified the trials from their study into these three types, finding that the patterns occurred with roughly the same frequency, and sometimes in the same trial. In trials where the distracter items were more closely related to the task, users were more likely to adopt the mouse-as-marker strategy. Interestingly, they also found that if participants were instructed not to move the mouse at all until they had located the target with their eyes, their search performance declined (e.g., they made a selection more quickly

but less accurately). This result suggests that it may be actively helpful for users to have the mouse pointer available while making a decision on where to click.

The first type of behaviors observed by Cox and Silver (2006) has also been observed in other studies. Mueller and Lockerd (2001) noted that many users in their study would "rest" the mouse in white space while reading, so that they did not cover up text or accidentally click a link. Arroyo, Selker and Wei (2006) described preliminary results from a study of the mouse movements of 105 users on a single web site. They observed similar types of behaviors as in the other studies, speculating that users who do not use the mouse as a reading aid may be characterized by leaving the mouse in the same position for long periods, followed by quick movements to click targets.

In addition to studies on the general mouse movement patterns and its relationship with eye movements, there were a few studies which examined the possibility of using mouse movement to infer users' interests on and preferences for information objects on web pages. Mueller and Lockerd (2001) described a study where they recorded and analyzed participants' mouse movements on general web pages. They found that 30% of the searchers tended to use the mouse pointer as a marker when looking through a list. On two shopping pages, where the first choice of item was indicated with a click, it was possible to predict the user's second choice 65% and 75% of the time, by looking at how long they hovered over the other links. Claypool et al. (2001) considered mouse actions on general

web pages, finding that the total time spent moving the mouse or using it to scroll the page is correlated with explicit user satisfaction.

Hijikata (2004) observed users' mouse use patterns while they browsed Web pages of their choice and identified 10 types of mouse operations: text tracing, link pointing, link clicking, text selection, scrolling, bookmark registration, saving, printing, window movement, and window resizing. They then focused on four types of mouse operations whose targets were text: text tracing, link pointing, link clicking and text selection, and extracted keywords based on mouse operations as representations of their interests. They found that the mouse-based method extracted keywords that the user was interested in about 3 times more accurately than random extraction of keywords and about 40% more accurately than the tf-idf method.

Rodden and Fu (2007) presented the only study that was found to be specifically focused on studying mouse movements during Web search. The details of the study were reviewed in the previous section. The general conclusion was that mouse movements have some potential as a method of determining whether the user has noticed the answer to their question on the results page itself, but are unlikely to tell us much about which aspects of the surrogate the user is taking into account when making a decision.

In sum, the research reviewed in this subsection acknowledges mouse tracking as an economic way to collect data on users' behavior at a fine granularity. However, the mouse

tracking technique is still in its infant stage and the current evidence is not strong enough to be used to reliably infer users' interests.

3.3.4 Video taping

Video taping is one of the well accepted user observation methods. It covers up for the inability of the physical presence of the entire research team at the real time user environment. It also remedies some shortcoming that direct observations have. For example, researchers do not have to sit next to the user in order to take notes. Having a camera instead of a person is less obtrusive. So, video taping can be used in some situations when direct observations are not possible. For example, Marshall and Bly (2005) had 3 participants video taping themselves reading a weekly magazine when and where they normally would. They subsequently viewed the videotapes to log different kinds of activities of interest.

Video taping can also capture peripheral activities, which may be of interest to the researchers. In the Marshall and Bly's (2005) study, they captured peripheral activities like reaching for a drink, shifting position, and face or head-scratching, as well as the way the participant held the magazine (e.g., one-handed or two). Having this information is beneficial for designers of electronic books. In many studies several aspects of user activity are monitored by different video cameras. For example, one camera may be focused on the keyboard and screen while another is directed at the user. Users' body language can provide useful clues about the way they are feeling about using the system.

In human computer interaction studies, video taping has often been used to capture the screen activities (i.e., screen recording). Recall that a limitation of logging is that logging can only capture the status of the user-system interaction at particular time points, so when it is used to capture users' interaction with Web contents, the Web dynamics makes it hard for researchers to replicate searches recorded in the log. For example, some URLs may no longer be accessible, and some queries may return a different set of results. Unlike logging, video taping captures exactly what happens during the study and that can be reviewed as many times as necessary.

Video taping involves relatively low cost. Special equipment is needed but it is relatively cheap to use once the equipment has been purchased. Video taping is also less obtrusive than most other methods. However, video taping is a very crude way to capture users' interactions. It only reflects high level features of the interaction, such as where the user hesitates. To get the details of the interaction, such as which query terms the user enters, and which results the user examines, researchers have to play the recording back and forth. Some of the features, such as where the user looks, can not be captured by this method. Moreover, the data generated by this method is generally qualitative in nature. It can be both difficult and time-consuming to analyze. For example, it is not often used to obtain the time that a user spends, although it is possible to use a stop watch to get the time information. Generally, it takes about three to five times the duration of a video to complete logging user interactions by pausing and playing the video (Oh & Lee, 2005). In practice,

video taping is often used as a supplementary method to some other logging methods with finer focus to maintain the big picture while other methods capture the details. For example, Preece et al. (1994) note that keystroke logging and interaction logging are often synchronized with video recording.

3.3.5 Verbal protocol analysis

Verbal protocol analysis is somewhat different from the observation techniques reviewed so far in this section in that it is not used to observe the user's behavior; instead, it seeks to reveal human information processing and the thoughts that underlie behavior (Wang et al., 2000). Since the thinking process is not directly observable, researchers have to rely on users to verbalize their cognitive activities.

Verbal protocol analysis is not a method by itself. Rather, it is often used to complement other techniques and gain unique insight on users' thinking processes and reasonings. A common shortcoming of observation techniques reviewed in this section is that data collected by these techniques alone only tell researchers what users did, but not why they did it. So, this type of data by itself may be open to interpretation (Ericsson & Simon, 1993). Thus, it is important to collect verbal data on the thoughts and feelings besides the physical movements. The value of verbal data lies in the fact that it can help to interpret nonverbal actions and activities more accurately.

The justification for using verbal protocols comes from human information processing theory (Griffiths et al., 2002). Ericsson and Simon (1980) maintained that

"verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are a valuable and thoroughly reliable source of information about cognitive processes" (p. 247).

There are two types of verbal protocols that are most commonly used: think aloud protocols and post-event protocols. In a think aloud protocol (also known as concurrent protocol), the user says out loud what she is thinking while she is carrying out a task or doing some problem solving. This enables observers to see first-hand the process of task completion, rather than only its final product. Observers at such a test are asked to objectively take notes of everything that the user says, without attempting to interpret her actions and words. Test sessions are often audio and/or video taped so that developers can go back and refer to what users did, and how they reacted.

As can be easily noticed, think aloud protocols place added strain on users, who are required to do two things at once: to perform the task itself and to describe what they are thinking about. Hence, there have been concerns that thinking aloud alters the cognitive process being studied. There have also been debates on the validity of think aloud protocols. I will briefly discuss some of these concerns here and leave the readers with a pointer for further reading.

Wang et al. (2000) suggest that searchers can only verbalize a subset of the thoughts occurring during the interaction because some thoughts are difficult to verbalize. However, they also suggest that when verbal report is combined with logging data (such as screen

captures), partial verbalization can reveal users' difficulties and problems at specific points during the search. The two methods together form a more complete picture of the interaction and provide insight into users' behavior and thoughts. Some researchers believe that the process of thinking out loud may introduce bias into the primary task and affect measurements. For example, Granka and Rodden (2006) maintain that think aloud should not be combined with quantitative measures (such as time to complete task) because of the bias. They suggest that think aloud protocols also affect eye tracking so that eye tracking data should be used purely qualitatively. For example, a user might pause in the middle of a task in order to explain why they were having a particular problem, and look around the screen far more than they would if simply getting on with their task in silence. For a more comprehensive discussion on reconciling theory and practice of think aloud protocols, please refer to Boren and Ramey (2000).

Another approach to using verbal protocols is to obtain them after the tasks have been completed. These are known as post-event protocols (or retrospective protocols). Post-event protocols are often used when it is important not to interrupt users. To implement post-event protocols, video equipment is often required to record the study session so that users can make comments while the video recording is played back. Users are given the opportunity to explain what they did and why. An example of the use of retrospective protocol is described in Choo et al. (1999). They used a piece of client-side logging software to record participants' Web-use activities during two-week periods. After

that, they conducted retrospective interviews, in which participants recalled critical incidents of using information from the Web. Since the study was over a two-week period, it was not feasible to videotape or review the sessions. Instead, participants relied on their memories and, where appropriate, were prompted by the researchers with the names of Web sites that were indicated in the log files.

With the benefit of not contaminating users' behaviors during study sessions, the use of post-event protocols nonetheless receives the criticism that they can contain recalled information that was not used during the task sequence and that hindsight can produce a rationalization of the user's own actions (Preece et al., 1994). So, strictly, a post-event protocol does not generate observation data; rather, it is good at collecting further explanations or rationales for what is observed. Despite so, some researchers (Monk, Wright, Haver, & Davenport, 1993) report that when users are invited to participate in data analysis, it is often very beneficial because they are stimulated to recall useful details about their problems.

In sum, verbal protocol is a much debated technique which is able to provide some unique insight into users' thinking process while performing a task. It is relatively easy to administer but can be time-consuming to analyze. The data is qualitative in nature and often only makes sense when used together with data collected by other methods, such as logging. When the goal of the study is to understand users' problem solving process or to find out where they have difficulties during the process, think aloud protocols are more reliable.

When the focus is on the reasoning for certain behaviors, and/or when it is important not to interrupt users, it is more suitable to use post-event protocols.

3.3.6 Setup of observational studies

In any data collection effort that involves observation of behaviors, users have to be engaged in some type of tasks, either their own tasks, or tasks assigned by researchers, while the data is being collected. Although this may sound obvious, the nature of the task and the context in which tasks are performed may have a strong impact on the interpretation of the data. Careful observation and analysis of real users in the context of actual use are invaluable (Wolf, Carroll, Landauer, John, & Whiteside, 1989). The major distinction in terms of study setup is laboratory study versus naturalistic study. In a laboratory study, some kind of experimentation is often designed and tasks are often assigned to the subjects, while in a naturalistic study, users normally perform their own tasks while researchers collect data.

Three important issues must be considered when designing laboratory studies: control, sampling and tasks. The gist of laboratory study is control, so sometimes laboratory study is also called controlled test. A laboratory study usually has a hypothesis that is tested through an appropriate experimental design by manipulating an independent variable and collecting data associated with dependent variables (Preece et al., 1994; Geisler, 2003). Although some of the same techniques are used to collect data (for example, video, audio, logging), as in naturalistic studies, the data that is collected is more rigorous and can be

analyzed quantitatively. If a test is carefully planned following the general experiment design principles (e.g., counterbalancing, randomization), statistical tests are often performed on the results to draw conclusions about the viability of the hypothesis. Because the number of factors that can practically be manipulated is limited, the controlled test is most often used to investigate very specific elements of a system or interface or to make general statements about particular interface principles. For example, Kelly and Fu (2006) designed a laboratory study to compare a term relevance feedback interface which displays terms in isolation with another interface which displays terms in the sentence context.

The fact that researchers have full control in laboratory studies is beneficial in several aspects. It allows the setup of the logging software that is required for data recording. It controls for the quality of Internet connection, hardware/software differences and creates an environment that is equal for all subjects. This makes it possible to compare the results between subjects and compute aggregate statistics. Conversely, if the study were conducted at subjects' own locations, researchers would not be to tell, for example, if a longer time spent on a task was due to pool search skills, or a slower network connection.

However, control over the study environment comes at the price of placing subjects at a search environment different than the one of actual use and this may change their behaviors (Wolf et al., 1989). For example, although Hargittai (2002) made efforts to allow for variation in subjects' computer experiences (e.g., they allowed subjects to choose between a PC and a Mac and to choose from three most popular browsers the one that they

were most familiar with), she noted requiring subjects to use a computer that is configured differently from the machine they usually use for browsing may influence the results, as certain settings (e.g., the default home page and bookmarks) are not equivalent to their own. She further noted that requesting users to travel to the study location affects response rates.

Moreover, control over the study environment also limits the generalizability of findings. Although certain results are found for a specific type of users under specific type of context (experiment environment, task, time constraint, and so forth), it is usually unknown if they can be generalized to the larger population in the real use condition.

In laboratory studies, sampling method also has an important impact on how results can be interpreted. Hargittai (2002) pointed out that an important limitation of many such studies is that they concentrate on the behavior of a small segment of the population by limiting participants to university faculty and students or long-term users from the information technology profession. Such sampling techniques limit the extent to which findings can be generalized to a larger segment of the Web user population. Preece et al. (1994) also pointed out that it is very hard to generalize results from laboratory experiments to other tasks, users or working environments.

The third issue is the choice of tasks. One option is to use assigned tasks to stimulate searches, as in Hargittai (2002). The tasks should be as natural as possible and should aim to mimic the problems which users are likely to encounter. Having people complete assigned search tasks while being observed is standard practice in laboratory studies of search

behaviors (Wildemuth, 2002). The advantage is again about the control: having multiple subjects complete the same task allows researchers to compare their performances and aggregate the results. However, this has several disadvantages. Subjects may not be motivated for assigned tasks, so some of them may not spend as much efforts as they would had the tasks been of importance to them. Subjects may have difficulty understanding what is required by the tasks. This is most problematic if the tasks involve some kind of relevance judgments. As subjects were not authors of the tasks, they may understand requirements of the tasks in a way different than the original authors, thus leading to inaccurate relevance judgments.

To summarize, laboratory studies allow researchers to have control over users and tasks so that a specific aspect of the design can be examined closely by observing real users in the context (although artificial) of real use. The price of the control includes the loss of some contexts of the search and limited generalizability of results. Wolf et al. (1989) summarize four aspects of actual use that a controlled experiment destroys: the motivational context (the person is not doing something of importance to them), the social context (in real use, people have a network of support to call on), the time context (lab studies usually do not let the subject work on something else, or try again the next day), and the work context (the person is doing your work, not theirs) and argue that the data generated in laboratory studies may be distorted reflections of the actual use.

Awareness of problems with laboratory studies encouraged researchers to explore techniques that collect data that reflects real usage more accurately, such as in the naturalistic environment. The critical characteristic of a naturalistic study is that users work in their normal working environment while performing some tasks and being observed. If the main purpose of the study is to observe user behavior, the study is often called naturalistic observation. By definition, naturalistic observation is an empirical method of study by which the researcher introduces no outside stimulus, instead witnessing behavior as it naturally occurs in the environment[1]. In a naturalistic observation, researchers take great care in avoiding making interferences with the behavior they are observing by using unobtrusive methods, without attempting to influence or control it. Therefore, the studies are often conducted in places like streets, homes, and schools. Thus, they are also called field studies.

Compared with laboratory studies, naturalistic studies have the drawback of higher time investment (researchers have to travel to the site of the user; it is only possible to conduct individual sessions; and the observations usually take place over extended periods of time) and lower chance of observing the behaviors of interest (Wolf et al., 1989) since researchers have little control over what the users do and how they do it. Data analysis can also be time consuming.

---

[1] http://en.wikipedia.org/wiki/Naturalistic_observation

An alternative setup to study users in their natural working environment is remote study. By using web-based communication techniques (e.g., screen share) and other tools (e.g., telephone, Web camera), it allows researchers to observe users remotely without incurring the complexity and cost of bringing them to a lab or traveling to their places. This makes it possible to have larger numbers of participants with more diverse backgrounds, and may add to the realism since participants do their tests in their own environments, using their own equipment (Shneiderman & Plaisant, 2004). The downside is that there is less control over user behavior and less chance to observe their reactions.

CHAPTER 4

PROBLEM DEFINITION AND RESEARCH OVERVIEW

The observations presented in Chapter 3 revealed several gaps in the current research on implicit feedback: the lack of studies focusing specifically on the implicit feedback for Web search, the lack of studies on a wider range of behavioral evidence other than clickthrough and time, and the lack of in-depth studies seeking to understand how each type of behavioral evidence relates to the searcher's interest. The empirical study described in the rest of this dissertation was designed to extend previous research and cast some additional light on these points. Section 4.1 describes the specific research questions and how they address the above gaps. It also defines the scope of the dissertation. Section 4.2 presents an overview of the study design.

4.1 Research questions and scope definition

The first gap identified in Section 3.2 is that few studies have focused on examining implicit feedback for Web search conducted through widely used general purpose commercial search engines; instead, many of the studies examined Web-based information seeking and discussed users' patterns of navigation across general Web page contents, not

necessarily associated with search. Web search is a distinctive information seeking environment. For example, for Web search implicit feedback, it is important to consider whether a piece of evidence for feedback is collected from behaviors on the results list page or external result content page because that distinction has significant impact on the implementation of monitoring techniques: if searchers' behaviors on the results list pages are the only valuable evidence, search engines can capture those behaviors much more easily than those on external results content pages since search engines have full control over the search results list pages while pages beyond them can only be tracked by client-side logging. Most previous studies focused on the search results list (e.g., Joachims et al., 2005; Shen et al., 2005a), while Jung et al. (2007) argued that it is important to collect data beyond the search results list and consider all pages visited in the entire search session. This dissertation aims to study searchers' behaviors in the normal Web search process. Although the study was conducted in a laboratory environment, the search environment was made as natural as possible. Searchers used their favorite search engine in their favorite Web browser and conducted the searches without any restriction or interruption (e.g., from the application of the think-aloud protocol). The study examines if the genre of the page affects the behaviors that can be captured and used. In particular, are the behaviors on the search results list page more useful than those on the result content pages?

Secondly, unlike many previous studies which focused on some particular types of behavioral evidence (mostly clickthrough and time), this study considers a wider variety of

behaviors and implicit measures of interests to support feedback for Web search. Table 4.1 summarizes the behavioral sources of implicit feedback that have been considered in previous implicit feedback studies which built predictive models based on a number of behaviors (such as Agichtein et al. 2006 and Fox et al., 2005), and studies of individual behaviors or measures, such as Joachims et al. (2005) on eye movements, Hijikata (2004) and Kerry and Fu (2007) on mouse movements, and White and Kelly (2006) on display time. Behaviors that were considered in implicit feedback studies on non Web search environments (such as Web browsing, catalog search, UseNet reading) are selectively included based on their applicability to Web search. Some other behaviors, *purchase*, *subscribe*, *translate*, *reply*, *link/cite/quote* and *forward* can also indicate the searcher's interest under certain special contexts (e.g., search for merchandise), but they were so rare in the general Web search that the decision was not to include them in the table.

Table 4.1. Behavioral sources of implicit feedback mentioned in the literature

Category	Behavior	Measure
Search	Submit query	Query length
		Number of search results pages
		Fraction of shared words between query and title, summary, URL, and domain
		Fraction of shared words with previous query
Select	Select results	Number of visits
		Time to first click
		Number of clicks to reach the page from the query
		Position of page in the results list
		Ranking of selected result on the results list page
		Absolute ranking of selected result
		Characteristics of the page (count of image, size of page, and number of scripts on page)
		Click on next result
		Click on previous result
Examine	View	Dwell time
	Scroll	Scrolled
		Scrolling count
		Average seconds between scroll
		Total scroll time
		Maximum scroll
	Eye movement	Eye fixations
		Eye movement patterns (reading versus browsing)
		Pupil dilation
	Mouse movement	Time spent moving the mouse
		Target text of text tracing, link pointing, link clicking and text selection
		Hesitation on links or text
	Mouse click	Follow links on result page
	Search within page	Searched within page (Ctrl-F)
	Exit page	Exit type (kill browser window; new query; navigate using history, favorites, or URL entry; or time out)
	Total amount of page activity	

Table 4.1. Behavioral sources of implicit feedback mentioned in the literature (continued)

Category	Behavior	Measure
Retain	Print	Presence/absence
	Bookmark	Presence/absence
	Email	Presence/absence
	Copy/Paste	Presence/absence

This table serves as the baseline model for this study. The central questions that are addressed in the study relate to the behavioral sources of implicit feedback: Which behaviors and measures listed in Table 4.1 are actually considered by human analysts? Are there any other behaviors or measures that have not been identified in previous research? Based on findings of the study, an updated model of implicit feedback for Web search will be presented.

Finally, the review of related work suggests that most of the studies found so far examine observable searcher behaviors as implicit interest indicators by somehow comparing them against explicit ratings. Unlike previous studies, this work does not focus on whether any single behavioral measure or combination of them correlates well with explicit measures of searchers' interests; instead, it seeks to gain better understanding of the process of inferring searcher interests from behaviors. Assuming a range of behaviors is observed by a human intermediary (such as a reference librarian or a search expert), which behavior(s) will she consider as evidence of interests? Does she use a single behavior or a set of behaviors to make the inference? Does more evidence consistently lead to more reliable inferences? Why does she believe that a certain behavior is useful? Are there any

rules that are commonly used? Answers to these questions do not only provide more evidence for the usefulness of behaviors as implicit feedback measures in Web search context, but also advance the understanding of why and how each type of evidence is useful. Such an understanding forms a foundation for improving search engine algorithms that exploit implicit feedback to deliver better results and create better user experience.

In sum, the following specific questions guide the study: (1) Which type(s) of searcher behavior is useful evidence of the searcher's interests? (2) How does the quality of inference about the searcher's interest evolve with more evidence available? Does more evidence consistently lead to more reliable inferences? (3) Does a single behavior indicate interest, or is it necessary to capture a set of behaviors? (4) Does the genre of the page affect the behaviors that can be captured and used? In particular, are the behaviors on the search results list page more useful than those on the result content pages? (5) Finally, why is a certain behavior useful? What are the rules to make the inference?

The scope of this dissertation is limited to the natural interactions that a searcher normally has with a typical high precision oriented search engine, such as Google and Live Search. This has two implications. First, there are other systems such as the exploratory search systems described in Marchionini (2006) that require complex or copious searcher interaction with results interfaces and support tasks other than high precision oriented retrieval. It is easy to conceive that searchers' behaviors when they interact with these systems are very different from those when they interact with Google; however they are

beyond the scope of this work. Second, behaviors incurred by special add-on interaction mechanisms designed to elicit user intention are not considered as implicit indications of interests. For example, White (2004) designed a search interface to actively engage searchers in the examination of search results. In addition to the full text, the results are also represented by a variety of snippets, such as titles, top ranking sentences extracted from the top 30 documents retrieved, and sentences in the document summary. When users interact with these representations, their behaviors are tracked and used to learn implicit feedback models. It may be the case that with the development of advanced search interfaces, some of the novel interaction features that are experimental today may become routine in the future and widely used by searchers. Here, the discussion is limited to the important and likely to continue to be used search engine interface that displays result summaries in a list and only allows searchers to click on the titles to navigate to result pages or modify their queries in the query box. Furthermore, it is only concerned with general text-based Web search. It does not consider searches over other properties, such as images and videos, nor does it consider specialized databases such as those for genomics or law.

The focus of the dissertation is on advancing the understanding of the relationship between the types of behavior that can be captured and the searcher's interests. Based on analyzing how inferences about the searcher's interests are made, the dissertation is expected to conclude with some common rules about such inferences and to put forward

design recommendations that can be applied in automatic systems. However, the actual implementation and evaluation of such algorithms are beyond the scope of this work.

4.2 Overview of study design

Research questions outlined above were addressed through a two-phase empirical study, summarized in Figure 4.1. The first phase was a laboratory study in which inexperienced searchers were paid to perform Web searches on assigned topics during 1-hour private lab study sessions. Logging and eye tracking techniques were used to collect recordings of searchers' behaviors during Web search activities. The outcome of this phase was a corpus of Web search cases from inexperienced searchers, with screen recording and eye tracking. From this corpus, a subset in which searchers experienced underspecification problems at the beginning and went through multiple rounds of query modification during the search process was selected. This resulted in a pool of search cases, and for each of the search case, four types of stimuli were created showing different types of behaviors during the search. The different types of stimuli corresponded to different experimental conditions in the second phase of the study.

In the second phase, reference librarians were recruited to analyze recordings of Web searches collected from the first phase. For clarity of the presentation, participants in this phase of the study are referred to as "analysts", as compared to the "searchers" who participated in the first phase of the study. Analysts examined search cases presented in

different types of stimuli and made inferences about searchers' interests based on behavioral evidence. The stimuli were presented as series of screen shots or video segments so that analysts' inferences and rationales were elicited at each screen shot or video segment. The data was generated from the second phase of the study. It consisted of analysts' inferences and rationales. The data was used for content analysis to inform the research questions.

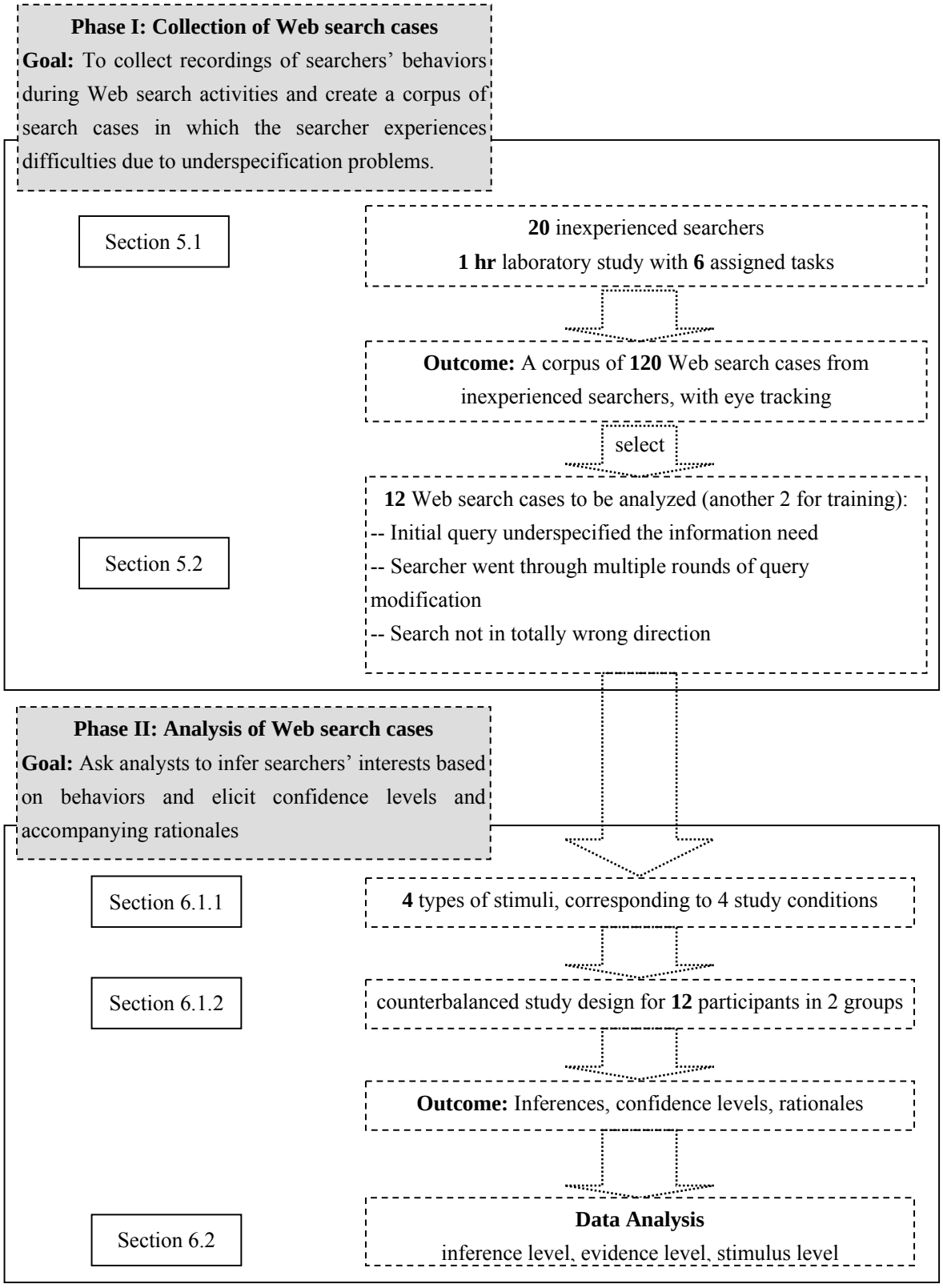


Figure 4.1. Overview of study procedure and structure of Chapters 5 and 6

CHAPTER 5

PHASE I: COLLECTION OF WEB SEARCH CASES

The goal of the first phase of the study was to collect recordings of searchers' behaviors during Web search activities and create a pool of search cases in which the searcher experiences difficulties due to underspecification problems. The recordings were then examined by search analysts in the second phase of the study.

5.1 Design of data collection

Issues involved in designing the data collection included: recruitment of searchers, choice of tasks, choice of data collection methods, the overall setup (i.e., the choice between collecting the data in a laboratory environment versus a naturalistic environment), and the procedure. Each of these issues is described in detail in the rest of this section.

5.1.1 Recruitment of searchers

The crucial task in recruiting searchers is to screen the candidates so as to get people who are more likely to experience search difficulties due to underspecification of information needs. The literature suggests that people are more likely to underspecify when they search in a new field and have complicated information needs involving multiple

aspects. Moreover, novice searchers are more likely to suffer from underspecification problems than experienced searchers due to their lack of knowledge about system vocabulary and/or system syntax and experiences on reformulating queries based on system feedback. These factors were taken into consideration when searchers were recruited.

A recruitment advertisement (Appendix A) was sent out via the UNC Mass Email to all UNC staff who opted into the email list. The email described the purpose of the study, its time and location, the compensation, and provided a URL to the online recruitment questionnaire. The questionnaire contained 12 screening questions, whose answers were used to screen respondents, and 3 demographic/contact questions (name, age, and email addresses). All people who completed the questionnaire were entered in a drawing for a $25 gift certificate to the UNC bookstore, no matter whether they were subsequently selected for the study or not.

Three strategies were used in the recruitment questionnaire to determine if a respondent was more likely to experience underspecification problems. First, respondents were asked to state their computer usage, Web usage, Web search frequency, and search skills. The questions, referred to hereafter as background questions, included (with options in brackets):

How long have you been using computers? [10 years or more, 7-9 years, 4-6 years, 1-3 years, less than 1 year]

How many hours do you use computer on a typical day? [4 hours or more, 3-4 hours, 2-3 hours, 1-2 hours, less than 1 hour]

How long have you been using the Web? [10 years or more, 7-9 years, 4-6 years, 1-3 years, less than 1 year]

How long have you been using search engines (such as Yahoo!, Alta Vista, and Google)? [10 years or more, 7-9 years, 4-6 years, 1-3 years, less than 1 year]

Which search engine do you use most often?

How often do you search typically? [more than 5 times a day, 1-5 times a day, a few times a week, every few weeks, less often]

How do you feel typically when you use search engines? [very relaxed, relaxed, a little nervous, stressful, very stressful]

When you search, how often do you find what you are searching for? [always, most times, sometimes, rarely, never]

Based on your experience, in general do you feel that using search engines to find information is [very easy, easy, neither easy nor difficult, difficult, very difficult]

In general, your experience with using search engines can be best described as [very satisfying, satisfying, neither satisfying nor frustrating, frustrating, very frustrating]

The second strategy consisted of a query formulation exercise. Four search problems were listed on the questionnaire. Respondents were asked to imagine that they would use a

Web search engine (such as Google) to find the information and asked to formulate a search query for at least one of the four search problems without conducting any search. Two sample search problems are given below.

> *Your friend is coming to visit you next week. You know she really likes Chinese cuisine. Please find a restaurant that you can take her to dinner during her visit.*
>
> *Your friend visited the Kennedy Space Center recently. When he was there, he watched a movie about the Apollo Project. The video included a segment showing President Kennedy announcing the lunar landing project. Your friend vaguely remembers that President Kennedy said something like the project was undertaken not because it was easy, but because it was difficult. Can you find the exact quote for what President Kennedy actually said and where he made the speech?*

All search problems involved multiple facets. To make them comparable, preferences were given to more close ended (fact finding or known item) search problems which involved about 3 facets. The first sample problem had a geographical facet (Chapel Hill) which is implied, in addition to the expressed topical facet (Chinese restaurant) and subjective quality facet (best). The second sample problem involved four facets, the subject (speech), the topic (lunar landing project), the person (President Kennedy), as well as the additional descriptors (such as the words "easy" and "difficult"; in the original speech, "easy" and "hard" were actually used). Given the inaccurate description of the quote, the

challenge is to find the correct vocabulary to describe the problem. There were many Web pages which contain the exact quote, but they used very different vocabularies to describe the context of the speech. For example, instead of saying “lunar landing project”, some described it as “moon landing project” or “the Apollo project”.

Two other principles were used in creating these search problems. First, all of them presented simulated task scenarios (Borlund, 2000; White et al., 2002b), instead of directly describing the topics themselves, as the TREC topics do. Simulated tasks are short search scenarios that are designed to reflect real-life search situations and allow searchers to develop personal assessments of relevance. This is believed to have several benefits. Simulated task scenarios position the searchers within a realistic context and help generate natural behaviors. Participants can provide their own interpretations of what information is required, what search strategies should be used, and which results are relevant. This approach also discourages searchers from simply choosing the exact phrase out of the problem description, a phenomenon observed in some laboratory studies involving assigned search tasks (e.g., Fu, Kelly, & Shah, 2007).

Second, whenever appropriate, search problems were presented in altruistic contexts in which the searcher was asked to look for information to help another person. It was generally regarded that participants became more motivated when the problems were described this way (M. Stone, personal communication, June 2006).

The third screening strategy asked respondents to describe a search they had done recently that was not fully successful. They were asked to describe what they were looking for (the search problems) and what they had done (the search strategies and queries). This screening strategy served two purposes. It helped to identify the kind of problems a respondent had so that respondents with underspecification problems could be recruited. In the meantime, problems that were due to underspecified queries could be used as tasks in the study.

When respondents were screened, most attention was paid to the last four background questions which were about past search experiences. An ideal participant was one who found it difficult to use search engines, normally felt stressful when she searched, and often got frustrated by unsuccessful searches. For respondents who met these criteria, their responses to other background questions were used as sanity checks. Previous studies (e.g., White, 2004) showed that experienced searchers found using Web search engines significantly easier than inexperienced searchers. Therefore, it was expected that respondents who reported generally negative search experiences (difficult to use, stressful, frustrating, unsuccessful) should also report middle to low levels of experience with Web searches.

Respondents who were selected from the first strategy were further screened based on their responses to the query formulation exercise. As all search problems involved multiple facets, queries missing one or more of the facets were considered underspecified.

Respondents who formulated an underspecified query or queries in addition to reporting negative search experiences were recruited first.

It must be acknowledged that none of the three screening strategies was guaranteed to get participants who would definitely suffer from underspecification problems during the study sessions. The purpose of the screening process was to increase the likelihood that such cases might be observed and collected in the study. Effort was also made to select respondents who indicated the same favorite search engine so as to control the possible impact of search engine on the searcher's behavior. In total, 34 respondents were selected and invited to participate in the study, of whom 22 participated. All 22 participants indicated Google as their favorite search engine except for one, who had used Google before, but mentioned Yahoo as the favorite search engine. The 15 women and 7 men ranged in age from 22 to 57, including 4 in their twenties, 6 in their thirties, 7 in their forties, 5 in their fifties, and the mean age of was 40.41. All of them had used computers for 10 years or longer and were using a computer at least 2-3 hours a day. All but 6 participants had used the Web for 10 years or longer while their search engine use experience represented a balanced mixture between 4-6 years and more than 10 years. Participants' search frequency averaged a few times a day. Their average perception towards search engine use was between neutral and satisfactory, as determined by the 4 questions on search experiences. In general, the participants represented a relatively less search-savvy sample from an academic institute where the average education level is high.

5.1.2 Tasks

The tasks were designed to encourage naturalistic search behavior by the participants. Each searcher worked on about 6 search problems during the study (depending on her pace), coming from two sources. The first source of tasks was a collection of search problems that the investigator maintained, which included the search problems used in the query formulation exercise in the screening questionnaire. When tasks in this category were used in the study, preference was given to those on which the participant formulated underspecified queries when she completed the screening questionnaire.

The second source was the collection of search problems obtained from the third screening strategy described above. Search problems contributed from all respondents, including those who had not been selected to participate in the study, were examined by the investigator. Those problems for which difficulties were likely to be caused by underspecified queries were selected to form a pool of search problems. Similar to the search problems used in the query formulation exercise on the screening questionnaire, selected tasks were multi-faceted search problems with fairly close-ended answers. For some tasks, the investigator modified the contributed search problems slightly and made up the scenarios.

Answers to some of the search problems could be personalized. For example, when a search problem asked the searcher to find a good restaurant, the searcher needed to

determine where the restaurant should be located based on where she lived. A complete list of the search problems is provided in Appendix B.

5.1.3 Data collection techniques

Observation techniques were mainly used to collect the data, as the goal here was to construct a pool of Web search cases which could later be examined by human analysts in the second phase of the study. In addition, a structured interview (described in Appendix C) was administered before each search to elicit the searcher's familiarity with the search topic, registered on 7-point Likert type scale, and a semi-structured interview (described also in Appendix C) was conducted after each search in which the searcher was asked to reflect on the search process, focusing on two aspects. First, did she think her initial query clearly stated what she wanted? Second, did she learn something in the search process which made her change her search strategy? If so, what were some of the critical instances which triggered the change? These data were later used when search cases collected in the first phase of the study were screened for use in the second phase.

The rest of this subsection will discuss observation techniques that were used to collect the search cases. Seven types of observation techniques have been identified in the literature review, including direct observation, logging (server-side logging, logging via proxy server, and client-side logging), eye-tracking, mouse tracking, physiological measures, video taping, and verbal protocol analysis. The selection of methods has been based upon the suitability of each method to capture the types of behavior of interest. Here,

the behaviors of interest include view, scroll, mouse-over a link, click, search within page, and query modification. Direct observation is too coarse to capture the intricacies of the actual behavior. Obtaining physiological measures is intrusive; additionally, physiological measures are mainly used to indicate the searcher's cognitive load, which is not the focus here. For similar reasons, verbal protocol analysis (which is used to capture the searcher's mental activities) is not relevant either. Although mouse-tracking can unobtrusively capture some of the scrolling activities (if done by mouse), the link hovering behavior, and the link clicks, the huge quantity of low level data on the mouse position and the objects under the mouse pointer is difficult to interpret without the help of computer analysis tools. Given the purpose of this phase of data collection and the way the data will be used, logging, eye tracking, and video taping are considered to be the most appropriate methods.

A Tobii 1750 eye tracker running Clearview software was used for the purpose of data collection. The eye tracker was embedded in a 17-inch screen set to a resolution of 1024x768. The Clearview software saved the time, URL and a screenshot for every page visit during the study as well as the eye positions every 200 milliseconds. It also recorded the screen contents into a video. With the saved eye positions, it was able to generate a video recording of the search session with eye gaze overlaid on top of the screen contents.

The data collection was conducted in a lab located in the School of Information and Library Science (Manning Hall) on the UNC campus. Although this sacrificed the natural search environment to which the searcher was accustomed, the arrangement was necessary

to use the eye tracking setup. Efforts were also made to minimize the difference in computer setup. For example, left-handed searchers were provided with left-handed mice and searchers were encouraged to use their favorite browser in the study. As the data were not used to compare the searcher's performance or the effectiveness of the search engine, but collected to capture searchers' behaviors while performing real search tasks, the laboratory setup should not have much negative impact on the data. Figure 5.1 is a picture of the study room. On the left is the table for the participant, with the Tobii eye tracker. On the right is the table for the investigator with a regular computer monitor. A dual monitor, dual keyboards/mice setup was used. A microphone was hung to the file cabinet in the middle to capture the voice.

Figure 5.1. Study room setup

5.1.4 Procedure

In each one-hour session, the investigator first provided a verbal overview of the study (Appendix D), answered any questions that the participant had, and obtained the informed consent from the participant. Eye tracker calibration was performed and the recording devices were started when the participant was ready to start. She then completed about 6 tasks in sequence. One of the tasks was the one that the searcher contributed in response to the third screening strategy. The rest were assigned by the investigator. For participants who did not respond to the third screening strategy, all tasks were assigned. For each search task, four steps were completed as follows.

First, search problems were read to the participant and repeated as necessary, but no clarification was offered. Search problems were read to minimize searchers' head movements, an action potentially causing problems to the eye-tracking system.

Second, the participant was asked to verbally indicate her familiarity with the search topic on a 7-point Likert type scale and the investigator wrote down the answer.

Third, the participant searched for the topic using Google. The investigator sat next to the participant during the entire session, but the participant was asked to work alone and not to talk to the investigator while searching, unless she was unclear about what to do. There were no restrictions on what queries the participant might choose, how and when to reformulate the query, or which links to follow. It was totally up to the participant to do whatever she thought she needed to in order to complete the task.

Finally, as soon as the participant felt she was done, or was ready to give up, she was instructed to close all additional browser windows that were opened during the search and brought the main browser window to the home page (which was set to be the Google home page before the study). This made it easier to delimit search sessions at the data cleaning step. The participant then signaled the investigator that she was ready to move on to the next task. At this point, the investigator discussed with the searcher what had been found as a check to make sure that she had made an earnest effort on the task. Then the semi-structured interview described above was administered.

When the study time was up, the participant was debriefed and compensated. The data were exported from ClearView to generate the video recording with eye movements.

5.2 Selection of search cases and preparation for Phase II

This subsection will first describe the pool of 118 search cases that were collected from this phase very briefly. A detailed analysis of the data set is not the goal of this dissertation. Instead, the focus will be on describing the selection process and characterizing the cases that were selected to be used in the second phase of the study.

5.2.1 Cases collected from Phase I

In total, the 22 participants completed 118 searches (mean=5.36, standard deviation=1.79). Among the 118 searches, 19 were from the searchers themselves[1]. The searchers had an average[2] familiarity of 2.60 on the search topics on the 7-point scale, suggesting that they were indeed not familiar with most of the search topics. For tasks coming from the searcher (i.e., own tasks), the average familiarity was much higher at 5.00. Ruling these 19 cases out, the remaining 97 search cases had an average familiarity of 2.13. The Mann-Whitney test suggested that searchers' familiarities with their own search topics were significantly higher than those with assigned topics ($U=230.0$, $z=5.418$, $p<0.001$).

Despite the fact that selected participants claimed to have relatively less experience with search engines and have difficulties when searching, their overall search performance (whether the answer was found, how much time was taken, and how efficient the search strategy was) was better than expected in the study. It was common that searchers were able to find the information they needed with one or two queries. Although no formal test was conducted, it seems to be the case that topic difficulty had a larger effect on search performance than the searcher, especially when search topics were difficult. In other words, for some difficult search topics, all searchers performed almost equally poorly; while for

---

[1] If a searcher searched on a topic that she contributed with respect to the third screening strategy, it was counted as a task from the searcher. There were 19 such cases. If a searcher searched on a topic that another participant contributed, it was counted as an assigned task.

[2] The average was based on 116 cases because the searcher did not indicate familiarity in 2 cases.

some easier topics, searchers performed differently. For example, the hardest search topic that was used in the study is Search Topic e:

> *My nephew is doing a school project on the deaf population. He wants to find out how many deaf people in the U.S. speak English, and in the same time, use the American Sign Language. Can you help him?*

This question came from a response to the screening questionnaire (the context was added by the investigator). Twelve participants searched this topic, but none of them found the answer, although some of them spent up to 30 minutes on it and used as many as 13 queries. The difficulty of this topic can be attributed to the seemingly non-existence of the data, at least on the shallow Web. One of the Web pages that some searchers found pointed out that "There is not an official statistic on how many persons with hearing loss or deafness live in this country; the U.S. Census Bureau stopped including deaf demographics in 1930. Individual surveys are rarely conducted, and they are not done on a large enough scale." Appendix E lists all the queries used by searchers who searched this topic. Some of the searchers were able to make adjustments to their queries, by adding search terms such as "usage", "statistics" and "census" which did not appear in the search problem, but had the potential of leading to good results; however, none of the queries led to results which could answer the question.

On the other hand, the quality of the search strategy did make a difference for some other topics. For example, Search Topic b says:

Your friend visited the Kennedy Space Center recently. When he was there, he watched a movie about the Apollo Project. The video included a segment showing President Kennedy announcing the lunar landing project. Your friend vaguely remembers that President Kennedy said something like the project was undertaken not because it was easy, but because it was difficult. Can you find the exact quote for what President Kennedy actually said and where he made the speech?

A searcher (064) was able to find the answer with just one query ([“apollo project” “president kennedy” “lunar landing” easy hard speech]) and examining two results while another searcher (206) issued four queries and clicked on 8 results before finding the answer. The first two queries that Searcher 206 used ([apollo project john kennedy speech] and [john kennedy quote about apollo project]) underspecified the search question by missing the keywords that were given in the scenario. Given that President Kennedy had given multiple speeches on the Apollo Project and that most of the participants did not use the search facility (Ctrl-F) within the browser, it took them much longer to reach the information that they needed.

Compared with the “topic effect”, the “searcher effect” seems less prominent, probably because searchers in the study were selected after the screening so that they were more homogeneous. No searcher performed poorly on all topics; instead, they had

difficulties on different topics, except for a few who completed almost all searches smoothly.

Before moving on to discuss cases which started with underspecification, it is interesting to note some patterns on how searchers started searching on multi-faceted topics. First, some searchers were not able to distinguish key concepts in the search topics from unimportant concepts describing the context. They mistakenly put contextual concepts into the search query, which led to irrelevant results. For example, a searcher included "kennedy space center" in the query for Search Topic 2 (Kennedy quote); another searcher included "history museum" in the query for Search Topic 7 (ATC spur). Although it is clear that terms like "kennedy space center" or "history museum" should not be included in the queries, there are situations where it was less clear whether certain concepts that are mentioned in the search topic should be kept in the query. For example, for Search Topic 4 (Roy Williams quote), "Roy Williams" seems to be an important concept in the scenario. However, as Roy was actually not the original source for the quote, putting his name in the query was not helpful for finding the author of the quote; instead, putting "author" in the query is a better strategy. Interestingly, one of the cases that were selected later to be examined in the second phase was on this topic and the searcher's inclusion and exclusion of "Roy Williams" was used by the analyst as an important clue to infer what the searcher was looking for.

Secondly, some searchers started the search with intentionally underspecified queries. They stated in the post-search interviews that they had done so to first acquire some knowledge on the general topic with which they were not familiar. For example, a searcher used “north central arkansas” as the first query for Search Topic j; another searcher used “june bugs” as the first query for Search Topic f. Related to this is the building block strategy, in which the searcher broken down the search topic into smaller questions and looked for the answer to each question first. An example was a search for Topic e in which the searcher first looked for the number of deaf people in the U.S. who use the American Sign Language and the number of deaf people in the U.S. who speak English.

Finally, when search topics were rather complicated involving more than 3 facets, searchers often posed over-specified queries, but found that few results would contain all the keywords in the query. Then, they would take out less important concepts, or concepts that had been implied by other concepts. For example, for Search Topic j, Searcher 64 found that when “biofuel” was already in the query, he would not need additional terms such as “crop type”.

This summary of the searches illustrates the diversity of search strategies even among this relatively homogeneous sample of searchers. Future work is planned to examine the 118 searches more closely to better understand how people search multi-faceted questions.

5.2.2 Selection of cases

The purpose of this step was to select 12 cases from the collection of 118 search cases which would be analyzed in the second phase of the study. Twelve cases were needed to populate the experimental design for the second phase, as will be explained in Section 6.1.3.

To make the selection, all 118 cases in the collection were reviewed by the investigator. Review of the 19 searches on searchers' own topics revealed that searchers' behaviors were somewhat artificial when they searched on topics that they had searched before: some searchers tried to repeat the search they had done and explained to the investigator what had gone wrong (although they were told not to think aloud during the search); many searchers recognized pages that they had visited before, thus were able to make judgments about the pages without clicking on them or by spending a much shorter time. The original intention of having participants search on their own topics was to collect search cases where searchers fully understood the context and were truly motivated to look for the information; however, considering that searchers were significantly more familiar with topics they generated than topics assigned to them by the investigator, as noted previously, and that the searchers had been "contaminated" by their prior search experiences, the inclusion of such cases would be unfair and would likely confuse the

analysts in the second phase of the study. Therefore, the selection was made on the remaining 99 cases.

The investigator first pruned off search cases which started with natural language queries that literally described the search topics. Considering that the selected cases would be analyzed by human analysts and the goal would be to infer the searcher's interests based on behavioral evidence, the presence of such queries would ruin the experimental design. An example of such a query is [how much vodka does a person in russia drink on average] for Search Topic i (Searcher 169). Next, search cases in which searchers moved in the wrong direction (e.g., they misunderstood the search topic, or did not exert good effort to find the answer) were ruled out so that the selected cases would not cause undue confusion for the analysts.

For each of the remaining cases, a judgment was made on whether underspecification occurred during the search. Preference was given to cases in which the initial query underspecified the information need (missing some of the facets), and then the searcher went through multiple rounds of query modification (by adding or changing query terms) and/or browsed through many result pages before the information was found or the searcher gave up. The topic familiarity data and searchers' reflections on the search processes were used to aid the selection by giving priority to cases where searchers were less familiar with the topics and indeed felt that initial queries were underspecified. This ruled out cases where initial queries were rather good or the information was found only by

browsing. Finally, whenever possible, search cases were selected so that they were from different searchers.

As a result, 12 cases were selected from 10 searchers on 8 different topics (Search Topic a-h). There were 4 topics which were covered by 2 searches to satisfy the experimental design of the second phase, as will be explained later. Three cases on 3 topics (Topic h, j and k) were selected to be used for training in the second phase.

Table 5.1 summarizes the 12 search cases that were analyzed in the second phase. They ranged from about 1.5 minutes to 12.5 minutes long, with an average length of 327.2 seconds, or about 5.5 minutes. The searchers used an average of 3.7 queries and viewed results 5.25 times[1] during a search.

[1] A searcher may examine the same result more than once during a search. The multiple views were counted separately.

Table 5.1. Selected search cases

ID	Topic	Searcher	Duration (sec)[1]	#Queries	#Result views
1	a. Chinese restaurant	219	89	3	3
2	b. Kennedy quote	188	310	3	3
3	g. ATC spur	137	159	4	2
4	c. Curling	036	364	8	5
5	d. Roy quote	219	154	4	3
6	e. Deaf comm.	108	753	5	9
7	h. Gas price	169	185	2	3
8	f. June bugs	165	277	3	3
9	f. June Bugs	172	323	3	6
10	b. Kennedy quote	206	394	4	8
11	d. ATC spur	169	236	3	4
12	c. Curling	072	684	2	14

Table 5.2 lists the initial query posed in each case as well as how queries were later modified. Initial queries are highlighted in bold. A quick examination of these queries would lead to the observations that they were not only shorter in length, but more importantly, they all missed some facets in the search topics. For example, the initial query for the first case missed the quality aspect; that for the second case missed the question qualifier "quote"; the third case missed two topic qualifiers "spur" and "American Tobacco Company". These observations were confirmed by the searchers' reflections on their own searches. For example, when Searcher 219 in Case 1 was asked the question "Do you think your initial query clearly stated what you wanted?", she answered "Not really. It came up

[1] The time is the sum of the length of all the search segments. Because extended page loading time and query typing time between two segments are trimmed off when creating the stimuli (as will be explained in Section 6.1.1), the duration of search reported here can be slightly shorter than the actual time the searcher spent.

with a list of restaurants, but didn't tell me if any of them were good." When Searcher 169 in Case 11 was asked the same question, she answered "The first one? No! Because it just gave me more information about the American Tobacco Company than its foundation or its history … It gave me more updated like what is going on with the American Tobacco Company properties now than that historical facts." In response to the question on critical instances in the search, she also explained the reason why she modified the query from [american tobacco company and the railroad] to [railroad spur and american tobacco company]. She said "… American Tobacco Company and railroad, it just gave me information about the railroad that runs behind it or near it, so I had to be more specific in the information I put into the search."

Table 5.2. Queries on selected search cases

Case	Topic#	Queries
1	a. Chinese restaurant	**[chinese restaurant chapel hill]** [chinese restaurant chapel hill best] [chinese restaurant chapel hill best voted]
2	b. Kennedy quote	**[lunar project president kennedy easy]** [lunar project president kennedy easy quote] [We choose to go to the moon. We choose to go to the moon in this decade... not because they are easy but because they are hard; - John F. Kennedy, 1962]
3	g. ATC spur	**[railroad durham, nc tobacco]** [railroad spur durham, nc tobacco] [railroad spur durham, nc] [railroad spur durham, nc american tobacco]

Table 5.3. Queries on selected search cases (continued)

4	c. Curling	**[ice sports]** [description of ice sports] [types of ice sports] [type "ice sport"] [ice sport with brush] [use of brush in curling] [use of brush "curling ice sport"] [rules for curling sport]
5	d. Roy quote	**[be led by your dreams quote]** [be led by your dreams quote roy Williams] ["be led by your dreams" quote] ["be led by your dreams" quote author]
6	e. Deaf comm.	**[+"american sign language" +deaf +population +speak]** [+us +deaf +population +"american sign language" +speak] [+us +deaf +population +statistic +"american sign language" +speak] [+us +"american sign language" +speak +english] [+deaf +statistic +us +"american sign language" +speak +english]
7	h. Gas price	**[gas prices in european countries]** [current gas prices in european countries]
8	f. June bugs	**[canine eating bugs]** [dog eating june bugs safety] [june bugs toxic]
9	f. June Bugs	**[june bugs]** [june bugs toxicity dogs] [june bugs harmful dogs]
10	b. Kennedy quote	**[apollo project john kennedy speech]** [john kennedy quote about apollo project] [john kennedy quote about apollo project was easy but difficult] [not because they are easy but kennedy speech was where]
11	d. ATC spur	**[american tobacco company]** [american tobacco company and the railroad] [railroad spur and american tobacco company]
12	c. Curling	**[ice sports]** [ice sports tools]

CHAPTER 6

PHASE II: ANALYSIS OF WEB SEARCH CASES

This chapter describes the second phase of data collection and its results. It begins with a description of the study design and procedure. Results of the study are then presented, which include characteristics of the analysts, descriptive analyses of data and findings from analyzing the inferences and accompanying rationales from three different levels.

6.1 Methods

The goal of the second phase of data collection was to capture the process in which analysts review recordings of previous searches and make inferences about the searcher's interests based on the observed behaviors. These searches were selected as stimuli because the searcher experienced difficulties due to underspecification problems. Analysts were asked to think aloud during the review process, which was recorded. All the data collection sessions were conducted individually.

6.1.1 Creation of stimuli

After the 12 search cases were selected, four types of search recordings were created for each case as the stimuli for this phase of the study. As will be explained below, different

types of search recordings showed different aspects of the search behavior and they corresponded to four experimental conditions.

The stimuli took one of the two formats, screen shot format or video format, and each format was further divided into two types. Figure 6.1-6.3 show sample pages in the screen shots format, Type A. This type of stimuli displayed screen shots of the Google results list pages (including subsequent pages when applicable) that were returned from all the queries used in the session and the external result content pages visited after each query. If a searcher followed links on external result content pages, these clicks were not displayed. In other words, only links one step from the search results list page were considered. This was used to mimic the clickstream data that is typically captured in the server-side log of a search engine.

On the interface, the left frame served as the "table of contents" for screen shots that were displayed at the right frame. Each item in this frame corresponded to the screen shot of a page that the searcher visited during the search. The sliding bar could be pulled down so that the screen shots were displayed one by one. For screen shots of Google results list pages (as in Figure 6.1), the right frame displayed the screen shot in its original size; while for screen shots of external result content pages (Figures 6.2 and 6.3), the right frame was divided into two sub-frames. The upper right frame displayed the URL, title, and keywords (assigned by page authors in the META tag) of the page and the lower right frame displayed the screen shot. When a link to an external result content page was clicked on the left frame,

a thumbnail of the screen shot (Figure 6.2) would first be displayed to fit the size of the lower right frame (i.e., height of screen shot equal height of the frame). This meant that, for longer pages, the text on the thumbnail might be illegible. However, this would provide the analyst with an opportunity to look at the structure and layout of the page before enlarging the image (Figure 6.3) to examine the page content.

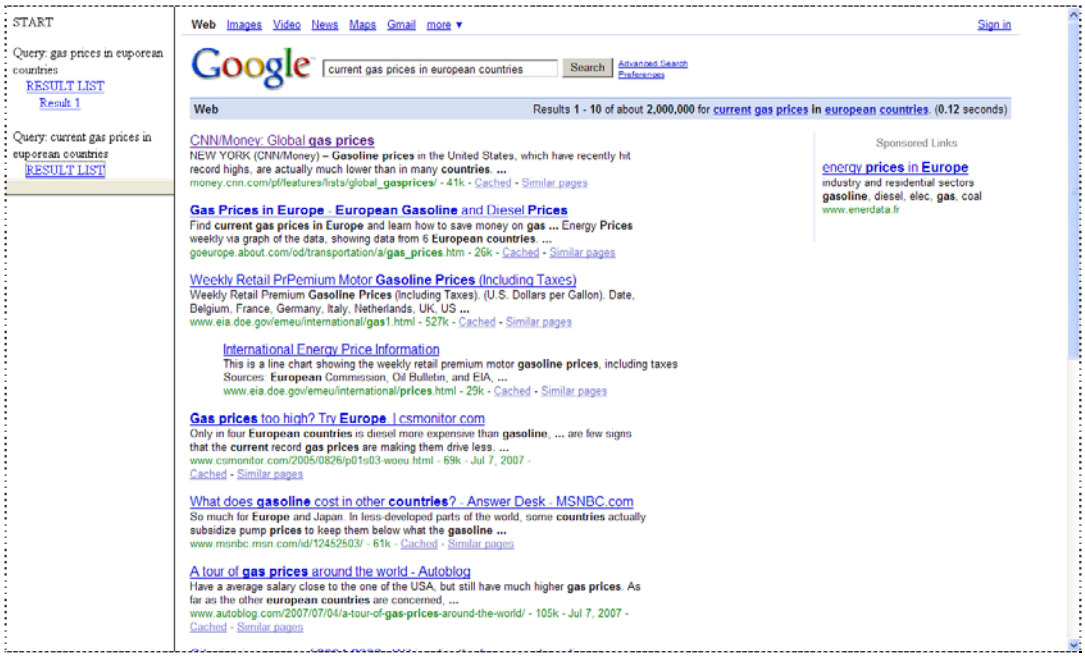


Figure 6.1. Sample page of Type A stimulus (screen shot format for Google results list page)

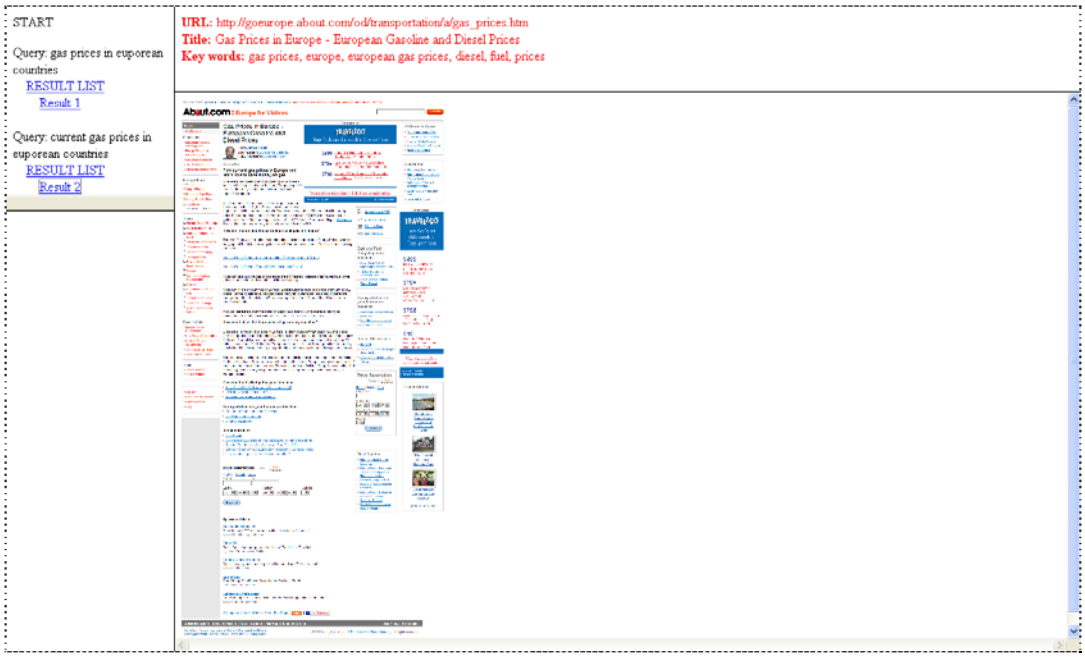


Figure 6.2. Sample page of Type A stimulus (thumbnail screen shot format for external result content page)

Figure 6.3. Sample page of Type A stimulus (screen shot format for external result page)

A Type B (Figure 6.4) stimulus not only provided the same information that Version A provided (queries, screen shots of search results list pages, and URL, title, keywords and screen shots of clicked results linked from the results list pages), but also included all the pages that the searcher visited by following links on external results content pages. In other words, it took into account the browsing activities in the entire session. It also displayed the amount of time that the searcher spent on each external results list page.

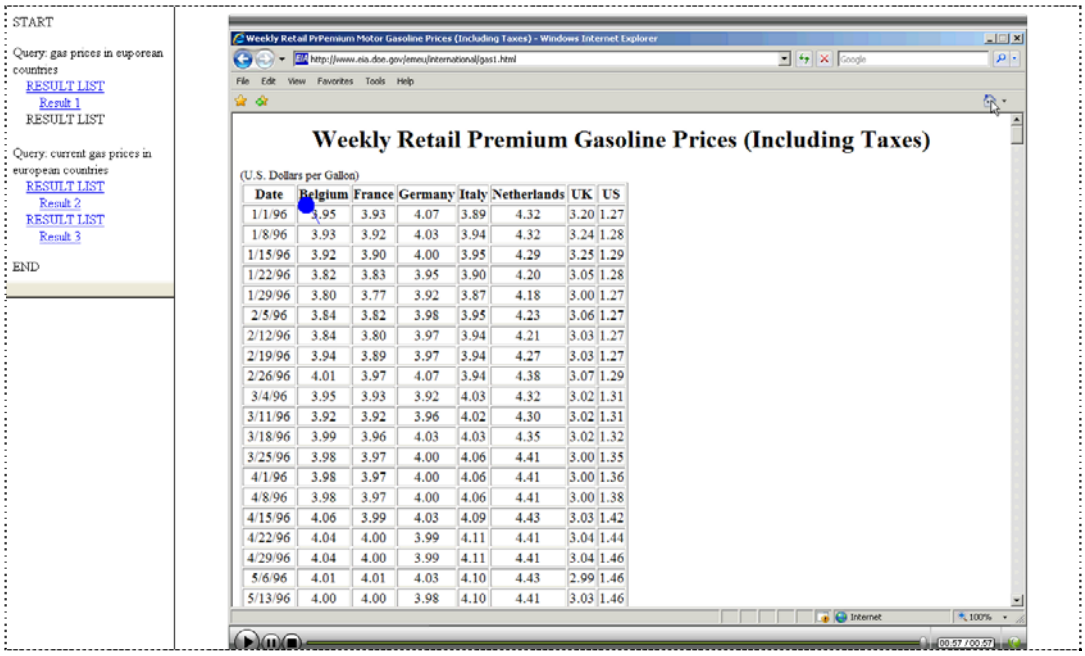
Figure 6.4. Sample page of Type B stimulus (thumbnail screen shot format for external result page)

Figure 6.5 displays a sample page of a Type C (video format) stimulus. It was similar to the screen shots format, except that video segments of the searches were played, instead of screen shots. The segments were created by making a cut to the original ClearView recording of the search session whenever one of these events happened: (1) a

query was issued; (2) a result was selected; (3) the searcher returned to Google results list page after examining a result. These events were chosen because they represented more critical instances in a search when changes were more likely to be made to search strategies. When the cuts were made, extended page loading time and query typing time were trimmed off, considering that analysts would be unlikely to learn anything about the searcher's interests by, for example, spending 5 seconds watching an external result content page being loaded or 15 seconds watching the searcher entering a long query. In addition to all the information available in Type A and B stimuli, a Type C stimulus also captured the scrolling behavior and the mouse movements within pages. The temporal nature of video recordings should also make the time spent on each page more salient to the analysts.

Figure 6.5. Sample page of Type C stimulus (video format)

Finally, a stimulus of Type D (Figure 6.6) was similar to Type C except that video segments in Type D also displayed the searcher's eye traces (the blue dot and line on the table).



Date	Belgium	France	Germany	Italy	Netherlands	UK	US
1/1/96	3.95	3.93	4.07	3.89	4.32	3.20	1.27
1/8/96	3.93	3.92	4.03	3.94	4.32	3.24	1.28
1/15/96	3.92	3.90	4.00	3.95	4.29	3.25	1.29
1/22/96	3.82	3.83	3.95	3.90	4.20	3.05	1.28
1/29/96	3.80	3.77	3.92	3.87	4.18	3.00	1.27
2/5/96	3.84	3.82	3.98	3.95	4.23	3.06	1.27
2/12/96	3.84	3.80	3.97	3.94	4.21	3.03	1.27
2/19/96	3.94	3.89	3.97	3.94	4.27	3.03	1.27
2/26/96	4.01	3.97	4.07	3.94	4.38	3.07	1.29
3/4/96	3.95	3.93	3.92	4.03	4.32	3.02	1.31
3/11/96	3.92	3.92	3.96	4.02	4.30	3.02	1.31
3/18/96	3.99	3.96	4.03	4.03	4.35	3.02	1.32
3/25/96	3.98	3.97	4.00	4.06	4.41	3.00	1.35
4/1/96	3.98	3.97	4.00	4.06	4.41	3.00	1.36
4/8/96	3.98	3.97	4.00	4.06	4.41	3.00	1.38
4/15/96	4.06	3.99	4.03	4.09	4.43	3.03	1.42
4/22/96	4.04	4.00	3.99	4.11	4.41	3.04	1.44
4/29/96	4.04	4.00	3.99	4.11	4.41	3.04	1.46
5/6/96	4.01	4.01	4.03	4.10	4.43	2.99	1.46
5/13/96	4.00	4.00	3.98	4.10	4.41	3.03	1.46

Figure 6.6. Sample page of Type D stimulus (video format, with gaze path)

The four types of stimuli correspond to the four experimental conditions. The potential differences between Type A and Type B reflect the value of monitoring the additional browsing paths beyond the search results list page as well as keeping track of the time spent on each page. The differences between Type B and Type C attest to the usefulness of capturing searchers' behavior within a page (such as scrolling and mouse movement), in addition to recording page level information, such as URLs. The differences between Type C and Type D will show the added value of eye tracking.

6.1.2 Recruitment of search analysts

Participants in this phase of the study were 12 search analysts who examined the search cases and inferred searchers' interests based on their behaviors. The expertise that is crucial to complete the task consists of experiences with observing people's Web search behaviors and skills to help searchers improve their search strategies. Reference librarians, who had high levels of expertise in those areas, were recruited through advertisements sent to listservs at the reference departments of several public and academic libraries in the Research Triangle Park area and individual invitations sent to a few potential participants with whom the investigator had personal contacts. In the recruitment advertisement, the possession of the above expertise was mentioned as the inclusion criterion.

6.1.3 Procedure

Overall, the task for the analysts was to view recordings of the searches and try to infer searchers' interests based on the evidence they discovered from the screen shots or the video segments, including queries and behavioral sources of implicit feedback. Descriptions of the original search problems were not shown to the analysts.

A study session lasted for about 2 hours. It started with a brief overview of the study (script in Appendix F) in which the investigator first explained what the participants would be asked to analyze and what their goal was. In particular, it was emphasized that the aim of the study was to understand how inferences would be made about searchers' interests and

what evidence would be helpful; therefore, analysts were requested to think aloud and always provide the best inference that she could make no matter how confident she was. The second purpose of the overview was to provide the analysts with information on the context of the searches that they were going to analyze. Specifically, it was pointed out that: the searches were done by paid searchers from the UNC staff who claimed to have less experience with Web search; searches were done in a lab with the investigator sitting next to the searcher, but without thinking aloud; search tasks were assigned and described in a scenario; all search problems were multi-faceted, close-ended with certain expected answers, although some of answers could be personalized. The analysts were also told that their focus should be put on inferring the search question embedded in the scenario correctly while it was less important to figure out the exact context around the question.

After the verbal overview, the investigator walked the analyst through a training task and pointed out the range of behaviors that could be considered. At the end of the training task, the original task description was disclosed to give the analyst another opportunity to understand the kind of search tasks that should be expected. After that, each analyst worked independently on 4 search cases, in the way described below and summarized in Table 6.1.

Table 6.1. Experiment design, showing assignment of analysts to search cases and types of stimulus. Each cell displays stimulus type and case ID.

		First Visit			Second Visit			
		Training	First Hour	Second Hour	Training	First Hour	Second Hour	
Analyst	1	B	A1, A5	B2, B6	C	C3, C7	D4, D8	First Group
	2	B	B6, B2	A5, A1	C	D8, D4	C3, C7	
	3	C	C2, C6	D3, D7	B	A4, A8	B1, B5	
	4	C	D7, D3	C6, C2	B	B5, B1	A8, A4	
	5	B	A3, A7	B4, B8	C	C1, C5	D2, D6	
	6	B	B8, B4	A7, A3	C	D6, D2	C5, C1	
	7	C	C4, C8	D1, D5	B	A2, A6	B3, B7	
	8	C	D5, D1	C8, C4	B	B7, B3	A6, A2	
	9	D	D9, D10	D11, D12				Second Group
	10	D	D10, D9	D12, D11				
	11	D	D11, D12	D9, D10				
	12	D	D12, D11	D10, D9				

The 12 participants were divided into two groups. The first group of eight analysts attended two sessions of the study and reviewed eight search cases in total, two for each type of stimulus. The two sessions were scheduled on two different days with minimum interval in between. For example, Analyst 1 examined Type A stimuli of Search Cases 1 and 5 and Type B stimuli of Search Cases 2 and 6 at her first session, then Type C stimuli of Search Cases 5 and 6, and Type D stimuli of Search Cases 7 and 8. The two cases of the same type of stimuli were always evaluated together, thus forming 4 pairs, but the within-pair order (i.e., the order of the first and the second case of the same Type, such as Search Case 1 and Search Case 5) was counterbalanced. The between-pair order (i.e., the order of stimuli type) was also counterbalanced. The eight search cases (first eight in Table

6.1) were on eight different search topics to avoid a learning effect. When Type A and Type B stimuli were used in a session, the training was done using a Type B stimulus (on a topic different than any of the eight topics). After the training, the differences between Type B and Type A were pointed out and the analyst was notified that she would be working on 2 cases of each type. It was decided to do the training with a Type B stimulus because the types of evidence provided by a Type A stimulus was a subset of the evidence provided by a Type B stimulus. When Type C and Type D stimuli were used in a session, the training was done using a Type C stimulus (on another topic different than any of the eight topics and different than the previous training topic). After the training, the differences between Type C and Type D were pointed out by playing two segments from the Type D stimulus for the same case. It was decided not to do the training with a Type D stimulus so that analysts would not get accustomed to having the more predominant type of evidence (eye movements) after the training task and overly rely on it when they worked on experimental tasks while not paying enough attention to more subtle types of evidence (such as mouse movements).

In order to observe how the analyst's inferences about the searcher's interests progressed with increasing amounts of evidence, an incremental presentation of the recordings approach was used. A similar approach was used before in Janes (1991) to study how users' judgments of document representations changed as more information about a document was revealed to them.

During the study, the investigator sat next to the participant and conducted structured interviews after the analyst had finished viewing each screen shot or video segment, or whenever the analyst revealed some useful piece of evidence in the middle of a video segment. The following questions were asked at the interviews (video would be paused if the interview took place in the middle of a video segment):

- Did you learn anything about the searcher's interest? If so, what is it?
- How did you learn it? What evidence was it based on?
- What would you say about the searcher's interest now? Please summarize in a sentence. How confident are you with this inference on a 10 point scale, with 1 being least confident and 10 being most confident?

The screen contents and the interviews were recorded with Camtasia.

Analysts were allowed to reexamine a previously viewed screen shot or video segment at any time and were encouraged to pause and/or rewind the video at any time they felt necessary to reexamine some potentially useful details of the video and discuss it with the investigator. However, they were not allowed to use the fast forward function to skip any part of the video that was unseen so as not to miss potentially useful behaviors. At the end of a search case, an analyst was asked to summarize her inference and form a final statement of the searcher's information need based on her best judgment. She was also given the chance to discuss any evidence that she had found and compare the usefulness of different types of evidence. When the analyst finished all four search cases, an exit

interview was conducted in which she was asked to reflect on the entire experience and compare the effectiveness of the stimuli and the usefulness of the different types of evidence that she had used. Finally, she was debriefed and paid $20 for participation.

The second group of four search analysts examined search cases 9 to 12. Unlike the first eight participants, analysts in the second group only viewed Type D stimuli. Type D stimuli presumably contained the richest set of behaviors, so it would be interesting to collect more instances of examining this type of stimuli. The procedure for these four participants was similar to the first group. The order of the search cases was also counterbalanced. Training was done with a Type D stimulus on a search case other than the four experimental ones.

6.2 Results and analyses

Raw data collected in the study exist in the form of recordings of the search analysts' responses to the interview questions asked at critical events and at the end of each search case. Although think aloud protocols were applied, there was a strong contention between two highly competing tasks that analysts were asked to perform: to focus on the traces (many of which were subtle) of searcher activities in the videos and to verbalize their thoughts. As a result, the verbal protocols did not generate much useful data since most of the time analysts were merely verbalizing searchers' actions, instead of their interpretations of the actions.

After recordings were transcribed, classificatory content analysis (Allen, 1989) was performed on the data. In classificatory content analysis, a typology or classification of topics, ideas, or themes is established. Then, written texts or transcribed recordings of oral communication are assigned to one or more of the classes of the typology. In this study, classificatory content analysis was based on existing classifications of the behavioral sources of implicit feedback, as discussed in the literature review. The behaviors that were looked for include *examine* (time, number of revisits, pattern of eye movement and mouse movement, exit type), *scroll* (time, speed, amount, depth), *mouse-over a link*, *click* (i.e., select a link), *search within page*, and *query modification*. For *examine* and *scroll*, items in parentheses are attributes of the behaviors. They suggest the different ways that the behaviors have been used in previous studies. For example, *examine* may be a useful behavior not by a single instance of viewing, but because of repeated visits to the same page. In this case, the "number of revisits" attribute of the *examine* behavior is used.

This section starts with descriptions of characteristics of participants, especially their previous reference librarian experiences, and characteristics of the data collected, such as the number of inferences that each analyst made for each search case, analysts' confidence levels and the accuracy of their inferences. Then, analyses were conducted at three different levels to inform the research questions. Firstly, analysis was conducted on the search case level to shed light on this research question: does more evidence lead to better inferences?

The aim was to identify the critical points when the quality of the inference is significantly improved.

A second analysis was done by using type of behavior as the unit of analysis. This kind of analysis was used to inform the following research questions: Which type(s) of searcher behavior is useful evidence of the searcher's interests? Are the behaviors on the search results list page more useful than those on the result content pages? How is each behavior used? What are the rules used to make the inference? Different types of behaviors can also be compared on how likely they would lead to good inferences.

Thirdly, comparisons were made at the stimulus type level. Inferences based on 4 types of stimuli were compared in terms of their effectiveness. As mentioned in the study design, the potential difference between Types A and B reflects the value of monitoring the additional browsing paths beyond the search results list page as well as keeping track of the time spent on each page. The difference between Types B and C attests to the usefulness of capturing searchers' behavior within a page (such as scrolling and mouse movement), in addition to recording page level information, such as URLs. The difference between Types C and D will show the added value of eye tracking.

6.2.1 Characteristics of subjects

One of the participants (Participant 05) in the first group was only able to complete one study session, so another participant (Participant 13) had to be found to complete the other session that was originally assigned to that participant. Therefore, 13 participants

attended the study. Participants 2, 3, 5 (13), 7, 8, 9, 10 and 12 were in the first group; participants 1, 4, 6, and 11 were in the second group.

Among the 13 participants, 3 were male. All participants had at least 1 year of professional reference librarian experiences: four of them had less than 5 years' experiences, two between 5 and 10 years, and seven more than 10 years. At the time of the study, seven participants were working as reference librarians in academic libraries, two in public libraries, three worked as reference librarians before but were currently in other occupations (one as a faculty teaching and research librarian, two as doctoral students in information and library science), and one was a Master's student in library science, but concurrently working as a part-time reference librarian at an academic library.

6.2.2 Characteristics of data

The study consisted of 20 sessions and 4 search cases were analyzed in each session. Each of the first 8 cases were presented and analyzed 8 times, twice in each type of stimulus condition, while each of the remaining 4 cases were analyzed 4 times. Topic-wise, the 12 search cases were on 8 different topics: all 8 cases in the first groups were on different topics while the 4 cases in the second group were on the same topic as one of the cases in the first group. Therefore, out of the 8 topics, 4 were reviewed 12 times and 4 were reviewed 8 times.

6.2.2.1 Number of inferences

The unit of observation in this dataset is an inference that the analyst made and the accompanying rationales that she provided in response to the interview questions listed in Section 6.1.3. For the two types of screen shot stimuli (Type A and Type B), an inference was elicited after each screen shot was displayed and at the end of the entire search. For the two types of video stimuli (Type C and Type D), an inference was elicited after each video segment and at the end of the entire search, but an elicitation could also take place if the analyst spotted an interesting piece of evidence in the middle of a segment. In the latter case, the video was paused and the interview questions were asked.

Table 6.2 lists the number of inferences that were made by each analyst for each search case. These numbers not only counted the instances when the analyst made a new inference, or kept the same inference but updated the confidence level, but also included those when the analyst wanted to keep the same inference and same confidence level, but suggested new understandings of the search scenario based on evidence that were observed. For example, in some cases, the analysts observed evidence which did not immediately affect their inferences, but was confusing, or conflicted with what they were thinking, so they put a question mark in their heads and kept that in mind when they watched later parts of the search. Sometimes the analysts noticed behaviors which were different from their expectations, but were able to come up with reasons that would account for the behaviors, which made them keep the same inference and confidence level. For example, in one

instance (11-10D[1]), the searcher spent a very short time on a page that the analyst expected her to stay on longer, but in the meantime, the analyst noticed that this searcher had also spent a short time on other good pages that she had visited, so the analyst inferred that the searcher might just not like scrolling. Therefore, the analyst decided to keep the same inference and confidence level. In some other cases, the analysts noticed both evidence which would make them more confident and evidence which would make them less confident, so they decided to stay at the same confidence level before they saw more evidence. For example, in one instance (11-10D), the analyst noticed that the searcher spent a lot of time reading some text on an external result content page and the part read should answer the predicted question, and this observation would make her more confident about her inference. However, she also noticed that the searcher went back to the Google results list page at the end of the segment, instead of ending the search, which made her less confident. As a result, the analyst decided to stay at the same inference and same confidence level. Occasionally, an analyst said "this makes me more confident" or "this made me less confident", but did not want to change the confidence level. Sometimes this was because they were already at 1 or 10. Other times, they would say things like "this makes me more confident, but I have to see the next thing", presumably because the evidence was not strong enough.

---

[1] The notation 11-10D refers to Analyst 11 analyzing Search Case 10 using the Type D stimulus.

Table 6.2. Number of inferences made for each search case (shaded analysts were in second group)

Analyst / Case	1	2	3	4	5/ 13	6	7	8	9	10	11	12	Mean
1		3	5		5		3	6	5	5		5	**4.63**
2		8	4		5		6	5	5	4		4	**5.13**
3		6	5		5		6	7	5	5		6	**5.63**
4		9	14		11		8	13	13	7		7	**10.25**
5		6	5		7		5	4	8	7		5	**5.88**
6		12	13		11		9	11	7	7		7	**9.63**
7		3	5		5		5	3	5	4		8	**4.75**
8		5	7		4		5	5	7	6		4	**5.38**
9	5			8		6					8		**6.75**
10	14			13		12					14		**13.25**
11	6			6		8					5		**6.25**
12	7			7		8					6		**7**
Mean	**8**	**6.5**	**7.25**	**8.5**	**6.63**	**8.5**	**5.88**	**6.75**	**6.88**	**5.63**	**8.25**	**5.75**	**Grand Mean =6.787**

An analyst made an average of 6.8 inferences for each search case. The average numbers of inferences ranged from 4.6 for a search on Chinese restaurants to 13.2 for a search on President Kennedy's quote on the Apollo Project. There was apparently a much wider variation among search cases than among analysts. Statistical tests confirmed this observation: there was a statistically significant difference in the number of inferences made for each case [$F(11, 68)=15.579$, $p<0.001$)], but not by analyst [$F(11, 68)=0.724$, $p=0.711$)]. This result should not be surprising considering the information in Table 5.1: it is reasonable that analysts made more inferences for search cases in which the searcher had spent longer time searching, used more queries, and examined more results.

Inference changes were also analyzed. Results of this examination are summarized in Table 6.3. An inference update includes both the change in the content of the inference and/or in the confidence level. Results showed that out of the 6.8 inferences an analyst made for each case, 89.5% differed from previous ones in either contents and/or confidence levels. Among these inference updates, 69.3% were changes to the contents, while the rest 30.7% were changes to the confidence levels on the same inference.

Table 6.3. Number of inference update instances (shaded analysts were in second group)

Analyst / Case	1	2	3	4	5/ 13	6	7	8	9	10	11	12	Mean
1		3	4		5		2	6	5	4		5	**4.25**
2		6	4		5		4	4	5	4		4	**4.5**
3		6	5		4		6	7	4	5		6	**5.38**
4		6	10		11		8	12	13	7		7	**9.25**
5		6	5		7		4	4	8	4		5	**5.38**
6		9	11		6		7	10	7	7		7	**8**
7		3	5		4		5	3	3	3		8	**4.25**
8		5	7		4		5	5	7	5		4	**5.25**
9	5			6		6					7		**6**
10	9			12		10					12		**10.75**
11	6			5		8					5		**6**
12	7			7		7					4		**6.25**
Mean	**6.75**	**5.5**	**6.38**	**7.5**	**5.75**	**7.75**	**5.13**	**6.38**	**6.5**	**4.88**	**7**	**5.75**	**Grand mean =6.075**

Finally, looking at the contents of the inferences, a total of 319 unique inferences were made on 12 cases. Table 6.4 lists the number of inferences made on each case. This data reflects the variation among topics. Presumably, the more inferences made on a search case, the more confusing the searcher's behavior was since it allowed more room for

interpretation or the more complex the topic was. On average, 26.6 inferences were made for each search case. It is interesting to note that the number of inferences and inference updates and the number of unique inferences exhibit different patterns on some cases. For example, Search Case 3 received the second most unique inferences, although only the 8th most inferences and the 7th most inference changes, while Search Case 10, although receiving the most inferences and inference updates, was only associated with a mid-range (6th) number of unique inferences. This discrepancy means that for some cases like Search Case 10, analysts were more stable in the contents of their inferences, but went back and forth with the confidence level, or they made frequent switches between 2 or 3 inferences, while in other cases like Search Case 3, the analyst considered a number of different things that the searcher might be interested in.

Table 6.4. Number of unique inferences made on each search case

Case	Topic	Number of Unique Inferences
1	a. Chinese restaurant	24
2	b. Kennedy quote	23
3	g. ATC spur	39
4	c. Curling	48
5	d. Roy quote	30
6	e. Deaf comm.	37
7	h. Gas price	17
8	f. June bugs	25
9	f. June Bugs	11
10	b. Kennedy quote	24
11	d. ATC spur	20
12	c. Curling	21

6.2.2.2 Confidence levels

Analysts were asked to indicate their confidence level for each inference that they made using a 10 point semantic differential scale. Table 6.5 lists the average confidence levels each analyst indicated across all inferences they made for each search case. Note that the row average and column average were computed based on the numbers in the cells, which are averages themselves. It was decided to compute the "average of average" so that all cases or analysts contribute equally to the row/column means, instead of allowing cases with more inferences to have higher weights. Likewise, the grand mean reported in the table was computed by averaging the row (or column) means so that cases and analysts contribute equally, although some search cases were analyzed 8 times while others only 4 times, and some analysts analyzed 8 search cases while others only 4. If the grand mean was computed by averaging the confidence levels of all 490 inferences, regardless of which analyst were they from and for which search cases they were made, the value will change to 6.306.

Table 6.5. Average confidence levels (shaded analysts were in second group)

Analyst / Case	1	2	3	4	5/ 13	6	7	8	9	10	11	12	Mean
1		8.333	8.375		6.800		6.200	9.000	8.250	8.800		7.125	**7.860**
2		6.833	6.000		6.000		8.000	7.125	8.400	7.500		7.250	**7.139**
3		3.833	2.400		1.750		6.333	5.143	4.250	7.500		4.667	**4.485**
4		5.833	6.150		5.429		5.500	7.375	7.000	7.500		6.643	**6.429**
5		8.333	5.800		6.800		4.286	8.500	7.000	8.875		6.125	**6.965**
6		7.222	3.455		3.583		7.786	6.350	7.500	7.429		7.429	**6.344**
7		6.333	7.600		5.750		8.240	8.333	7.667	6.833		5.500	**7.032**
8		5.800	7.571		6.750		6.375	6.900	6.000	7.643		7.100	**6.767**
9	5.800			6.417		8.200					4.571		**6.247**
10	8.111			6.958		8.150					2.833		**6.513**
11	4.667			5.400		5.750					2.800		**4.654**
12	4.286			4.786		5.000					2.250		**4.080**
Mean	**5.716**	**6.565**	**5.919**	**5.890**	**5.358**	**6.775**	**6.590**	**7.341**	**7.008**	**7.760**	**3.114**	**6.480**	Grand mean =**6.210**

The ANOVA showed that there is a statistically significant difference among analysts in their confidence levels [$F(12, 67)=3.988$, $p<0.001$], and Scheffe's post-hoc test indicated that Analyst 8 and Analyst 10 assigned significantly higher confidence scores than Analyst 11 at the 0.05 level. Although the small number of participants in this study is insufficient to support a robust examination, an ANOVA was performed to assess the potential impact of an analyst's reference experiences on her confidence level. The 13 analysts were divided into 3 groups according to their years of experiences as reference librarians (less than 5 years, 5-10 years, and more than 10 years). The test suggested no statistically significant difference in confidence levels of participants from different groups [$F(2, 10)=0.733$, $p=0.505$]. Therefore, the difference seems to be due to individual

differences. Kelly, Harper and Landau (2007) discussed the limitations of using closed questions with Likert-type scales or semantic differentials in usability studies and pointed out that since scale measures reduce the respondent's options to one or more numbers and scale values are subject to individual interpretation, response sets provided for closed questions do not always capture the extent of a person's opinions. The same limitation applies to the capture of confidence levels in this study. Fortunately, the main interest here is to examine the relative change in confidence levels, rather than the absolute confidence level. When comparisons are made "within subject", the response bias should not be of concern as long as the same analyst had the same interpretation of the scales when analyzing different search cases. When comparisons are made "between subjects", standardized confidence levels (z-scores) were used in lieu of the raw levels that analysts assigned. The normalized confidence levels were computed as:

$$z = \frac{x - \mu}{\sigma},$$

where x is a raw confidence level to be standardized, μ is the average confidence level of the analyst of concern, and σ is the standard deviation of all the confidence levels that the analyst assigned. The quantity z thus represents the distance between the raw confidence level and the mean in units of the standard deviation and z is negative when the raw confidence level is below the mean, positive when above. As can be easily seen, the standardization does not affect any comparison within subject.

The ANOVA showed that there is also a statistically significant difference in analysts' confidence levels among different search cases [$F(11, 67)=4.368$, $p<0.001$], and Scheffe's post-hoc test indicated that Search Case 3 was associated with significantly lower confidence levels than Search Case 1 at the 0.05 level. The three search cases that analysts felt least confident with were Search Cases 3, 11 and 12. A common feature with these three cases was that the queries, even those after modifications, did not describe the search topics well. The final queries in these three search cases were: [railroad spur durham, nc american tobacco], [railroad spur and american tobacco company], and [ice sports tools], respectively. Unlike in most other search cases where later queries in the search sessions described the main concepts in the search questions reasonably well, it is quite unlikely that one could figure out the search questions or the scenarios from any of queries posted for these three cases. Instead, analysts had to rely more on behavioral evidence. That probably explains why the confidence levels were lower on these search cases.

6.2.2.3 Accuracy of inferences

In order to assess the accuracy of the inferences, all inferences were graded by human reviewers based on how close they were to the original search scenario. Three reviewers were involved in the grading. The investigator graded all the inferences. Two other reviewers each graded inferences on two topics. One of the reviewers was involved in pilot testing the interface for the second phase of the study, while the other reviewer participated in two sessions of the study as an analyst, so they were familiar with the

context of the inferences. Even so, the investigator explained the rules for grading in detail to the other two reviewers. They were reminded that analysts had been given the context of the search and been asked to infer the scenario that motivated the search. They were also reminded that analysts had been instructed to focus more on getting the search question correct than figuring out the context for the search question; therefore, for Topic f, for example, an inference saying "my dog ate June Bugs" should be ranked the same as "my neighbor noticed that his dog had eaten some June Bugs", if both inferences suggested the same search question "will this cause any problem to the dog". However, the reviewers were asked to pay special attention to the number of facets in the search question that each analyst correctly figured out in the inferences and rank the inferences accordingly. For example, for Topic a, "look for a good Chinese restaurant" should be ranked higher than "look for a Chinese restaurant" since the former one correctly suggested the quality facet.

Before the additional reviewers started working, the investigator met with them separately in person to explain the grading rules, and worked together on one topic for training purposes. The investigator worked with Reviewer A on Topic f and with Reviewer B on Topic h. So, grading for inferences on these two topics was the result of a consensus method. Then, Reviewer A worked independently on Topic c and Topic g; Reviewer B worked independently on Topic b and Topic e. The investigator worked on all 8 topics and compared the ranking with the two additional reviewers on the 4 topics that they graded. All three reviewers graded the inferences by weakly ordering them based on how close they

were to the original scenario (i.e., ties were allowed between inferences). Two methods were used to assess the extent of agreement between the investigator and the additional reviewers. First, the Spearman's Rho statistic was computed for the 4 pairs of ranked lists. The results were 0.848 (Topic b), 0.914 (Topic c), 0.637 (Topic e), and 0.742 (Topic g), all of which were statistically significant (all $p<0.001$). Secondly, Joachims et al. (2005, p. 156) described a method to compute the inter-judge agreement when ratings are in the form of two weakly ordered lists. The method counts the percentage of cases that the two judges agreed in the direction of preference whenever they expressed a strict preference between two items in the set. Using this method, the inter-reviewer agreements were 0.85 (Topic b), 0.93 (Topic c), 0.75 (Topic e), 0.85 (Topic g). Then, the investigator met with the other two reviewers again in person to discuss the inferences with large discrepancies in ranking and resolve the disagreements. Based on the adjusted ranking, an accuracy score was assigned to each inference so that the worst inference for each topic received a score of "1", the second worst inference received a "2", and so on. Tied inferences received the same score. Within each topic, the higher the accuracy score is for an inference, the closer it is to the original search question. As there were different numbers of inferences on each topic and different numbers of ties, the highest accuracy scores for the inferences ranged from 9 (Topic a) to 39 (Topic c); this suggests that accuracy scores can only be compared within topics.

Table 6.6 lists the mean ranks of each analyst's inference accuracy by search case. Analysts with the highest mean ranks for each search case are shaded. The table suggests that analysts performed differently on different topics and that no analyst performed well or poorly on all topics.

Table 6.6. Mean rank of inference accuracy (best analyst shaded for each search case)

Analyst \ Case	1	2	3	4	5	6	7	8	9	10	11	12
1									9.60	28.33	13.33	7.71
4									17.00	21.63	16.30	16.50
6									11.75	14.00	11.63	13.07
11									11.36	24.29	9.10	16.00
2	16.00	25.17	25.33	40.67	21.08	46.00	32.00	15.80				
3	10.75	11.00	12.20	37.80	20.50	6.68	20.00	24.57				
5		15.00	19.25			49.42	10.25					
7	22.00	23.50	19.67	38.06	21.50	43.57	15.80	23.50				
8	18.50	18.00	22.86	33.63	12.75	40.30	6.67	21.60				
9	20.20	9.90	32.00	34.54	21.25	25.64	28.00	17.43				
10	19.00	27.88	24.60	47.00	19.63	26.86	13.00	21.00				
12	20.40	17.25	21.17	39.43	21.60	31.50	17.00	18.88				
13	14.00			35.55	31.93			31.00				

6.2.3 Analysis on the inference level

This subsection presents the analysis on the inference level to examine the evolution of confidence levels and inference accuracy across time in each case. This aims to answer the research question: Does spending more time watching the search and exposure to more evidence lead to better and more confident inferences? It is most interesting to identify the critical instances when inference accuracy was significantly improved and confidence levels

were increased. Only those 490 instances in which the analyst changed their inference contents and/or confidence levels were included in this analysis.

6.2.3.1 Evolution of confidence levels across time

Figure 6.7 plots the evolution of confidence levels across time. The x-axis represents the normalized timestamps and the y-axis displays the average standardized confidence levels at corresponding time points. As different search cases vary significantly in duration (89 seconds to 753 seconds), the raw timestamps when inferences and accompanying confidence levels were elicited were normalized by dividing the raw timestamps by the duration of the corresponding search case. For example, all timestamps associated with inferences for Search Case 1 (which was 89 seconds long) were divided by 89. As the divisions led to fractions, all normalized timestamps (rounded to the third decimal place) were multiplied by 1000, so the final timestamps used in analysis were integers ranging from 1 to 1000. If the analyst updated her inference after seeing that the search had ended, the timestamp for that final inference was defined as the duration of the search plus 1 second, and in the normalized version, as 1001. For inferences made with the two screen shot types of stimuli, the time associated with an inference was defined as the time corresponding to the *end* of the search segment when the searcher was on that page. The 4 graphs display the change at different granularities: Figure 6.7(a) averages the confidence level every second; Figure 6.7(b) divides the 1000 standardized seconds into 100 segments, 10 standardized seconds each, and computes the average for each segment. Figures 6.7(c)

and 6.7(d) are similar to 6.7(b), but at larger scales. Overall, all four figures suggest that confidence levels increased as more evidence was shown.

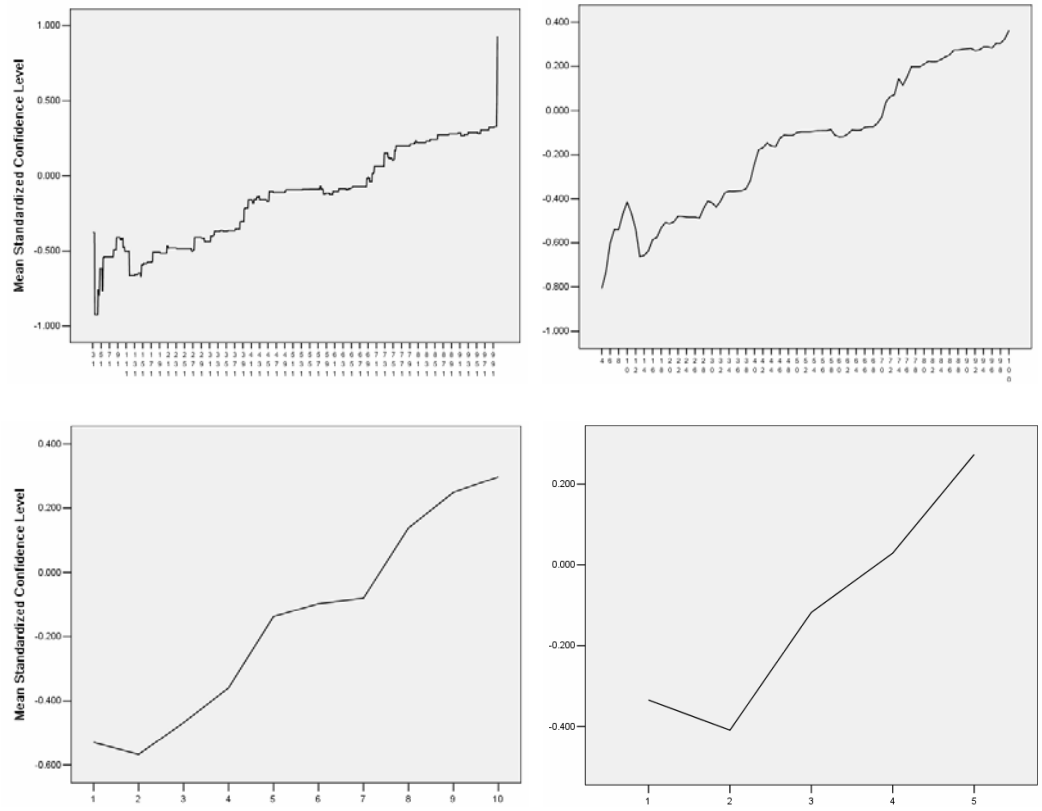


Figure 6.7. Evolution of confidence levels across time (upper left: a, every standardized second; upper right: b, every 10 standardized seconds; lower left: c, every 100 standardized seconds; lower right: d, every 200 standardized seconds)

In addition to the average, it is interesting to see how confidence levels evolved in each individual case. Figure 6.8 plots each analyst's change of confidence levels across time in each search case. Figure 6.9 presents a histogram that summarizes the magnitude of all changes in confidence levels (disregarding contents of the inferences). The magnitude of change was the difference between the raw confidence level of a later inference and that of

its predecessor. A positive change means that the analyst was more confident with the later inference (might be the same inference or a different inference) and that corresponds to an upward stair in Figure 6.8.

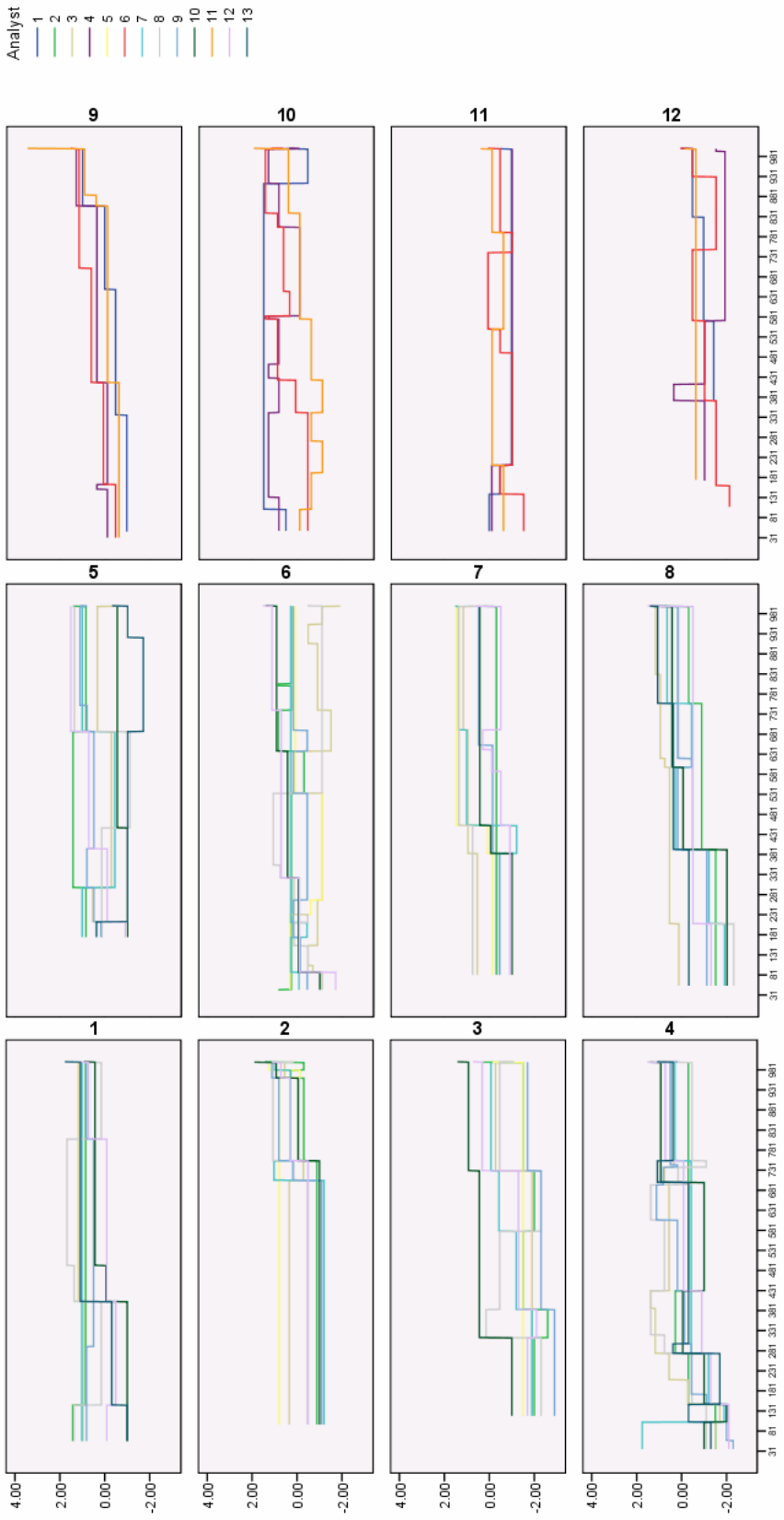


Figure 6.8. Evolution of confidence levels across time by search case

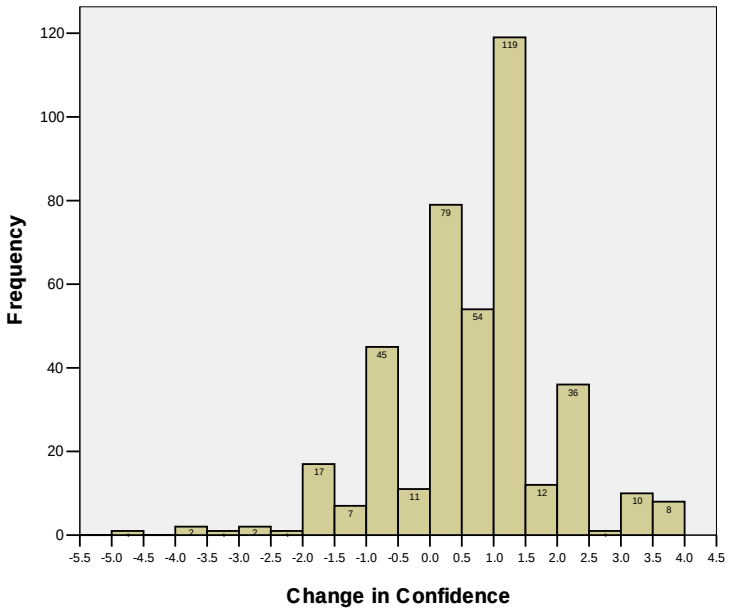


Figure 6.9. Magnitude of changes in confidence levels

Several messages can be learned from Figures 6.8 and 6.9 about the change of confidence levels. First, more evidence led to a mixture of more confident and less confident inferences. There are search cases (such as Search Case 8 and 9) for which more evidence led to steadily more confident inferences for most analysts, while for other cases, the confidence levels fluctuated for most analysts. The numbers of analysts who never decreased their confidence levels for each search case are summarized in Table 6.7. Shaded cases were analyzed 4 times while others were analyzed 8 times.

Table 6.7. Number of analysts whose confidence levels never went down

Case	# of analysts
1	4
2	3
3	4
4	1
5	2
6	2
7	3
8	6
9	3
10	0
11	0
12	2
Total	30

Second, some evidence was interpreted differently by different analysts so that they led to more confident inferences for some analysts and less confident ones for others. Many examples can be found in, for instance, Search Cases 5 and 11, as shown in Figure 6.8 (an upward stair meeting a downward step). In contrast, some evidence was interpreted more uniformly by different analysts and caused almost all analysts to increase their confidence levels. For example, at normalized timestamp 404 (raw timestamp 36) for Search Case 1 (Chinese restaurant), five out of the eight analysts increased their confidence levels (shown in Figure 6.10). That corresponds to the time when the word "best" was added to the original query [chinese restaurant chapel hill]. At normalized timestamp 730 (raw timestamp 116) for Search Case 3 (ATC spur), when "american tobacco" was added to [railroad spur durham, nc], seven out of the eight analysts increased their confidence levels

(shown in Figure 6.11). By seeing these additional query terms, most analysts felt that they were more confident about their inferences.

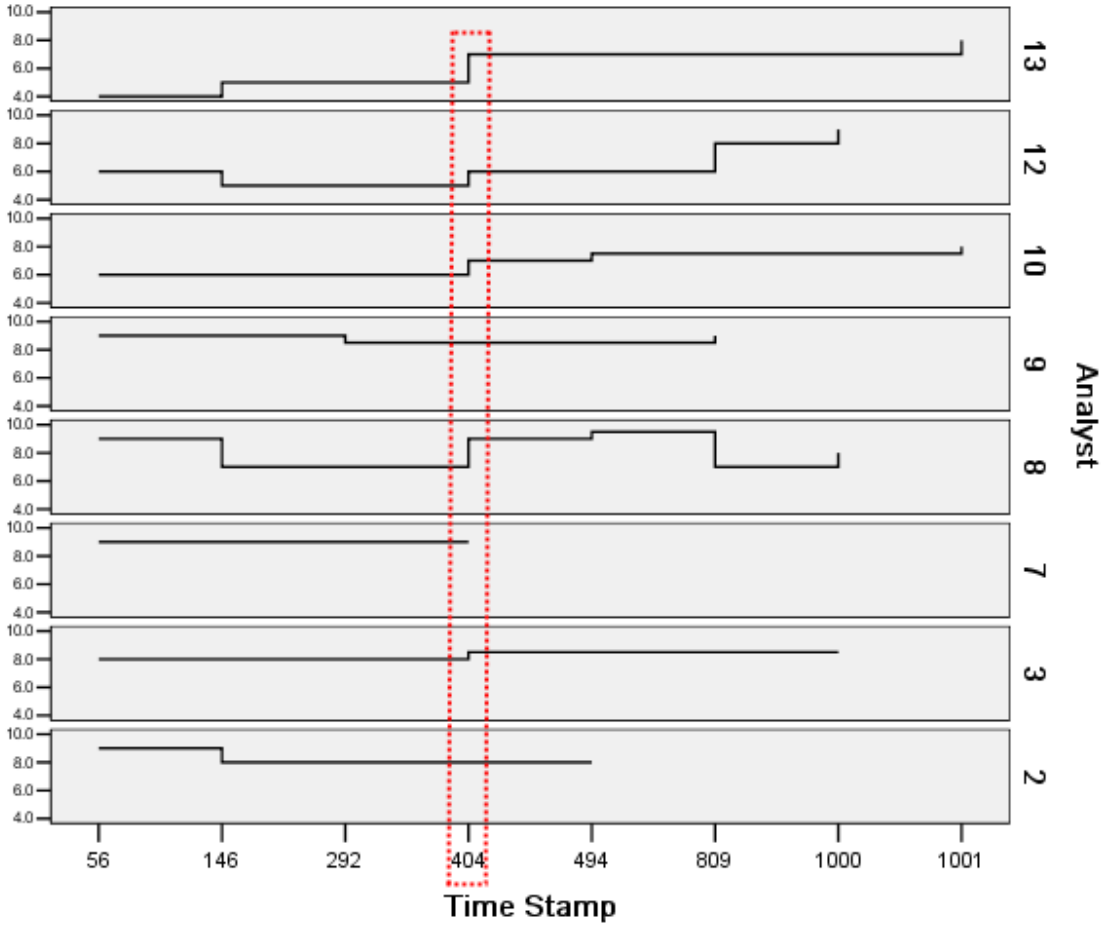


Figure 6.10. A sample instance with mostly positive confidence level changes in Search Case 1

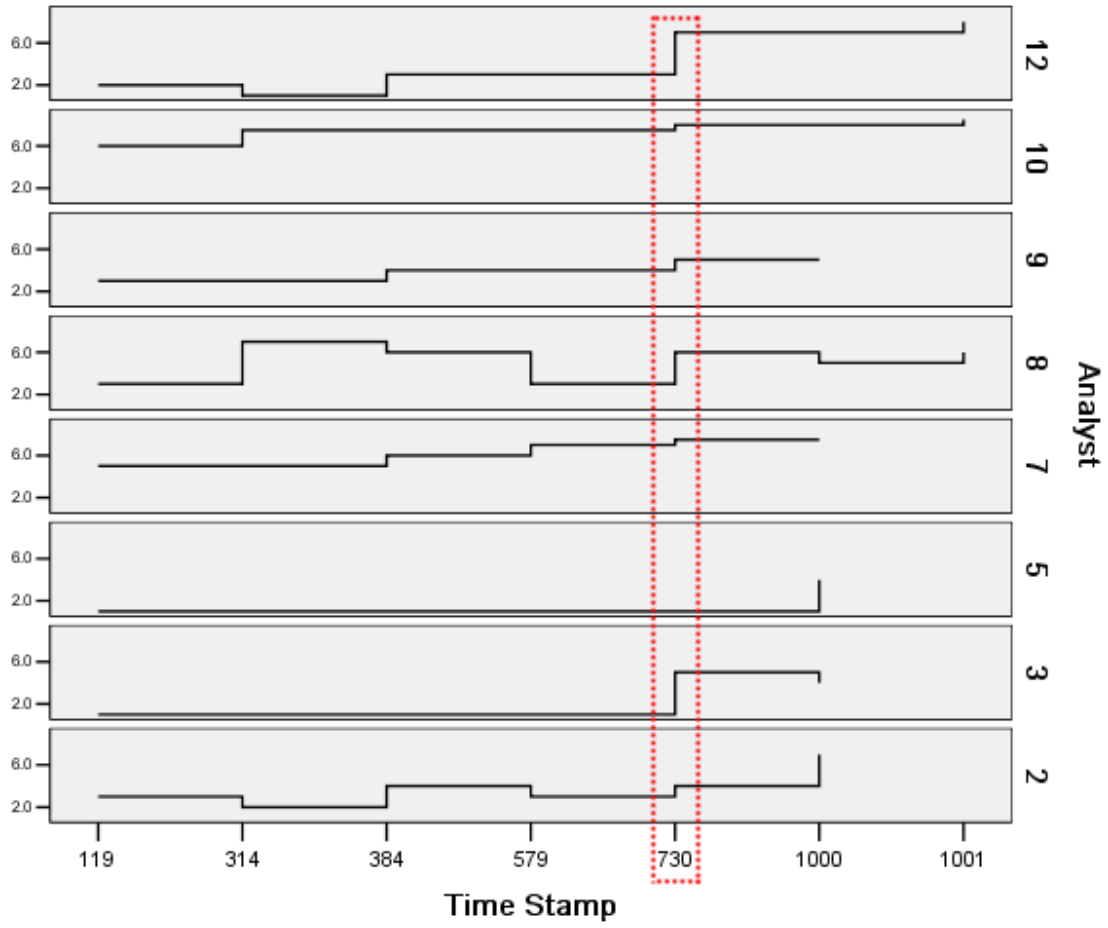


Figure 6.11. A sample instance with mostly positive confidence level changes in Search Case 3

Another way to look at the changes in confidence levels is to break them down by analyst. Similar to Table 6.7, Table 6.8 lists the numbers of search cases for each analyst in which she never decreased her confidence levels. Shaded analysts analyzed 4 cases while others analyzed 8. Also, in no case did an analyst constantly become less confident each time she updated the inferences and/or confidence levels.

Table 6.8. Number of search cases without decrease in confidence levels

Analyst	# of search cases
1	2
2	2
3	2
4	0
5	1
6	1
7	2
8	1
9	3
10	7
11	2
12	5
13	2
Total	30

Notably, Analyst 10 contributed 7 of these cases. That is to say, out of the 8 search cases that she analyzed, there were 7 in which she constantly became more confident with her inferences as more evidence were presented or stayed at the same level of confidence. Also, Analyst 12 did so 5 out of the 8 times. Both of them have more than 20 years of experience as reference librarians. Based on this information and that from Table 6.5, it seems that more experienced analysts did not necessarily assign higher confidence levels on average, but they were more likely to make sense out of new evidence while less experienced analysts were more likely to get confused or challenged. Again, this observation is based on a very small sample without considering the content and the accuracy of the inferences, so it should not be generalized.

6.2.3.2 Evolution of inference accuracy across time

Figure 6.12 plots the evolution of inference accuracy across time. The x-axis represents the normalized timestamp and the y-axis displays the accuracy score of the inference made by each analyst (represented by colors) for each search case (in different panels) at corresponding time points.

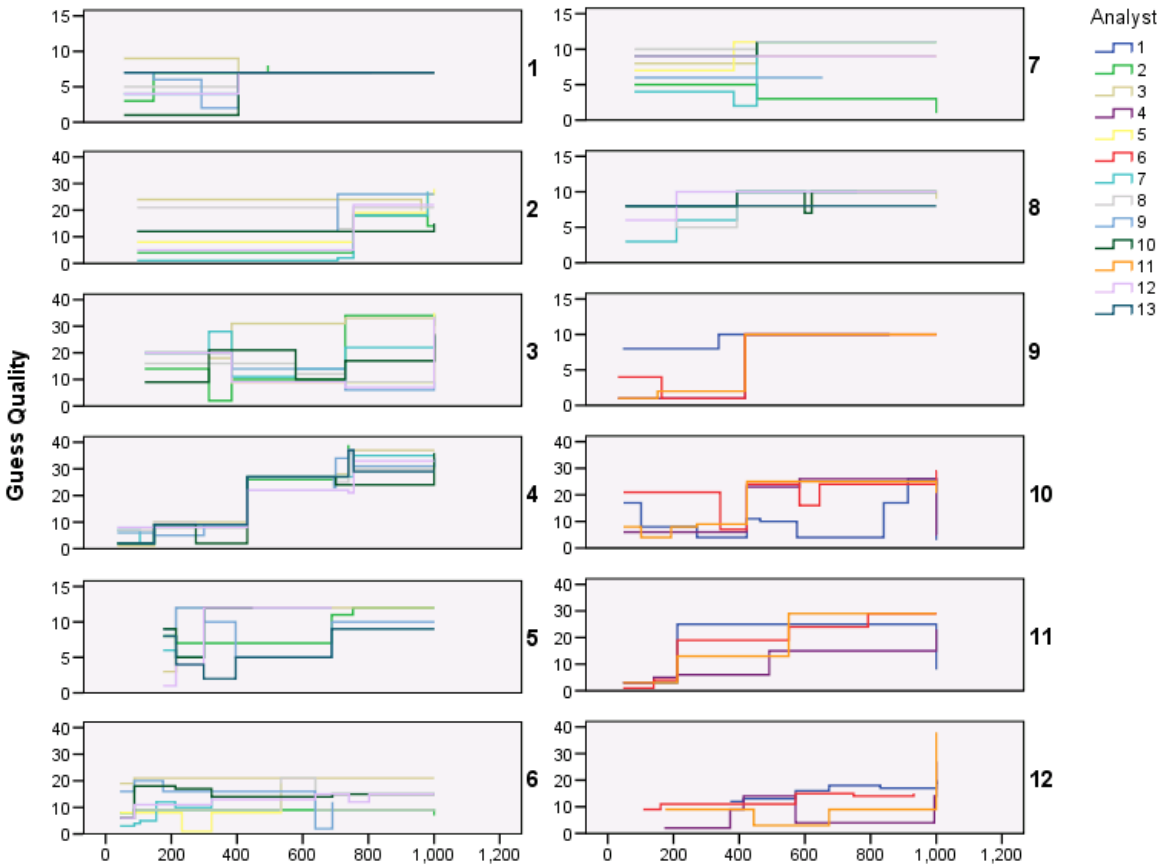


Figure 6.12. Evolution of inference accuracy across time by search case

Like confidence levels, more evidence led to a mixture of better and worse inferences. For some search cases (such as Search Cases 2, 8, 9 and 11), more evidence led

to better inferences; while for other cases, the evolution of inference accuracy was less uniform. Table 6.9 lists the number of analysts whose inference accuracy stayed the same or improved with more evidence available. Note that these numbers do not correlate with those in Table 6.7 about change of confidence levels.

Table 6.9. Number of analysts whose inference accuracy never went down

Case	# of analysts
1	6
2	5
3	0
4	2
5	3
6	1
7	6
8	5
9	3
10	0
11	3
12	0
Total	34

Also, some evidence led to both better and worse inferences; but probably more interestingly, there were a few evidence instances which resulted in jumps in accuracy scores for almost all analysts. The most striking one was in Search Case 4. At normalized timestamp 431 (raw timestamp 157) when the query changed from [type “ice sport”] to [ice sport with brush], the additional qualifier “with brush” boosted the accuracy of the inferences tremendously for all 8 analysts. Similarly, in Search Case 11, when the query changed from [american tobacco company] to [american tobacco company and the railroad]

at normalized timestamp 212 (raw timestamp 50), all 4 analysts were able to make better inferences. Clearly, both “with brush” and “and railroad” in the above cases are highly discriminating terms that represent crucial new facets and thus add a lot more value than modifying an existing facet.

Table 6.10 lists the numbers of search cases for each analyst in which her inference accuracy never went down. Shaded analysts analyzed 4 cases while others analyzed 8. In no case did an analyst consistently make equally good or better inferences each time she updated the inference and/or confidence levels. Again, these numbers do not correlate with those in Table 6.8 on change in confidence levels.

Table 6.10. Number of search cases without decrease in inference accuracy

Analyst	# of search cases
1	1
2	3
3	4
4	2
5	2
6	1
7	3
8	3
9	3
10	3
11	2
12	5
13	2
Total	34

6.2.3.3 Overall relationship between time and inference

Similar to the concepts of "calibration" and "resolution" in the confidence judgment literature (Liberman & Tversky, 1993; O'Keefe, 2000), accuracy of inferences and analysts' confidence in the inferences often counteracted each other. As Liberman and Tversky (1993) pointed out, a prediction can be perfectly calibrated without being informative. In this study, analysts also mentioned that they sometimes became less confident in their inferences when they *attempted* to be more accurate by making a more specific inference. Table 6.11 summarizes the directions of changes in inference accuracy and confidence level. It should be noted that this table only includes instances where either the content of the inference (not necessarily the accuracy score because of ties) or the confidence level had changed. Therefore, the "same-same" cell refers to instances where the analyst changed her inference and was as confident in the new inference as in the previous inference, while the new inference, although different, was as accurate as the previous one.

Table 6.11. Number of change instances broken down by directions of change in accuracy and confidence level

		Accuracy of Inference			
		Up	**Same**	**Down**	**Total**
Confidence Level	**Up**	84	130	28	242
	Same	43	16	18	77
	Down	27	41	19	87
	Total	149	188	69	406

The three cells in the upper left region mark the most positive changes, in which the analyst either made a more accurate inference and increased confidence, became more confident in the same or an equally good inference, or maintained the same confidence level on the same or an equally good inference. These instances counted for 63.3% of the total (67.2% if adding the 16 "same-same" instances). This means, about one third of the time, exposure to more evidence did not translate to better inferences. It can be argued that some of the 41 instances where analysts became less confident in the same (or a different but equally accurate) inference might be interpreted as positive changes if the inferences were inaccurate, because the fact that an analyst challenged herself about a bad inference was presumably a more positive action than the one where the analyst somehow became more confident about an inference which was inaccurate. A further analysis on this issue will be left for future work because it requires a more accurate measure of the inference accuracy at the interval or ratio level than the ordinal level measure used here.

6.2.4 Analysis on the evidence level

This subsection presents analysis of the behavioral evidence that was used to support the inferences. Through in-depth examination of the rationales behind the inferences, it seeks to understand the types of searcher behavior that were considered as useful evidence of searchers' interests and how they were used to inform the inferences. Note that the unit of analysis so far has been on the inference level since both confidence levels and accuracy scores are associated with inferences, but in this subsection, the unit of analysis shifts to a

finer level: the behavioral evidence level. Sometimes, the analyst may use multiple types of evidence that she has observed to support an inference; sometimes, an inference was not supported by any behavior from the searcher, but based on the analyst's background knowledge. Therefore, there was not a 1:1 correspondence between the inference and the evidence that supports the inference. There were cases in which analysts mentioned some behavioral evidence that they observed from a screen shot or a search segment, but did not update their inferences or confidence levels. These cases were included in this analysis, making the total number of instances 550, larger than that in 6.2.3 (which was 490).

The investigator performed a content analysis on analysts' responses to the question "How did you learn it? What evidence(s) was it based on?" First, 10% of the data was randomly selected to be analyzed using the baseline model of searcher behavior discussed in Chapter 4, which considers the following types of behaviors: search, select, examine (view, scroll, eye movement, mouse movement, mouse click, search within page, exit type, total amount of page activity), retain (print, bookmark, email, copy/paste). By analyzing this sample of data, the investigator revised the baseline model to reflect some preliminary observations. Firstly, the analysts heavily relied on comparing what they observed with what they had predicted. They often assumed the role of a searcher and tried to predict which link should be clicked based on the surrogate or whether an external result content page should be helpful to the searcher based on features of the page (content, structure, or layout). Then, they would compare the behavior of the actual searcher against what they

would do themselves if they had been searching. If the observation matched the prediction, they normally would keep the inference and increase their confidence level; otherwise, they would either change the inference if enough new information had been learned, or stick to the same inference but lower the confidence level. Thus, two elements were added to the coding schema for the "click" behavior. One is the "goodness" of the selection (i.e., the relevance of the selected page) as perceived by the analyst assuming her inference was correct. The other is page feature that the analyst considered in association with the click, such as types of surrogates (e.g., title, summary) on a Google results list page, or the content, layout, and structure of an external result page. Neither of these two items was based on the searcher's behavior, but they were closely associated with the "click" behavior. Secondly, the preliminary observations also revealed more nuances of evidence that analysts used in making inferences. For example, when considering results selection, analysts not only considered which one had been selected, but also ones which had been skipped, or the fact that no selection had been made on the page. This suggests attributes that should be coded for the "click" behavior, such as "select", "skip", and "lack of select". Sometimes, a combination of attributes was considered within the same instance. For example, some analysts used the difference between the selected result and the skipped ones as evidence of the searcher's interest. In these instances, both "select" and "skip" should be coded. With more data being analyzed, the coding schema was also further developed to cover new types of behaviors and new approaches to using the behavioral evidence.

6.2.4.1 Types of evidence

Table 6.12 lists the types of behavioral evidence considered by the analysts and the perspectives from which they were used. In the rest of this section, each type of evidence will be examined in detail with examples of usage.

Table 6.12. Types of behavioral evidence considered by the analysts

Category	Behavior	Perspectives taken by analysts
Search	Submit query	enter new query, add terms, remove terms, put terms back, modify query (difference between new query and old query), linguistic features of query terms, natural language query
Select	Select	result level – title of selected result, summary of selected result, URL of selected result, relevance of selected result based on surrogate page level – select next page of results list
	Skip	result level – title of skipped result, summary of skipped result, relevance of skipped result based on surrogate page level – lack of select (skip all results on page)
Examine	View	time spent on page, relevance of selected result based on page content, page structure (text? list?)
	Mouse movement	terms that were hovered over
	Eye movement	on all pages – eye movement speed (reading vs. scanning), fixation position, places where searcher spent a long time (focus), place that was focused on repeatedly, lack of focus on the page, exit position (where searcher looked at last) on results list page – relationship between click and scanning
	Scroll	scrolled, lack of scroll, scroll speed, scroll depth, scroll fixation position, number of repeated scrolling
	Search within page	search terms
	Exit page	exit type (Back, END)
Search session	Behaviors w.r.t. other behaviors in the search session	continued searching instead of ending stage in session past behavior in the same search session

6.2.4.1.1 Search

The first type of behaviors, "search", encompasses the explicit activities that searchers performed on the query, including issuing new queries and modifying existing queries (by adding terms, removing terms, changing to new set of terms, and putting back terms which were previously removed), as well as certain features of the queries that searchers used, such as the linguistic features of certain query terms, or the fact that some queries were in natural language.

Issuing the first query marked the beginning of a search session. As easily conceivable, seeing the first query always helped analysts to make their initial inferences. It should be mentioned that some human analysts were able to pick up subtle ***linguistic features*** from the queries (most often from the initial queries) and use them to inform the inferences. For example, three of the eight analysts working on Search Case 8 commented on the use of the word "canine" in the first query [canine eating bugs]. Their comments included "the word 'canine' was confusing" (02-08D), "the use of 'canine' suggests that this may be a medical request because 'canine' is a technical term" (03-08D), and "they used 'canine' maybe because they wanted authoritative information" (08-08B). Other comments on linguistic features were made on:

- "toxic" – "it's a very medical term" (08-08B)
- "ice sports" – "plural form, meaning they were looking for more than one sport" (10-04C)

- "population" – "more like grouping or statistics" (10-06A); "the use of the word 'population' is interesting; they didn't use 'people'" (09-06D)
- "US" – "they added the 'US' although the query already has 'American Sign Language'" (10-06A)
- "American Sign Language" – "they used '**American** Sign Language', not just 'sign language'" (09-06D)
- [+"american sign language" +deaf +population +speak] – "they used '**speak** American Sign Language'" (03-06B)

Some of these observations on linguistic features were correct and gave the analysts useful information about the search question, such as the use of "canine" (leading to medical request and authoritative information) and "population" (leading to statistical information), but some were simply faulty, such as inferring multiple sports from the phrase "ice sports", and understanding the word "speak" as collocating with American Sign Language. Another interesting use of linguistic features of queries is the attention to natural language queries, such as [description of ice sports] and [gas prices in european countries], which contained prepositions or connector words. All together, linguistic features were used 10 times and the existence of natural language queries was mentioned 4 times.

A much more prevalent use of queries was to look at query modifications. Analysts were good at comparing later queries with earlier queries in the session and when doing so

they paid special attention to which terms were added, which were removed, and which were removed and then put back.

Adding query terms was mentioned 120 times. The added terms represented the concepts that the searchers realized were missing from previous queries and search results. Therefore, in most cases, seeing these concepts added to the queries helped the analysts disambiguate what the searcher was looking for. For example, when seeing the query changing from [+"american sign language" +deaf +population +speak] to [**+us** +deaf +population +"american sign language" +speak], Analyst 3 commented that "It makes me more confident that they were trying to get a statistic because usually when you are looking at statistics, you talk about a particular country or region. So they included 'US' because they expected to find the word 'US' in the result" (03-06B). Some of the added terms could make even more critical contributions to the inference because they represented a whole new concept which was not covered in previous queries. Examples included the adding of "author" to ["be led by your dreams" quote] in Search Case 5 and the adding of "was where" in the last query in Search Case 10. Such instances were generally mentioned by multiple analysts and they corresponded well with searchers' self-reflection of critical instances in the first phase of the study.

Added terms did not always provide confirmatory information. Instead, they could challenge previous inferences sometimes. For example, when the query changed from [apollo project john kennedy speech] to [john kennedy quote about apollo project], Analyst

4 felt that her inference was challenged, because she had been thinking that the searcher had been searching on the full text of Kennedy's speech (04-10D).

It is also interesting to notice that the same added term can be interpreted differently by different analysts. For example, when the query changed from [railroad durham, nc tobacco] to [railroad **spur** durham, nc tobacco], the added term "spur" helped one analyst to figure out that the searcher was "not interested in the entire line" (10-03B). However, it also confused some analysts, because "I didn't understand why the searcher added the word 'spur'" (09-03A, 12-03B). Another example of the analyst being confused by new terms is "they added 'with brush', but I'm not getting more confident because I don't understand why they didn't use 'brush' at the beginning" (13-04A).

Sometimes, the introduction of new terms could be related to behaviors before or after the change. For example, with regard to the query change from [+us +deaf +population +statistic +"american sign language" +speak] to [+us +"american sign language" +speak **+english**], Analyst 10 commented that "I didn't do it [adding "English" to her inference] right away until I looked more of what they looked at …why they bothered to put in 'English' has more importance than I was thinking at the first given their behavior and what they looked at" (10-06A). In another instance, Analyst 11 commented that "when they added 'toxicity dogs', previous clicks now make sense" (11-09D).

A special case of adding "terms" to queries was adding quotation marks. This was reflected in some of the rationales analysts mentioned on the two search topics involving

quotes (Topic b and Topic d). For example, one analyst said "they put the quotation marks in, so they are really looking for that phrase together" (03-05A).

Finally, another perspective that quite a few analysts took when they saw new terms in queries was to examine the source of the added terms, which turned out to be very informative. Analysts made the distinction between terms coming from the searchers (i.e., terms that were given in the search question) versus terms that searchers picked up from search results. An example of the former case was reflected in this comment after the analyst saw the word "spur" being added to the query: "I think 'spur' was part of your original question. I don't see any other way how they would have come upon that. I didn't see it in any of the search results until now" (06-11D). An example of the latter case was the comment made after the query change from [ice sport with brush] to [use of brush in curling]: "they added 'curling', which means they found what they wanted from the search" (13-04A). Interestingly, a similar rationale was used when the searcher did not add words from results to the query. For example, Analyst 13 once noticed that the searcher of Search Case 4 (on curling brush) did not carry over terms from the results of previous queries, and commented that, "if the searcher was still searching for (ice sports) facilities, they should take some words from the pervious search when they made this new query, such as 'arena', but they didn't. So, I don't think they were looking for facilities" (13-04A).

Removing query terms was mentioned as evidence of searchers' interests 14 times. The fact that some terms were removed from the query challenged analysts to think why the searcher would have done that. Here are three examples:

> "Because I'm trying to think why the user thought that tobacco was really important at the beginning of their search, but it didn't turn out to answer the question, and so they took it out. Why would you do that? I think they took it out because they realized that wasn't part of the answer, and so I was trying to think of a question that would make it conditional: tobacco, yes or no. Well, they thought yes at first, and then as they did some research, they thought no." (08-03A)
>
> "… took out 'tobacco', maybe because the searcher thinks tobacco is no longer part of the picture, so it probably won't appear in the answer to the question; therefore, the question has something to do with the use of railroad today" (10-03B).
>
> "sometimes you search and you add something in, then you get from the results, and find that, OK, I knew that, and take it back out and change it … because it's like I already knew THAT part of the story … this makes me think that Roy Williams is given while the question asks for something else about the quote" (08-05C).

The three analysts used similar reasoning to explain why some terms were removed from queries: imagine a term X has been removed from the query, then this probably suggests that the question is about whether X is involved in some kind of relationship (e.g., whether railroad was or is used to ship tobacco; whether Roy Williams was the original

source of the quote). The analyst in the third case was correct in this observation and that was indeed the reason why the searcher had taken the terms "Roy Williams" out of the query. However, the removal of "tobacco" was a bad search strategy (the searcher put it back in the next query) and misleading evidence, so the inference based on this action was incorrect.

In some cases, a searcher would put a word or phrase ***back into the query*** after taking it out. This was deemed to be a strong indication of the searcher's interest on that concept. For example, three analysts commented on the fact the searcher in Search Case 6 put the word "statistics" back after taking it out and they thought this was an even stronger indication that the searcher was interested in finding statistics than when they first saw this word in the query.

Finally, if a query was ***modified*** (not just adding or removing terms), the analyst would still concentrate on the difference between the two queries. Specific comments on query modification (instead of just mentioning the fact that the query was modified) seemed to be made more on occasions when the new query was similar to the old one. Some comments include "they changed from 'safety' to 'toxic', but it's in the same line" (13-08A), "the searcher experimented with another query term 'harmful', which was similar" (11-09D), and "the fact that the person did not adjust the search substantially … it seems that what they got the first time was OK" (12-06A). As the last example suggests, this type of query modification has been largely viewed as reinforcing behaviors.

6.2.4.1.2 Select

Select

The fact that the searcher ***click****ed* a certain link was mentioned as evidence of their interests 131 times. Among them, the great majority referred to the selection of results on the Google results list page, while a few were about link clicking on external result content pages, which did not happen often in the collected searches.

There were two ways in which link selection helped the analysts. One of them was straightforward: the analyst learned of the searchers' interests based on some content features of the selected pages, mostly the summary and sometimes the title of the page in the Google results lists. This often resulted in an update to the content of the inference. The second way involved comparing the selection against the analyst's prediction. If the selected page matched the analyst's prediction (i.e., a good selection), the analyst became more confident in the inference; otherwise, a bad selection would challenge their inferences, which resulted in either the change in inference content and/or the drop of confidence levels.

There were 18 instances where analysts commented that the searcher made ***good selection****s*. The judgment was mostly based on the fact that the selected page would be helpful to answer the search question that the analysts were predicting at that moment (which was not necessarily correct though). Their comments went like "they clicked on this one, which would potentially be the answer to that" (02-07C), "they are going back to

places where they should be" (11-10D), and "they clicked on 'curling' which would reinforce my guess" (13-04A). In all instances but three, the analysts raised their confidence levels because of the match between the prediction and the observed behavior. As for the three instances without increase in confidence levels, one was the first inference in the session so that no comparison could be made (02-07C); another was due to an additional observation that the searcher was still in the middle of search session[1] so that the analyst felt there could still be changes to the search strategy (07-04A). There was only one instance where the analyst did not feel more confident and did not provide a reason.

It has also been observed that analysts sometimes considered a selection good not only based on the content of the selected page or the Google result snippet (title or summary) that informed the selection, but also based on features such as the URL of the selected page. For example, when the searcher in Search Case 2 (Kennedy quote) clicked on a NASA page, one analyst commented that "the searcher went to NASA showing that they were looking for more credible source … when I see answers to the question found on a more credible source, that made me more confident" (10-02A). Other features that have been considered include whether the page contained a list of items or contained mostly text, an important distinction when analysts were pondering about whether the searcher was looking for a quote or a speech which contained the quote. It is interesting to notice that linguistic

[1] The use of "position in search session" as evidence will be discussed later.

features of terms again played a role in defining the inference sometimes. For example, an analyst made a comment about the word "museum" in the title of a selected result: "the title of the selected page has the word 'museum', so this must be a historical question" (12-03B). This additional information was correct and helpful.

There were 10 instances where analysts made comments on ***bad clicks*** and all but two were associated with drops in confidence levels. Some comments were: "that made me confused, because as a librarian, I would have clicked on other pages" (02-06B), "Based on the blurb here, I have no idea why they chose this particular one. It's pretty far down on the list of results. It doesn't seem to provide any additional information … This makes me doubt my scenario … the fact they chose this, because this little blurb here, in my mind, gives me no reason to choose it." (03-05A), and "they seemed to be looking for other sign languages, which confused me" (03-06B). In the other two instances, the analysts admitted that they were challenged, but they did not lower their confidence levels immediately because in one case, the analyst felt that "it's just a bad choice at the beginning" (08-04B), and in the other case, the analyst would like to see more evidence before making the change:

> "They are clicking on pages which conflict with their query. At this point, without, you know, seeing a couple of more pages, to know they seem to like the ones that are less about statistics, and more about learning sign language, I can't really tell ...

they queried one thing, but they were attracted to another thing. Where do they go back and redo the search ... can mean one thing or the other." (02-06B)

Skip

There were 42 instances where the analysts considered the ***skip of certain results*** as useful evidence of their interests. In some instances, this was useful by ruling out certain possibilities. For example, when the searcher in Search Case 5 (Roy Williams quote) skipped the first result and clicked on the second one, the analyst commented that "they skipped the first result, which was about a book, so they were not looking for books titled 'be led by your dreams'" (12-05D). In Search Case 12 (curling brush), the skip of links on ice climbing and mountaineering also helped some analysts to rule out these sports. In Search Case 3 (American Tobacco Spur), the skipping of several results on the American Tobacco Trail helped analysts to figure out that "they don't seem to be interested in the walking trail" (02-03C).

The skipping behavior was also used by making the distinction between skipping pages that the analysts would have also skipped based on the predicted question and those they would have clicked. Other than what the searchers skipped, another way to consider the skip of links was ***the lack of any click*** on a page, mostly on the Google results list page (i.e., no result was selected before going to the next page of results or modifying the query). More than 20 such instances were considered by the analysts, most of which resulted in changes in confidence levels. For example, the observations below led to increases in

confidence levels because the searchers' behavior (lack of click) matched the analysts' expectations:

> "None of the results on this page were what they were looking for and I would agree based on my guess about what they were doing" (03-08D)
>
> "… and because they didn't click any of these that didn't have to do with dogs; all these results were not about dogs … that leads me to believe that they are still interested in dogs eating June Bugs and whether June Bugs are toxic to dogs." (03-08D)
>
> "They didn't click any of the results from previous query, which were all irrelevant; then they went to modify query." (13-04A)

On the other hand, the observations below caused the analysts to lower their confidence because what they observed challenged what they thought about the searcher:

> "… lack of click makes me less confident, because lack of click means the searcher didn't like something about their query and I didn't know what it is" (11-10D).
>
> "… they exited without clicking … didn't click on the results that I expected them to click if my guess was correct". (02-06B)
>
> "They didn't click on results with numbers [statistics]." (03-06B)
>
> "If what I was thinking was the query, I would have thought they would have clicked on a couple of the others, but they didn't and I don't know why … That to

me says there was something more specific that they were looking for, but I don't know what it is and I would want them to put into the search query." (05-06C)

Finally, there were 12 instances where the analysts explicitly mentioned that they had considered both the results that had been selected and those that had been skipped. In these instances, they focused on the differences between selected results and skipped once, which is well reflected in the comments below:

"There was the word 'speech' in the summary of the clicked result, but not in the summaries of the two skipped results." (01-10D)

"They chose the one about the railroad museum, but not the one about tobacco trail. So, they seem to be interested in the railroad itself, rather than any later thing." (02-03C)

"They selected the first one which didn't have specific restaurant and skipped others which mentioned the names of specific restaurants." (10-01D)

Noticeably, all 12 mentions of both selection and skipping were made on Google results list pages except for one made on an external result content page. That page contained a list of different ice sports and the analyst noted that "they picked 'other', but skipped 'ice hockey' and 'figure skating'" (06-12D) which made her more confident that the searcher was looking for sports other than ice hockey or figure skating.

6.2.4.1.3 Examine

This category encompasses all searchers' behaviors that examined a page in the search session, including Google results list pages and external result content pages. The sources of evidence related to examination that have been considered include time, eye movement, mouse movement, scrolling, searching within page, and exit type.

Time

Time is a source of evidence that has been widely studied in previous research. The way time was considered by the analysts in this study was mainly to make the distinction between "long" time and "short" time and compare that to the features of the page and its predicted usefulness.

First, it will be interesting to look at how much time has been considered to be "long" and how much time has been considered to be "short". The 27 mentions of "short" time ranged from 2 to 36 seconds and the median was 7.5 seconds, while the 34 mentions of "long" time ranged from 5 to 189 seconds and the median was 57 seconds. Sample comments on the short time spent on a page included, "they didn't spend much time on the page" (05-06C), "they were back right out" (04-10D), "within 2 seconds, they realized it's a movie site" (06-10D), and "they only spent 3 seconds … the layout tells that there is no full text" (01-10D). Comments on the long time spent on a page included, "they spent some time here looking through it" (03-07C), "he spent a whole lot of time on it" (07-07D), and "he is interested because he is spending time here" (07-06C).

Just like the way analysts considered link selection and skipping, they also made the distinction between spending time on a page that should help and spending time on a page that did not contain useful information to answer the predicted question. For example, Analyst 3 noticed that the searcher spent 70 seconds on a page that contained information relevant to the predicted question and thus increased her confidence level from 3.5 to 4.5 (03-06B). A similar comment was, "they spent some time reading it ... I think that's because it's helpful and they are supposed to find out about this" (08-04B). In comparison, another comment said, "that knocks my confidence down a little bit to a 7 just because they spent so much time here and it seems totally unrelated to my scenario".

Additionally, features of the page (structure, layout and so on) have also been frequently used to interpret the time. For example, consider the comments below:

> "I'm gonna guess this didn't really help them because they had spent 12 seconds on this page, which wasn't a long time, while the statistic is over here, it's not substantiated in anyway, so I'm gonna guess this didn't give him what they needed." (03-06B)

> "It's possible for the searcher to notice within 6 seconds that the page contained the full text of the speech, because the heading says 'TEXT OF PRESIDENT JOHN KENNEDY'S RICE STADIUM MOON SPEECH'." (02-02B)

In the first case, the layout of the page (text, without tables, figures or listings) helped the analyst to determine that 12 seconds was not long enough for the searcher to find

what she was presumably looking for (numbers or statistics). In the second case, although the searcher spent less time (6 seconds), the presence of a heading in large font made the analyst believe that it was possible for the searcher to capture that visual cue quickly.

Mouse movement

Mouse movement was only available in the two types of video stimuli (C and D). It was mentioned as a useful evidence only once by Analyst 9 when she was examining the Type C stimulus for Search Case 1 (Chinese restaurant). She noticed that the searcher moused-over the address of the restaurant quickly, which made her believe that the searcher was caring about the location of the restaurant and updated her inference to be "My family is going to dine out tonight and we're going to Chinese food in Chapel Hill. We're looking for a restaurant within walking distance." Although mouse movements were available in Type D stimuli, analysts reported that the eye movements were more straightforward and salient so that they did not normally pay attention to mouse movements.

Eye movement

Eye movement was only available in Type D stimuli. Out of the 550 instances examined in this section, 243 were associated with Type D stimuli. Analysts mentioned eye movements as useful evidence of searchers' interests 58 times, among one third of which referred to eye movements on Google results list pages and two thirds referred to external result content pages.

The most common way of using eye movements was simply to consider where the searcher looked while they were on the page, such as "he looked at the footnote to see where the information came from" (07-03D), "they kept their eyes on equipment names" (06-12D), "they looked at numbers in the summaries" (08-06D) and "they stopped a lot on climbing and mountaineering" (11-12D). In these cases, ***positions of the searchers' eye fixations*** gave analysts information about what they were interested in. In one instance, the analyst felt that where the searcher last looked on the page before they ended the search was indicative of her interest (05-03D).

The fixation evidence has also been used in different ways. On Google results list pages, analysts often compared the results that searchers had looked at and selected with those they had looked at but skipped. For example, consider these three comments: "they looked at part of the quote in a result summary and selected that result … they looked at but rejected some other results (such as IMDB) which were bad results" (04-10D), "they looked at but skipped links on American Tobacco Trail" (05-03D) and "they were not looking for snow sports, because they didn't spend much time on snow sports" (06-12D). In these cases, the availability of the eye movements enabled analysts to form better pictures of the searchers' interests by knowing not only where they clicked, but also which other results they considered or which aspects of the result surrogates the user paid most attention to (e.g. the title of the page, or the summary).

On external results content pages, analysts frequently considered which part of a page the searcher spent most time focusing on, which is a more accurate description of the searcher's behavior than how much time she spent viewing the page. In other words, knowing how the time has been spent is more informative than just knowing how much time has been spent. Below are some comments which reflected the use of eye movements from this perspective. In each of those instances, analysts clearly got some information which would not have been possible without knowing the eye movements.

> "He spent most of the time looking at what's the explanation, rather than the actual prices." (07-07D)
>
> "He didn't spend time on historical prices, but went down to the current prices." (07-07D)
>
> "They spent a lot of time on the quote, so they were interested in the speech itself, instead of information about the speech." (04-10D)
>
> "They spent a lot of time on the tools for curling." (04-12D)
>
> "The entire time they were on the curling entry, they were on the brush." (04-12D)
>
> "The user spent some time looking at 'Rice University'." (09-02D)

One analyst even commented on the usefulness of eye movement evidence directly. She first noted that "it's clear that they did see it and stopped there for a minute" and then went on to comment that "if you don't have the eye movement, I don't think I'll have a 9 [for confidence level]" (12-01D).

In addition to positions of eye fixations, analysts also paid attention to and learned from different ***patterns of eye movements***. Below are some examples of how patterns of eye movements informed the analysts about what the searchers were looking for. Again, all of them testified to the usefulness of eye tracking.

> "They scanned the entire page and seemed like they were looking for something specific about the June bug." (11-09D)
>
> "They spent enough time to scan the whole thing that didn't answer the question, instead of looking for a table or numbers, which made me less confident" (08-06D)
>
> "They looked at the location, but kept looking for a while. So obviously, it's not just that." (04-11D)

Most noticeably, analysts identified some eye movement patterns which seemed quite generalizable. First, there was a pattern about result examination and result selection:

> "Well, when they were looking over the [Google results list] page, they didn't just look at the first one and modify the search or go somewhere else; they looked carefully around all of them [results]. People usually do that if the results are fairly close to what they want. If the first one, which is supposedly the best ... if that was not good, they will probably go somewhere else, modify the search or give up." (12-05D)
>
> "They spent a long time reading the summary before clicking, which means what they read [i.e., summary] should be relevant." (11-11D)

Analyst 11 noted further that clicking and keeping reading while the page was being loaded means the searcher was more confident about the result than if the searcher moused-over a result, kept reading, and then clicked it.

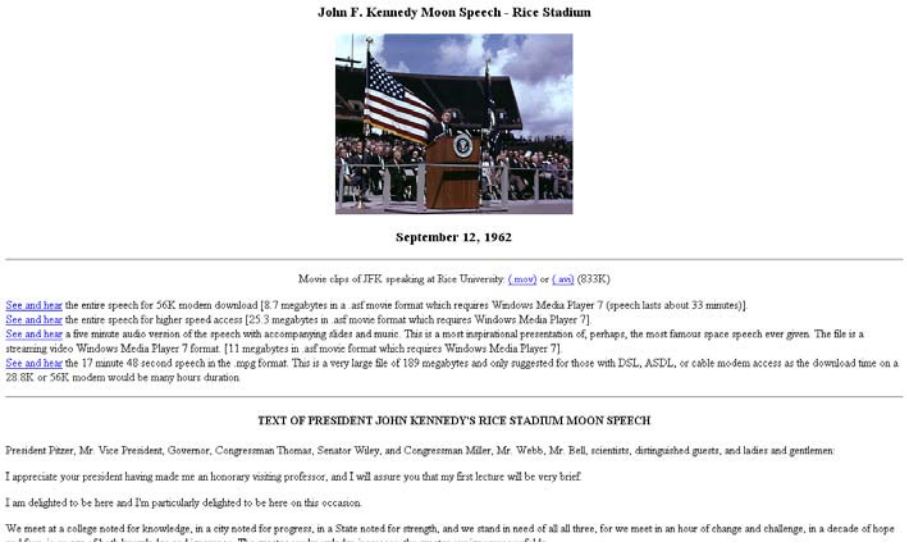
Second, among the patterns of eye movements, several analysts found ***repeated eye fixations on the same place*** to be the most reliable indication of searcher interest, as exemplified by the comment below:

> "They spent a lot of time reading and going back to the 'sweep' and 'broom' sections … and the fact that they kept coming back to the word 'sweep' and the word 'broom' three or four times … so I think these words mean something to them". (03-04D)

As another example, Analyst 11 successfully inferred that the searcher was looking for a quote based on the fact that she had read a quote on the page repeatedly. As will be discussed below, the repetitiveness of visits was also noted from scrolling behaviors when analysts used stimuli of Type C and regarded as highly indicative of their interests.

Overall, eye movement was able to give analysts information about searchers' interests at a more subtle level than click streams and viewing time through revealing which parts of the page searchers considered and how their viewing time was spent. On pages without clicks, knowing where the searcher looked was even more important. A good example was the last search segment for Search Case 2 (Kennedy quote) in which the searcher only spent 6 seconds on the page, quickly picking up the location of the speech

from the first line of the page, which says “John F. Kennedy Moon Speech – Rice Stadium”. Before seeing this page, Analyst 2 using Type B stimuli made the inference “In President Kennedy’s speech about sending a man to the moon, he used the word “easy”. Can you find the text of the entire speech?” (02-02B). After seeing the page and knowing the searcher only spent 6 seconds on the page, she thought it was possible for the searcher to notice within 6 seconds that the page contained the full text of the speech, based on the heading in all capital letters shown at the bottom of Figure 6.13. So, she used this as positive evidence and became more confident in her inference. In comparison, Analyst 9 who used Type D stimuli noticed that “the user spent some time looking at ‘Rice University’”, which helped her correctly figure out the location part of the search question. In fact, several analysts who used other types of stimuli after using Type D stimuli explicitly expressed that they missed the eye movement.



John F. Kennedy Moon Speech - Rice Stadium

September 12, 1962

Movie clips of JFK speaking at Rice University: (.mov) or (.avi) (833K)

See and hear the entire speech for 56K modem download [8.7 megabytes in a .asf movie format which requires Windows Media Player 7 (speech lasts about 33 minutes)]
See and hear the entire speech for higher speed access [25.3 megabytes in .asf movie format which requires Windows Media Player 7]
See and hear a five minute audio version of the speech with accompanying slides and music. This is a most inspirational presentation of, perhaps, the most famous space speech ever given. The file is a streaming video Windows Media Player 7 format. [11 megabytes in .asf movie format which requires Windows Media Player 7]
See and hear the 17 minute 48 second speech in the .mpg format. This is a very large file of 189 megabytes and only suggested for those with DSL, ASDL, or cable modem access as the download time on a 28.8K or 56K modem would be many hours duration.

TEXT OF PRESIDENT JOHN KENNEDY'S RICE STADIUM MOON SPEECH

President Pitzer, Mr. Vice President, Governor, Congressman Thomas, Senator Wiley, and Congressman Miller, Mr. Webb, Mr. Bell, scientists, distinguished guests, and ladies and gentlemen:

I appreciate your president having made me an honorary visiting professor, and I will assure you that my first lecture will be very brief.

I am delighted to be here and I'm particularly delighted to be here on this occasion.

We meet at a college noted for knowledge, in a city noted for progress, in a State noted for strength, and we stand in need of all all three, for we meet in an hour of change and challenge, in a decade of hope

Figure 6.13. Screen shot of the top of a page from Search Case 2

Scrolling

Scrolling was cited as useful evidence of searchers' interests in 20 instances and it was used in a variety of different ways. First, analysts considered the ***presence or absence of scrolling***, and most often, they considered it together with the content or features of the page where the scrolling took place, to determine if the searcher was interested in the page. Here are three examples in which the presence or absence of scrolling together with the layout of the page helped the analyst to interpret the searcher's behavior and infer her interest:

> "They were going through the article, and they scrolled; they didn't back out like they did with the quotes. So, the format [paragraphs of text] was what they were looking for, rather than bullets, or a bunch of quotes. So it seems to me they are looking for the speech that has the quote in it. … The fact that they scrolled through the whole thing tells me that this is the kind of format that they were looking for." (04-10D)
>
> "Based on the layout of the page, the searcher quickly realized that it did not contain the text of the speech; so he did not read the text carefully, but quickly went back to the results list. This reinforces the inference that the searcher was looking for the text of the speech." (01-10D)
>
> "I don't know why they don't scroll down. If they were looking for quotes, they would have scrolled down [on the page with a list of quotes]." (04-10D)

Here is another example in which the presence of the scrolling behavior, the content of the page, and how the searcher exited the page were used to support the inference:

> "The searcher scrolled down to look at the page and clicked 'back' to return to the Google results list. The page did not contain the text of the speech, so the searcher might have been looking for the specific text." (01-10D)

As these examples demonstrate, the presence and absence of scrolling is tied to the searcher's interest on the page. The presence of scrolling suggests that some features of the page shown before the fold made the searcher believe that the page had the type of content and genre that she was looking for. Further evidence, such as viewing time, and scrolling speed should be considered to determine what most likely interested the searcher. The lack of scrolling can either mean that the searcher noticed something clearly wrong about the page from the top fold (such as the wrong format, or a 404 error), or that she found what she wanted without having to scroll down (as in the Kennedy quote example given at the end of the discussions on eye movement). Again, further evidence such as viewing time and eye movement would be needed to get a more accurate interpretation.

When the searcher scrolled, analysts gained further insights from more specific patterns of scrolling such as the speed, depth, repetitiveness and place of focus. Almost all of these types of evidence were used in Type C instances. As can be noticed below, analysts often made attempts to infer searchers' eye movements based on patterns of scrolling.

Speed of scrolling is an important factor that analysts considered. Comments were often made when searchers scrolled down on a page quickly without focused reading, which was interpreted either as a sign of lack of attention or an indication that the searcher was looking for specific information such as numbers or names which easily stand out. Here are some examples:

> "They scrolled down really quickly as if they weren't reading, unless they were reading to recognize names … so maybe they were looking for some specific one and they didn't find. This makes me less confident." (08-01C)
>
> "Looks like they were scanning and looking for something that they think will jump out at them and that's gonna be a word, phrase, or number, but they didn't see what they wanted to see; otherwise, I would expect them to linger longer on the page and read more carefully. They were scrolling too quickly to read carefully." (05-06C)
>
> "They were scrolling, but didn't spend some time focusing on something. *They didn't even spend enough time to actually see where they mentioned the spur.* They sort of looked here … they spent some time, but not really enough time to read it thoroughly." (02-03C)

The italicized part in the last comment represented a very important observation that several analysts made: whether the searcher spent time around the part of the page where the keywords in the summary appear may be a good indication of whether the searcher found the page helpful. A rational searcher makes a clicking decision based on the surrogate

that Google provides for the result. Based on this assumption, places on the results page which contain the keywords in the summary would most likely be where the searcher should focus. If the searcher scrolled quickly and passed by those places without reading carefully, that probably means that the searcher was not satisfied with the page and did not benefit from viewing the page. Even worse, if those places were lower on the page but the searcher did not scroll far enough to reach there before going back, analysts could be even more certain that the searcher did not like the page, as suggested in this comment:

> "They didn't spend much time on the page. They didn't even scroll down to see where that quote was that was in that teaser. So this apparently right off was not what they wanted." (05-06C)

This suggests another approach to examining the scrolling behavior: the ***depth of scrolling***. Below are two more examples highlighting how depth of scrolling has been used to inform searchers' interests:

> "They only scrolled down to the middle of the list [instead of finishing the entire list]. This confirmed that it's a general question." (08-01C)

> "Given the amount of time they spent here, it looks similar to what they want. It either doesn't have the city or country they want, but they didn't scroll down far enough to see that. So, I'm more tempted to say it probably doesn't cover the time range they want." (05-07D) [This helped the analyst to understand why the searcher did not like the page and thus what she was looking for.]

Next, when searchers scrolled to some part of the page and then spent a significant amount of time resting there and doing some focused reading, the content of the focused area was considered by analysts in a somewhat similar way as they considered the positions of eye fixations. Therefore, the fact that searchers scrolled to a certain place on the Web page and spent time reading it is referred to here as ***scroll fixations*** and it has been regarded by the analysts as an indication of the searcher's interest in the content around that area. Mentions of scroll fixations include:

> "They spent most time on the 'method of play' section." (10-04C)
>
> "They went up and they went down, and they settled down on this [paragraph], which would potentially be the answer to it." (02-03C)
>
> "They scrolled down to recommendation [of Chinese restaurants], but didn't spend time reading other things like address or hours." (08-01C)
>
> "He spent a lot of time on it [the page], which confirmed my guess … he didn't spend time on historical [gas] prices, but went down to the current prices." (07-07D)

Like eye movements, analysts have also considered the ***repetitiveness of scrolling*** to a certain part of the page as a strong indication of interest, although without knowing searchers' eye movements, occurrences of such scrolling patterns could confuse the analysts sometimes, as evidenced by the example below:

> "[the searcher had] a weird scrolling behavior ... scrolling up and down for three times on a table that I expected them to focus on the bottom." (03-07C)

In this case, the searcher was scrolling through a lengthy table of weekly gas prices in 6 European countries over the past 10 years. When she scrolled to the end of the table to look at the current prices, she forgot the country names listed as the table headings. So she scrolled back and forth to check country names and current gas prices. Both analysts using Type D stimuli were easily able to understand the behavior because they could see that the searchers looked at the heading of the table which contained country names.

Search within page

Although searchers in the study were generally less experienced with Web searching, a few of them were able to use the search facilities within the browser (Ctrl-F). When the search did happen, most analysts noticed it, but few used it to infer the searcher's interest. However, there is one interesting example which shows a successful use of search within page as an evidence of the searcher's interest. Before seeing the segment which contained the search within page instance, the analyst made the inference that the searcher was "looking for something easy to understand about Kennedy's project of going to the moon" (07-02C). Then, on the next page, the searcher searched for the word "easy", which made the analyst feel that "I'm not sure if he's still looking for something easy to understand" and dropped that part of the inference. This change was due to the realization that search terms used in within page searches were expected to literally appear on the page, so they must be descriptors of the search topics, rather than qualifiers.

Exit Type

Exit type refers to the way in which the searcher exited a page. In general, there are seven types of actions that can be performed to exit a page during Web searches: issuing a new query, typing in a new URL, returning to the previous page (using the "back" button), clicking on a link (including selecting a result, navigating further from an external result page, and navigating to subsequent pages of Google results list), clicking the "home" button, clicking an item in the favorite list, and closing the browser window. As searchers in this study conducted the searches in a laboratory environment, none of them used the bookmarking function or customized the home page setting in the browser (which had been set to the Google home page prior to the study). Neither did they need to type in new URLs because they had been asked to use Google for all the searches. Searchers had been instructed to click the "home" button when they finished searching on a topic or when they wanted to return to the Google home page and start a new search strategy. They had been further instructed not to close the browser window due to the requirement of the recording software[1]. Therefore, the exit actions that could have been observed in this specific search setting included issuing a new query, returning to the previous page (clicking the "back" button), clicking a link (on the Google results list page or on external result content page),

[1] Searchers were allowed to close additional browser windows if they opened new windows during the searches, but they should not close the last window. However, none of the 12 search cases selected to be analyzed in the second phase involved opening of additional browser windows.

going to the next page of the Google results list, and clicking the "home" button to end the search. Among these actions, issuing new queries and selecting results have been discussed in earlier sections on "search" and "selection", so the focus here will be on the other three types of exit behaviors: returning to the previous page, clicking "home" to end the search and clicking "next page" to navigate to subsequent pages of the Google results list.

Analysts commented on ***returns to the previous page*** in 20 instances, most of which involving the searchers clicking the "back" button to return to the Google results list. This type of exit action has been mostly interpreted as a signal that the external result content page did not contain what the searcher wanted to find, as evidenced by comments like "they didn't find what they were looking for, so they clicked 'back'" (01-09D), "clicking 'back' suggests that the page didn't contain answer to the question" (11-10D), and "[the searcher] clicked 'back', which made me less confident … maybe he was looking for something other than the quote" (11-10D).

Then, depending on whether the analyst thought the content of the page that the searcher returned from contained what the searcher was looking for, the returning behavior could either reinforce or challenge the inference. For example, after seeing that the searcher returned to the Google results list after visiting a page that contained an answer to the predicted question, the analyst commented that, "I was a little bit less sure because they went back to this page … why do they need to come back here?" (09-05C).

There were 2 instances in which analysts mentioned that the searcher went to ***the second page of the Google results list*** and they took the same perspective in interpreting this behavior. They noted that the searcher chose not to change the query, but view more results and thought that it was because the results on the first page of the list were close to what the searcher was looking for.

Ending the search was mentioned as useful evidence of the searcher's interest in 29 instances, among which analysts became more confident in 27 instances. The increases in confidence levels may be due to the fact that most of the search cases selected to be analyzed ended with the searcher successfully finding the information, so by the time the searches ended analysts had been able to make a reasonable inference about the searcher's interest. A typical comment said: "I'm more confident knowing that the search ended here and it seems that the location sealed the deal" (06-10D). Analysts could also become more confident because they realized that they had watched the entire search so that the searcher's behavior would not change any more. For instance, here is a comment reflecting that notion: "I'm more confident because this is the end of search, knowing that I already have all the information about the search" (08-03A).

In one instance, the analyst kept the same inference and the same confidence level, but commented on the ending behavior. She said:

> "they either found what they wanted or they gave up, but I think they found what they wanted, [because] they read it too many times to give up there … I just think

it's the repetitiveness of the reading … if they were going to give up, they would have given up earlier on this Wiki page, or after reading what a broom does once, but they read it several times." (11-12D)

In the other instance, the confidence level dropped by 2 when the analyst knew that the search had ended. Her comment was:

"… he wouldn't have listened to the speech if he were looking for the location, so he couldn't have been looking for where … also, he didn't put in when/year/date, or where/place in the query, so, he was still looking for text. [On this page], he didn't look at the text, but clicked home. I'm confused." (08-02D)

In this case, the analyst did not realize that the question consisted of two parts, one on the quote and one on the location of the speech, so she ruled out the possibility of searching for the location of the speech based on the fact that the searcher listened to an audio recording of the speech (presumably to judge if the quote was contained in it).

6.2.4.1.4 Search session

In addition to considering individual screen shots and search segments, analysts also considered the evolution of the search session as a whole through comparing the searcher's later behavior with her behavior earlier in the session. In 10 instances, analysts noted that the searcher was still ***continuing the search*** in the same track, but they made different interpretations. In 4 instances, they did not change their confidence levels because they were not sure if the searcher would change the search strategy in the rest of the search. In

the other instances, the analysts started to question their inferences because the behavior contradicted their prediction. For example, here is a comment from Analyst 9: "my confidence dropped because they kept searching and I don't know why" (09-05C). In another instance, the analyst even modified the inference after noticing that the searcher did not exit the search. She explained that "the fact that they were still looking makes me believe they were looking for something specific, not just the name [of the sport]" (08-04B). Another analyst also felt that the searcher must be looking for something different when she noted that the searcher was continuing the search after viewing a page which could have answered the predicted question. She said "nothing on the page helped … [I'm] just trying to think of some question that has not been answered by the previous page" (12-04C).

Analysts made many interesting observations when they ***related searchers' behaviors across the session***. For example, an analyst inferred that the searcher must be looking for something general because "he clicked on results that have little in common" (11-12D). They sometimes used previous behavior to explain a later one. For example, in one instance (11-10D), the searcher spent a very short time on a page that the analyst expected her to stay on longer and did not scroll down the page, but in the meantime, the analyst noticed that this searcher had also spent a short time on other pages that she had visited which contained useful contents, so the analyst realized that the searcher might just not like scrolling. Sometimes, a later behavior can also help the analysts to better understand an earlier behavior, as evidenced by this comment: "I didn't do it [adding

"English" to her inference] right away until I looked more of what they looked at …why they bothered to put in 'English' has more importance than I was thinking at the first given their behavior and what they looked at" (10-06A).

Finally, some analysts were even able to take advantage of some negligence in stimulus design and use ***position in session*** as evidence to support the inference. Since some search cases involved so many screen shots and search segments that a scroll bar became necessary to display links to them in the left "table of contents" frame, some analysts used the position of the scroll bar to judge if the searcher was close to the end of the search. For example, an analyst noted that the searcher had selected a good page, so she should have become more confident in the inference, but "since he is still in the middle of the session, I want to stay at the same confidence level" (07-04A). Had this problem been noticed before the study, work could have been done to make sure that all left frames contain a scroll bar and the position of the scroll bar would not reveal the position of the screen shot or search segment in the search session.

6.2.4.2 Combination of evidence

In addition to considering the types of evidence that were used and the perspectives that were taken to use the evidence, it should also be noted that analysts have frequently used multiple types of evidence to support an inference. This subsection serves to highlight this phenomenon through aggregating previous discussions and providing more examples.

Page features (such as content, snippet, and layout) were often considered when analysts used time, link selection or the lack of it as evidence of the searcher's interest. For example, selection was compared against the analyst's prediction: a good selection made the analyst become more confident in the inference while a bad selection would challenge the inference, leading to either a change in inference content and/or a drop of the confidence level. Similar logics were applied to the interpretation of time: spending a long time on a good page would confirm the inference, as spending a short time on a bad page would; on the contrary, spending a long time on a bad page or spending a short time on a bad page would challenge the inference.

On Google results list pages, eye movements and result selection were used together to determine which results searchers considered but skipped. These results were then compared with the selected one to infer the searcher's interest. On external results content pages, when eye movements were not available, scrolling (especially speed and depth) was used in combination with the time to tell which part of the page the searcher read more carefully; the presence or absence of scrolling was also used in combination with the content and layout of the page and exit type to infer the searcher's interest.

In the exit interview, some analysts gave thoughtful suggestions on how certain types of behaviors could be used in combination to better interpret search behavior based on behavioral patterns that they observed, but may have felt premature to mention during the study. For example, Analyst 12 commented on how query modification can be used

together with search result page flipping to infer how closely a query represented what the searcher wanted:

> "Going to second page [of the search results list] means the results are close to what they wanted. If the searcher does not go to the second page, it may mean that he is satisfied, or the search is totally off – that can be further judged by if searcher puts in a new search and how different the new query was from the previous one." (Analyst 12)

In addition, it has been observed that the analyst's background knowledge of the search topic and the search context (i.e., the searchers, the study setup) were frequently referred to as factors (although not behavioral evidence) that affected the inference. For example, several analysts questioned their inferences and decided not to assign a higher confidence score because "I don't think you'll make the question so simple" (09-07A). Other more direct use of background knowledge varied from considering linguistic features of certain query terms to drawing upon the type of questions that one has received from patrons in reference interviews to predict specific aspects of the searcher's interest. For example, some analysts automatically thought about the author of the quote simply because they had received lots of such questions at work. Some analysts inferred that the searcher was looking for something broad when she clicked on a Wikipedia page. Such considerations did not necessarily lead to more accurate inferences, but reflected the unique perspectives of human analysts which would be for very hard for machines to simulate.

6.2.4.3 Analysts' perceptions of evidence usefulness

In the exit interview, analysts were also given the opportunity to discuss and compare the usefulness of different types of evidence that they considered. When they were asked which type of evidence had been most helpful, their responses focused on queries, eye movements, time and result selection.

When analysts mentioned queries, most of them emphasized that it was most helpful to see the process of query modification. Their comments included: "it was helpful to see the query development" (Analyst 2), "what really helps is the tweaking of their terms" (Analyst 9), "often it's just the way they modified the search: select, reject, and come back to modify the query" (Analyst 4), "seeing what they added and what they subtracted helps you to narrow the search and get a closer idea for what they were looking for" (Analyst 12) and a longer comment from Analyst 7:

> "the process of the search ... again, it's like a reference interview, you go back and forth ... it's a combination of the query, the behavior of picking the site, what the site tells them, how they modify the query based on viewing each site ..."

In general, the emphasis on the process, rather than individual queries, was highly consistent among analysts.

Analysts had mixed attitudes towards eye movements. Some of them mentioned that "knowing where they looked at is most helpful" (Analyst 10), "seeing what they read is

really helpful" (Analyst 12) and "knowing what they were doing is more helpful than just knowing how long they spent on the page" (Analyst 8). A more detailed comment said:

> "the eye movement thing helped so that I can see if they were actually reading something ... they'll take the time to read it ... and then see if they will go down more, or they were ready to go back if they were looking up at the top of the browser and looking for the back button" (Analyst 9)

However, several analysts commented that they did not benefit much from eye movements because "I'm not a very visual person" (Analyst 2, Analyst 3). Analyst 2 even felt that she had been distracted by the gaze path sometimes and concentrated less on the search process.

Analysts had general consensus on the difficulty of interpreting the usefulness of time. They agreed that knowing how long searchers spent on the page was helpful, but they also pointed out that the fact that the searcher spent more time on a page did not necessarily mean that the page was better. They also agreed that the distinction between "long time" and "short time" was very fuzzy and varied greatly across situations. Some of their comments on time are listed below.

> "The time helped because you could get more of a sense of how useful they found a page by how long they stayed on it ... short time could be very useful or not useful at all, but at least it's giving you that much ... longer time means that they were interested enough to explore further." (Analyst 7)

"… typically if they spent more time, they think they might be closer to the answer, but not always ... but it helps me judge, if they're spending time, what might be relevant, what might be part of their question … more time means more interest, not necessarily better. If they spend more than 10 seconds, there might be something that they were interested in" (Analyst 9)

"It can be the case that when they spend the time going to the end, they found it's not something they like, but in general, more time indicates more interests." (Analyst 12)

Analysts also had some consensus on the usefulness of the selection behavior. A typical comment said "selection is less trustworthy than query, but if results are very different from each other, selection can be helpful" (Analyst 13). A similar comment said "selection is especially helpful if there were pretty different web pages ... possibilities they can choose ... [selection] helps you see, oh, they meant this aspect" (Analyst 9). She further commented that "it can sometimes be in lieu of adding additional terms to see which one they click and which one they didn't".

There were comments on some other types of behaviors, such as exit type, mouse movements and scrolling. Several analysts mentioned that they were getting feedback from "knowing what they did next" (Analyst 8). In fact, almost all the analysts said things like "I want to see what they did next" and "I want to wait till I see more" during the study.

Only two analysts commented on mouse movements and both of them related mouse movements to eye movements. One of them said "[mouse movement is helpful] in lieu of the eye movement thing ... sometimes it's a little bit confusing because maybe they were hovering their mouse on this but their eyes were down here" (Analyst 9). The other said "that's helpful, too, but not as helpful as eye movement, because it does not always move … you can't tell if they read through the page, or didn't read through the page" (Analyst 13). These comments echoed the observation from previous studies on the difficulty in inferring eye-mouse coordination (e.g., Rodden & Fu, 2007). Another participant mentioned mouse movement, but actually referred to scrolling. She said "mouse movement helps … for example, knowing that they scrolled down makes a difference" (Analyst 10). Another comment on scrolling said "scrolling down means there is enough on the first fold which makes them want to scroll down" (Analyst 12).

In general, analysts considered queries and eye movements as the most reliable sources of evidence to infer searchers' interests. They also found time, selection, scrolling, exit type, and mouse movements to be useful in certain situations, but they often found it necessary to have additional evidence to interpret such evidence, as noted in the previous subsection.

6.2.5 Analysis on the stimulus level

This section presents comparisons at the stimulus type level. First, numbers of inference changes (in content and/or confidence level) made from each type of stimulus have been aggregated in Table 6.13. The fewest inference changes happened with Type A stimuli. This was partly due to the fact that Type A stimuli did not present screen shots or search segments resulting from clicks made on the external results content pages. However, there were only 5 instances of such clicks among the first 8 search cases and the investigator's observation suggests that the absence of those screen shots and search segments did not have a strong negative impact on the analysts. Instead, analysts made fewer inferences mainly due to the lack of the time information, which made them less ready to update their inferences. There were many more inference changes with Type D stimuli, but mainly due to the fact that analysts in the second group only used Type D stimuli. In total, Type D stimuli were analyzed 32 times while the other three types were only analyzed 16 times. So, the average number of inference change instances in Type D was slightly smaller than that in Type B. Overall, analysts updated their inferences more often when they worked with Type B and Type D stimuli. The percentages of positive change instances (shaded cells) were 62.3%, 59.3%, 57.7%, and 68.0% for the A, B, C and D types of stimuli respectively.

Table 6.13. Number of change instances broken down by stimulus type and directions of change in accuracy and confidence level

Stimulus A

		Accuracy of Inference			
		Up	**Same**	**Down**	**Total**
Confidence Level	**Up**	11	24	7	42
	Same	8	3	2	13
	Down	6	5	3	14
	Total	25	32	12	69

Stimulus B

		Accuracy of Inference			
		Up	**Same**	**Down**	**Total**
Confidence Level	**Up**	14	35	6	55
	Same	5	2	5	12
	Down	6	14	4	24
	Total	25	51	15	91

Stimulus C

		Accuracy of Inference			
		Up	**Same**	**Down**	**Total**
Confidence Level	**Up**	16	14	8	38
	Same	11	1	4	16
	Down	2	11	4	17
	Total	29	26	16	71

Stimulus D

		Accuracy of Inference			
		Up	**Same**	**Down**	**Total**
Confidence Level	**Up**	43	57	7	107
	Same	19	10	7	36
	Down	13	11	8	32
	Total	75	78	22	175

Table 6.14 lists the mean inference accuracy (rank score, as defined in 6.2.2.3) made from each type of stimuli on each search case in the first group. The higher the rank score is, the more accurate the inference was. The type of stimulus with the highest mean rank score of accuracy (i.e., the most effective stimulus) for each search case has been highlighted. Results suggest that none of the stimulus types excelled in all search cases. Overall, these results provide some evidence that the effectiveness of the four types of stimuli was comparable.

Table 6.14. Mean inference accuracy by search case and stimulus

	Search Case							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
A	6.86	14.38	16.64	19.37	9.18	13.29	8.33	7.89
B	6.57	14.70	18.73	21.88	7.36	15.30	9.18	9.08
C	6.09	13.22	21.36	16.79	10.08	8.08	6.50	9.00
D	5.67	20.33	20.50	17.56	8.78	11.59	8.78	9.08

There were some discussions from analysts in the first group who experienced different types of stimuli on the usefulness of some certain behaviors that are unique to a certain type of stimuli, both while they were analyzing some search cases and at the end of their participation. For example, when Analyst 3 continued to analyze Search Case 7 with the Type C stimulus after analyzing two search cases with the Type D stimuli, she mentioned that she missed the eye movements when the searcher spent 47 seconds on a page without scrolling. Similar comments have been made about time by analysts who

shifted to Type A stimulus after using Type B stimuli. These comments provide some evidence for the value of eye movements in Type D stimuli and time in Type B stimuli.

While using Type B stimuli, several analysts commented on the lack of time on the Google results list page. For example, Analyst 2 asked the investigator "you didn't have anything saying how much time they spent on Google results page? … because it would let me know if they rejected what would be my next best guess, or just go to the first one" (02-06B). There has been little research that studies time on search results list as a source of implicit feedback (for example, Fox et al. (2005) considered initial activity times, including time to first click), but it will be interesting to explore this issue further in future studies.

6.3 Summary of results

In the second phase of the study, 12 analysts evaluated a total of 80 search cases. Their goal was to make inferences about what the searcher was looking for based on the evidence from the recordings (screen shots or video segments). Inferences were elicited after each screen shot was shown or each video segment was played. Analysts were also encouraged to suggest any new evidence that they noticed and update their inferences at any time. Analysts were also required to indicate their confidence levels for the inferences on a 10-point scale and provide rationales for the inferences, especially the evidence that supported the inferences. The entire review process was audio recorded in sync with the screen contents.

Analysts took two approaches to making inferences. First, they directly learned of the searchers' interests based on some kinds of evidence, such as the query terms, snippets (mostly titles and summaries) of the selected results and the result contents. This was straightforward and often resulted in an update to the content of their inferences. The second approach relied on comparing expected behaviors with observed behaviors. Once the analysts saw the first query, they started making predictions about the searcher's interest and, based on that, making predictions about the searcher's next action if she was indeed searching for the predicted topic. Such predictions were updated every time a new piece of evidence was noticed. The second approach to making inferences thus involved comparing the observed behavior against the analyst's expectation. For example, when a selection was made by the searcher, the analysts would quickly determine if the selection was in the expected direction. If the selected page matched the analyst's prediction (i.e., a good selection), the analyst became more confident in the inference; otherwise, a bad selection would challenge their inferences, which resulted in either a change in the content of the inference and/or the drop of the confidence level. The same applied to the consideration of time spent on the page. Spending longer time on a page that contained useful information on the predicted search topic or spending shorter time on a page that did not contain such information would make analysts more confident, and vice versa.

In this study, analysts considered a wide variety of behaviors as indicators of searchers' interests. Most behaviors were considered from multiple perspectives. The

exposure to more evidence in a search case and the use of stimuli which provided richer evidence did not lead to consistently better inferences, but there were critical instances in some search cases which resulted in the increase of inference accuracy and confidence level for most analysts. In many instances, analysts referred to a combination of multiple behaviors or multiple aspects of the same behavior as evidence to support inferences. They also considered evidence instances through the search session to better interpret what some of the behaviors indicated. A number of rules for making inferences have been identified, some of which were more reliable and consistent while others were highly context dependent.

CHAPTER 7

DISCUSSION AND CONCLUSIONS

This chapter begins with a discussion of the major findings of the study. Each research question is presented along with the major findings that addressed the question. Potential explanations for the findings are discussed along with their implications. Comparisons are also made with findings from previous studies. Next, the chapter discusses the limitations of the study. Last, the chapter presents the conclusion of this dissertation and suggests directions for future work.

7.1 Discussion of results

This study was novel in that it was one of the first studies to examine the details and nuances in the use of behavioral evidence as implicit feedback for Web search and to achieve an in-depth understanding of the implicit feedback process through human reasoning. Five research questions were raised focusing on what types of behavioral evidence were considered, how they were used, and how effective they were in supporting the analysts' inferences about searchers' interests.

In the study, analysts considered a wide variety of searchers' behaviors as indicators of their interests throughout the search process, from behaviors on search pages, search results list pages, and external result content pages, to behaviors with regard to the search session. Searchers' behaviors on search pages mainly concern the submission of queries, but analysts considered not only texts of the queries, but also the query modification process, focusing on which terms were added, which were removed, which were removed and then put back, and how different the new query was from the previous one. These observations provide additional empirical support for some of the previous research on query modification, such as Jones and Fain's (2003) work on query term deletion, and Jones, Rey, Madani, and Greiner's (2006) work on query substitutions. In addition, analysts also benefited from analyzing the linguistic features of some queries and seeing some natural language queries.

Searchers' behaviors on search results list pages include mainly examination behaviors and selection behaviors. Eye movements, mouse movements and scrolling were useful examination behaviors that analysts considered. In addition, the relationship between searchers' eye movements and click behaviors was used to infer how confident the searcher felt with the selection. Searchers' result selection behaviors were considered from multiple perspectives, including not only the item which was selected, but also ones which were skipped, or the fact that no selection had been made on the page. This "no-action" perspective is particularly interesting as it represents some new opportunities to understand

searchers' interests in situations that have been typically considered as lacking informative evidence. It was also noted that analysts used texts in the surrogates, such as the titles, summaries and the URLs, to interpret the selection behavior as positive or negative search moves following the expectation notion.

Searchers' behaviors on external result content pages are most diversified, including viewing, mouse movement, eye movement, scrolling, searching within a page and page exit. Eye movement and scrolling were considered from several perspectives, most notably the speed and the repetitiveness of the actions. Content and structure of the page that the searcher examined was considered in association with behaviors and measures such as viewing time, eye fixations and scrolling fixations. The way the searcher exited the page (going back versus ending the search) was also used to interpret the examination behaviors on the page. However, mouse movement was rarely considered, perhaps an artifact of analysts not being used to seeing mouse moves and so not sensitive to what they might mean, whereas eye movement was more dramatic and obvious evidence of conscious behavior.

In addition to considering the query modification process, analysts also compared other types of behaviors through the session and compared the direction of the searcher's movements in the session with their expectations based on the predicted search topics. They sometimes related an earlier behavior to a later one, or used a later behavior to gain better understanding of an earlier one. Some analysts even took advantage of unintended evidence,

such as the position of the scroll bar in the left frame of the study interface, to infer the position of the session and interpret searchers' behaviors. In general, results demonstrate that analysts gained useful information about searchers' interests from many different types of behaviors and they considered the behaviors from multiple perspectives. This provides the answer to the first research question. Practically, this suggests that it is important for implicit feedback systems to monitor the additional browsing paths beyond the search results list page and capture searchers' behaviors both on search results list pages and on external result content pages. It also suggests that it is valuable to capture searchers' behaviors within a page, such as scrolling, in addition to page-level activities, such as link selection.

Some behavioral evidence of searchers' interests suggested by the analysts in this study are similar to those mentioned in previous studies. For example, in the prototype attentive agent, Suitor, that Maglio and colleagues (Maglio, Barrett, Campbell, & Selker, 2000; Maglio & Campbell, 2003) designed, eye movements were monitored while the user viewed web pages in order to determine whether the user was reading or browsing. Maglio and colleagues defined a document as relevant if reading was detected. Analysts in this study also considered the speed of eye movements and scrolling to determine if searchers were reading or scanning. Although this study did not identify new behavior that has not been mentioned in previous studies, it did reveal perspectives of considering some behaviors that were not mentioned before. Some of these perspectives could be very useful

in indicating searchers' interests. For example, taking a query term out and then putting it back into the query was mentioned as a strong indication of the searcher's interest in that concept. Due to the laboratory setup of the study, some types of behaviors that have been observed in other studies were not observed in this study. Such behaviors are mainly retention behaviors such as bookmarking and printing, but they also include some customizing behaviors such as resizing windows. These behaviors have been observed in other studies. The lack of them in this work should be considered as an artifact of the study design.

Based on these observations, a model of implicit feedback for Web search is summarized in Table 7.1. It extends prior classifications of behavioral evidence for implicit feedback proposed by Oard and Kim (2001) and Kelly and Teeven (2003) in three aspects. First, it focuses on Web search, so all the behavioral evidence considered in this model is related to the state-of-the-art commercial Web search systems, such as Google. Secondly, it is grounded in the data collected through an empirical study which captured real use of the behavioral evidence to infer Web searchers' interests by human analysts. Thirdly, it introduces a new and important level to the model, analytical lens, which reflects the wide range of perspectives that can be taken to use the behavioral evidence for implicit feedback.

Table 7.1. Model of implicit feedback for Web search

Search State	Strategic Evidence	Tactical Evidence	Analytical lens
Search page	Submit new query	Add initial terms	linguistic features of query terms, natural language query
	Modify existing query	Add terms	difference between new query and old query, linguistic features of query terms, natural language query
		Remove terms	
		Put terms back	
Search results list page	Page level tactics		time on the page
	Examine	Move eyes	relationship between click and scanning
		Move mouse	links that were hovered over
		Scroll	speed, depth
	Select	Select item (result)	title of selected result, summary of selected result, URL of selected result, relevance of selected result based on surrogate
		Select page	select next page of search results without changing the query
	Skip	Skip item (result)	title of skipped result, summary of skipped result, relevance of skipped result based on surrogate
			skip all results on page (lack of select)
External result content page	Page level tactics		time spent on page, relevance of page based on its content, page structure (text? list?)
	Read	Move eyes	speed (slow), fixation position, places where searcher spent a long time (focus), place that was focused on repeatedly, lack of focus on the page, exit position (where searcher looked at last)
		Move mouse	links that were hovered over
		Scroll	scrolled, lack of scroll, scroll speed, scroll depth, scroll fixation position, number of repeated scrolling
		Search within page	search terms
	Scan	Move eyes	speed (fast)
	Exit	Exit page	exit type (Back, END)
Search session	Not directly observable		continued searching instead of ending, stage in session, past behavior in the same search session

The model consists of four levels: search state, strategic evidence, tactical evidence, and analytical lens. It first groups observable behaviors for implicit feedback according to where and when in the search process (labeled "search state") that the behaviors can be captured. Different types of behaviors are available when searchers are in different states or on pages of different genres; the same type of behaviors may also be considered in different ways when they are captured on different types of pages. This distinction has implications for the implementation of implicit feedback technologies: behaviors on search pages and search results list pages can be captured through server-side logging techniques, which search engines can easily deploy, while behaviors on external result content pages, especially those more than one step from the results list pages, can only be captured through client-side techniques. In the second column, strategic evidence, behaviors are grouped according to the higher level search strategies that searchers took to achieve the information seeking goals when the evidence was observed. For example, two strategies were taken to search: to enter a new query and to modify an existing query. Analysts considered the differences between old and new queries as an evidence of searchers' interests and this was observable only when the "modify query" type of strategy was used by the searcher. At the next level, a search strategy consisted of several search tactics, conscious actions taken by the searchers to implement the search strategy. For example, to implement the "examine result page" strategy, searchers used a number of tactics, including moving eyes, moving the mouse, and scrolling. The searchers' uses of different tactics provided different types of

opportunities for the analysts to infer their interests. These tactics are summarized in the third column of the table. The totality of strategic evidence and tactical evidence represents the behavioral evidence that analysts considered. The last column of the table lists the analytical lens applied by the analysts when they considered the behavioral evidence. They were not searchers' behaviors by themselves, but rather the types of evidence that analysts used to interpret the behaviors.

A logical extension of the model is to consider metrics that can be used to capture and measure each type of tactical evidence. Some of them are more straightforward, such as the differences between two queries, while others are less clear, such as the identification of natural language queries. A systematic examination of the metrics will be left for future work.

To answer the second research question, analyses in Section 6.2.3 suggest that more evidence did not always lead to more accurate inferences and higher confidence levels. Instead, there were search cases for which more evidence led to steadily more confident inferences for most analysts, while for other cases, the confidence levels fluctuated for most analysts. The same applied to the change in inference accuracy. In total, about one third of the time, exposure to more evidence did not translate to better inferences (a better inference includes a more accurate inference and increased confidence, increased confidence in the same or an equally accurate inference, or the same confidence level on a more accurate inference). Working with stimuli which presumably contained richer evidence (Type B

compared to Type A, Type C compared to Type B, and Type D compared to Type C) helped some analysts with their decision making, but did not result in better inferences on average. The effectiveness of the four types of stimuli was comparable. This finding was a little surprising, but can possibly be attributed to two factors. First, all search cases were collected from inexperienced searchers and involved multiple rounds of query modification. A major problem with inexperienced searchers is that they are less skilled in modifying queries based on examination of initial results. Therefore, their search moves were often not well planned and did not accurately represent what they were actually looking for, which would easily confuse the analysts who were trying to infer the aims of their search. Secondly, many of the analysts were more used to viewing text than visual materials. Therefore, they were less sensitive to some of the evidence embedded in the videos and found it hard to follow the gaze path.

Some evidence was interpreted differently by different analysts so that it led to more confident inferences for some analysts and less confident ones for others, or a mixture of better and worse inferences. However, there were a few evidence instances which resulted in jumps in inference accuracy or confidence level for almost all analysts. These instances were mainly those when highly discriminating terms that represented crucial new facets were added into the queries, thus adding a lot more value than modifying an existing facet.

Finally, in no case did an analyst constantly become more confident each time she updated the inferences and/or confidence levels. Similarly, in no case did an analyst

consistently make equally or more accurate inferences each time she updated the inference and/or confidence levels. Analyses demonstrated that individual differences among analysts did not affect the results. Thus, it can be concluded that the process of inferring goals from raw search behavior traces is both complex and fluid.

The third research question was concerned with how behavioral evidence was used to support inferences. Results suggested that analysts frequently used multiple types of evidence to support an inference. Sometimes, a combination of multiple behaviors or a combination of attributes for the same behavior (e.g., results that were selected and those that were not selected) was considered within the same instance. Sometimes, behaviors from different instances were related to support an inference. These observations are explained by the theory of polyrepresentation (Ingwersen, 1996) which suggests that obtaining multiple representations of a single information need is a better approach to representing user needs than solitary, isolated queries. They also echo the findings from several other studies. For example, Claypool et al. (2001) found that the combination of time and scrolling led to the most accurate predictions of searchers' interests. Fox et al. (2005) found that the best predictive models at the page level combined clickthrough, time spent on the search result page, and how a user exited a result or ended a search session.

In addition, analysts' background knowledge of the search topic and the search context (i.e., the searchers, the study setup) were frequently referred to as factors (although not behavioral evidence) that affected the inferences. This provides empirical support for

the observation that implicit feedback should be interpreted within the larger context of the searcher's characteristics, tasks and search environment (Kelly & Belkin, 2001; Kelly & Teeven, 2003; White, Ruthven, & Jose, 2005; White & Kelly, 2006).

Although the search questions that were assigned to the searchers were quite homogenous (for example, there was no question involving finding multimedia information and no question for online shopping), the genre of the page still had some impact on the set of behaviors and measures that analysts considered on the page. Special page features, including images, lists and tables, offered further evidence for consideration in addition to the text. Analysts focused on different sets of behaviors on search results list pages versus external result content pages and gained information about the searcher's interest from both places. This again suggests that it is valuable to capture searchers' behaviors beyond the search results list pages and that pages in certain genres, such as those containing mostly bullet points, provide additional angles for understanding searchers' interests.

Finally, in response to the last research question, some common and more consistent rules for making inferences are summarized below. Some other rules (e.g., linguistic features of query terms) correctly provided useful information in some cases, but misled the analysts in other cases. They warrant further investigation, but are not included here.

Rule 1: Natural language queries can be used to directly interpret searchers' interests.

Rule 2: If an added term represented a new facet in the search topic and was from the results that the searcher examined before adding the term, the results were very likely to be relevant.

Rule 3: A term removed from the query was probably involved in some kind of relationship that the searcher was investigating. The searcher first assumed that the term was part of the relationship, but after examining the results felt that it actually was not. This is especially likely if the term represents a concept that is not expressed by other terms in previous or future queries.

Rule 4: A term removed from the query but later put back was a strong indication of the searcher's interest in the concept represented by that term.

Rule 5: If the searcher modified the query and the new query was close to the previous query, the query was probably close to what the searcher was looking for.

Rule 6: Selecting a result from a "credible" source (e.g., NASA, government sites) suggested that the searcher was looking for authoritative information. Selecting a Wikipedia page suggested that the searcher was looking for general information on the topic.

Rule 7: Seeing a good selection increased confidence levels while seeing a bad selection decreased confidence levels unless there was competing evidence.

Rule 8: If the searcher considered but skipped results which were significantly different from the one selected (e.g., providing new concepts), the searcher was probably

not interested in that additional concept. The difference between selected and skipped results should also be considered.

Rule 9: Positions of eye fixations suggested searcher interest, and repeated fixations in the same place were much stronger indications of interest than a single fixation. When eye movements were not available, scrolling could also be used to indicate the area of focus and the repetitiveness of focus.

Rule 10: A long time spent on examining a results list page before the first click or going to the second results list page without modifying the query provided indications that the results were close to what the searcher was looking for. A long time spent on the summary of a clicked result indicated that the text in the summary was relevant.

Rule 11: Clicking on a result and keeping reading other results while waiting for the page to be loaded meant a stronger confidence about the result than if the searcher moused-over a result, kept reading, and then clicked it.

Rule 12: The presence of scrolling suggested that some features of the page shown before the fold made the searcher believe that the page had the type of content and genre that she was looking for. The lack of scrolling could either mean that the searcher noticed something clearly wrong about the page from the top fold (such as the wrong format, or a 404 error), or that she found what she wanted without having to scroll down. Further evidence, such as viewing time and scrolling speed, could be considered to make better interpretations.

Rule 13: When searchers scrolled down on a page quickly without focused reading, the page either did not contain what they were looking for, or they were looking for specific information such as numbers or names which easily stood out.

Rule 14: If a searcher scrolled quickly on an external result content page and passed places where the keywords in the result summary appeared without reading them carefully, or if those places were low on the page but the searcher did not scroll far enough to reach there before going back, the searcher was most likely not satisfied with the page.

These rules suggest opportunities for designing implicit feedback algorithms to infer searchers' interests from their behaviors. The rules vary in feasibility and difficulty of implementation. Some of them were stand-alone, involving a single type of behavioral evidence, so they were most feasible to implement given the current logging techniques. For example, Rule 3, 4 and Rule 5 only involve the capturing of the query modification process. Rule 14 is also quite easy to implement, but would require client-side techniques to capture the searcher's scrolling behavior on external result content pages. The lack of attention to the places on an external result content page where the keywords in the result summary appeared would then suggest that the page can be used as negative feedback. To help searchers locate these places more easily, search engines can probably consider highlighting the sections where they extracted the summaries.

The implementation of some rules, such as 8, 9, and 11, requires the use of eye tracking techniques which are not currently feasible with end users, but the development of

more accurate and less intrusive eye tracking techniques may provide better support for these rules. Moreover, scrolling can also be used as a proxy to eye tracking, at a more coarse level.

Some rules were more complicated, involving a combination of different evidence that contextualize each other. The implement of these rules relies on observing and relating searchers' behaviors across several segments in the search session. For example, Rule 2 suggests an opportunity for positive feedback by observing the searcher adding terms from a previously examined page to the query, and its implementation involves monitoring both the query modification and contents of selected pages. The implementation of Rule 12 involves capturing the scrolling behavior and time, and Rule 13 involves capturing scrolling and page features. The implementation of some rules is contingent upon the availability other non-behavioral prerequisite information. For example, the implementation of Rule 1 requires the identification of natural language queries; the implementation of Rule 6 requires the identification of credible websites.

In some cases, a combination of rules needs to be considered to interpret some behaviors. For example, when a lack of selection is observed on a search results list page, the searcher's next behavior is needed to interpret this "no-action" action. If the searcher modifies the query, Rule 2, 3, 4 and 5 about query modification come into play. If the searcher goes to the second page of results list, part of Rule 10 can be applied. In general, this suggests that search engines should base their implicit feedback algorithms on the

totality of available evidence for the search session and actively update the representation of the searcher's interest not only based on the current query or the behaviors in the current search segment, but also on evidence from the searcher's past behaviors in the same search session, or even evidence from the search history (discussed in the user modeling and search personalization literature) and the behaviors from like-minded searchers (discussed in the collaborative search literature).

Another set of opportunities for monitoring and interpreting searchers' interests comes from search result pages with special features. Just like comparing search results which were clicked versus skipped on the search results list page, if a click is made on an external result content page and the clicked item is in a list (which can be detected from HTML list tags), it is useful to consider other items on the list that the searcher dismissed. Likewise, the occurrence of some special search terms in the query also represents special opportunities. Examples of such terms observed in the study include "statistics", "population", and "quote".

An important theme, which was not included in the original research questions, emerged from observing how analysts worked and analyzing the transcriptions. Analysts took two approaches to making inferences: a data driven approach and a knowledge based approach. In the data driven approach, analysts directly learned of the searchers' interests from some kinds of evidence, such as the query terms, snippets (mostly titles and summaries) of the selected results and the result contents, while the knowledge-based

approach relied on comparing observed behaviors with analysts' expectations of the searcher's actions based on their knowledge of the search context, the searcher, and the past behaviors. According to the available literature, the data driven approach is the basis of most current search engines' operations, while the knowledge-based approach is only implemented in limited domains (such as analyzing clickthrough data to model searchers' long term interests). This work suggests that search engines should continue to evolve from simple query-oriented IR systems to knowledge intensive operations that capture massive amounts of data to infer knowledge about the searcher's interest. By showing how human reasoning was used to obtain knowledge from the data, this study reveals possible avenues for automatic generation and use of the knowledge about searchers' interests through monitoring their behaviors. Although it is hard for machines to emulate all the human reasoning capability, some of the human reasoning processes can be captured through studies like this one. Moreover, machines can take advantage of their strengths in processing speed and memory size to access and analyze the past behaviors of large number of searchers that human analysts do not have access to and do not have the cognitive resources to process. It is conceivable that the development of search engines that leverage vast amounts of knowledge in and beyond the ways exemplified in this work is the key to taking the search technologies to the next level.

7.2 Limitations

Like all laboratory studies, this study suffered from the artificiality of searchers' behaviors since search questions were assigned and the search was conducted in a different environment than what the searchers were used to. This setup not only limited the types of behaviors that could have been observed, but also may have impacted the interpretations of some behaviors that were exhibited.

In the study, the elicitation of inferences and confidence levels were mostly made at predefined critical instances, although analysts were also encouraged to suggest any new evidence that they noticed and update their inferences at any time. This design was a result of the trade off between allowing the analysts to watch a relatively complete search segment and having them verbalize their thoughts as soon as possible. The decision was to use search segments and screenshots as the unit of presentation, but technically, this unit can be made smaller to allow for more zoomed-in examination of the implicit feedback process. For example, analysts sometimes updated their inferences once they saw a new query, without seeing the searcher's behavior on the results list page that Google returned for the query. However, the experimental setup did not accommodate the recording of such instances, especially for screen shot versions where the time corresponding to the end of the search segment when the searcher was on that page was used as the timestamp for any inference made using the screen shot.

Additionally, some specific aspects of the design of the experiment could be further improved. A few search questions had two parts while others only had one. This inconsistency affected analysts' inferences in some instances. Some analysts took advantage of the scroll bar (when available) at the left panel to judge the position of the segment in the entire search session, which, although pointing to the usefulness of this evidence, was artificial and should have been avoided.

7.3 Conclusions and future work

This dissertation presented a study which was designed to formally examine the range of evidence that searcher behavior offers and to understand how each kind of evidence can be used to infer the searcher's interest. The goals of this dissertation were accomplished by conducting a two-phase study in which Web search cases involving underspecification of information needs and modifications of search strategies were collected from inexperienced searchers as screen shots and videos in the first phase and analyzed by reference librarian analysts in the second phase. Analysts used evidence available from the recordings to infer the searchers' interests and explained what evidence they considered and how the evidence was used, in addition to making the inferences and stating their confidence levels with the inferences.

This is one of the first studies to gain in-depth understanding of the implicit feedback process and it used a novel approach to observing human analysts' reasoning

process when they simulated the role of an implicit feedback system. Results demonstrated that analysts considered a wide range of behaviors and most of the behaviors were used in multiple ways. Although all the behaviors have been mentioned in previous studies, the study discovered several new and useful ways of using searcher behaviors for implicit feedback that have not been studied before and revealed the nuances of evidence that analysts used in making inferences about searchers' interests. Key findings from this study are integrated as a model of Web implicit feedback presented in Table 7.1. The model consists of four levels: behavior category, strategic evidence, tactical evidence, and the analytical lens analysts used to make inferences about the intents behind the behavioral evidence. It bridges previous discussions on observable behaviors that can be used for implicit feedback (e.g., Oard & Kim, 2001; Kelly & Teeven, 2003) and those on implicit measures (e.g., Fox et al., 2005) or features (Agichtein et al., 2006). It introduces a new and important level to the model, analytical lens, providing a road map for future research on implicit feedback for Web search by suggesting the perspectives in which data should be collected when empirical studies are conducted on a particular behavior. For example, when scrolling is studied, it is not enough to just detect the presence of the scrolling event; instead, data should also be collected on the pattern of scrolling, such as its speed and depth. It also suggests directions for future work that will elaborate the analytical lens with the aim of identifying measures of evidence that may in turn be incorporated into algorithms.

Findings of the study suggest that it is important for implicit feedback systems to monitor the additional browsing paths beyond the search results list page and capture searchers' behaviors both on search results list pages and on external result content pages. They also suggest that it is valuable to capture searchers' behaviors within a page, such as scrolling, in addition to page-level activities, such as link selection. The study put forward design recommendations for implicit feedback systems based on some of the rules that analysts used in making references. Some of these design recommendations can be readily turned into algorithms.

This study is part of an overall attempt to develop technologies that assist searchers with difficulties in formulating effective search queries. It complements research in explicit feedback and other approaches. There have been suggestions for combining implicit and explicit feedback techniques. For example, Nichols (1997) suggested combining implicit ratings with existing rating systems to form a hybrid system and using "implicit data as a check on explicit ratings" (p.5). Gadanho and Lhuillier (2007) also argued for a hybrid system using both explicit user modeling and implicit user modeling. In the future, different approaches to achieving such integration can benefit from the results presented here.

This dissertation addresses some key questions in Web implicit feedback, namely the nature of the behavioral evidence for searchers' interests and how it can be used. There are other important issues that remain to be addressed for this topic. First of all, this dissertation relied on the manual review process to select search cases that involved

underspecified queries. In practice, underspecification is not a clear cut concept; instead, queries can be viewed as existing in a continuum of specification. Therefore, when search assistance is provided, automatic techniques need to be developed to determine the level of underspecification and deploy implicit feedback techniques only when the level exceeds a certain threshold so as to minimize system cost and distraction to the searcher. There has been little research on how to identify underspecified queries. One possible method is to analyze the diversity of the search results using clustering techniques (Zamir & Etzioni, 1999; Dumais, Cutrell, & Chen, 2001). Given the same clustering parameters, if results are clustered into a small number of clusters, it can be assumed that the query was well specified so that it returned a homogenous set of results. On the contrary, when the query is broad and open to different interpretations, the results should be diversified and more clusters should be formed. This represents the situation when search assistance should be provided. Other possible approaches to identifying underspecified queries include query clarity measures (Cronen-Townsend, Zhou, & Croft, 2002) and query difficulty prediction methods introduced in the 2004 TREC Robust Track (Voorhees, 2005).

Time heuristics can also be used to trigger implicit feedback algorithms. In this study, several analysts considered the time that searchers spent in the search sessions to infer if they were more likely to be frustrated with the results. They felt that a searcher spending too much time in a session was an indication that she might have difficulty, but they were not able to quantify the threshold for "too much time". However, this did suggest

a useful perspective to analyze the searcher's progress. A more effective way may be to consider time in combination with other features of the search session, such as the number of queries, the average amount of time the searcher spent on a result page, and the overlap between queries. If the searcher spends a short amount of time on most results while repeatedly trying new queries which were similar in content, that is a clear indication that the searcher has difficulty adjusting the search strategy to get better results. To get more insights into such situations, search cases ending with failures should be collected and analyzed in the future.

Another issue to be addressed arose from the discussions with the analysts. It was concerned with the delivery of search assistance. Two general principles were suggested by the analysts.. First, no matter what type of assistance is provided, caution must be used when communicating the intention of the assistance with the searchers so that they understand why the assistance is provided and they have the option of declining the assistance if they so wish. Some of the analysts reflected upon their experiences during reference interviews and pointed out that if assistance was offered too quickly, or in an intrusive way that the patron felt that they did not have control over the assistance, the patron would feel disrespected and not pay due attention to the assistance. For Web search assistance, presenting search recommendations like Google spelling error suggestions ("Did you mean …") and offering the control to the searcher are some viable options.

Second, as behaviors are less reliable indicators of interests in general, it is more appropriate to use them for less radical changes to the search strategy, such as re-ranking of the results, than more radical ones, such as query expansion. For example, Shen et al. (2005a) designed the UCAIR toolbar to capture a searcher's search context and history information and use it to re-rank unseen results when the searcher clicks the "Back" button or the "Next" link. Promoted results are indicated with an up arrow at the end of the result surrogate. There are other options, too, such as using the last (usually the 10^{th}) space at the bottom of each results list page to display the top result after modifying the query or re-ranking the unseen results based on incorporating implicit feedback collected from the searcher's behavior on this page. The 10^{th} space is a good position because by the time the searcher reaches there, she should have left a relatively rich set of behaviors for consideration (scrolling through results list page, result clicks and skips). Alternatively, the first space on the top of each subsequent results list page beyond the first page can also be a good candidate position to place promoted result after capturing all the evidence from the first page and knowing that the searcher has decided to examine more of the results, rather than modifying the query. In either case, client-side techniques such as Ajax can be used to support dynamic display of the promoted result if certain implicit feedback conditions are met. Clearly, more research needs to be conducted to test the different options and select a better approach to delivering search assistance based on implicit feedback.

There are also various ways to extend this work. On the one hand, the data generated from this study can continue to be explored to gain further insights into the research questions. For example, only 10% of the search cases collected in the first phase have been analyzed for the purpose of this study. It will be interesting to analyze the remaining search cases to gain a better understanding of how people search complex topics.

On the other hand, some future data collections are planned to further evaluate the useful evidence and useful rules to make inferences identified in this study. Firstly, to directly compare with the findings of the current work, a future study will be carried out to recruit people who are familiar with the search engine algorithms (i.e., search engine designers) as analysts to examine the same search cases and see whether they interpret the behaviors differently from the reference librarians given their different expertise. Secondly, the rules for making inferences about searchers' interests that were identified in the study will be tested empirically. The plan is to collect a new data set of Web search cases which include not only screen recordings but also quantitative logs of searcher activities (such as key strokes, mouse clicks, and positions of the mouse). Algorithms that implement some of the 14 rules will be applied to this new data set to automatically make inferences about searchers' interests. These inferences will then be compared with human inferences and/or end user evaluation to test the effectiveness of the rules.

Finally, this dissertation only studied the implicit feedback from a relatively homogeneous group of searchers (university staff, inexperienced searchers) searching on a

small and relatively homogenous set of search topics (multi-faceted questions with close-ended answers). To develop more robust implicit feedback systems, more work needs to be done to study more searchers with different characteristics searching on more diversified topics and compare people's behaviors in these different environments.

In conclusion, this research has contributed to a better understanding of the different behavioral evidence of searchers' interests in Web search, what they mean and how they can be used as implicit feedback. The research findings have practical implications for designing implicit feedback techniques that provide assistance to searchers with difficulties in specifying their information needs. They also suggest future research agendas that can further address the issues involved in Web implicit feedback.

APPENDICES

Appendix A:

Recruitment advertisement for Phase I

We are soliciting volunteer participants for a study from June 25 to July 20 investigating how people use Web search engines, such as Google. The purpose of the study is to inform designs of intelligent interfaces that are more adaptive to people's behavior when they search the Web. Your participation is very important to us. It will help the development of personalized search engines that deliver better results and bring better user experiences. This study has been approved by the UNC Behavioral IRB (IRB Study 07-0944).

If you are interested in participating in this study, you will first need to visit this web site: http://www.ils.unc.edu/webstudy and fill out a brief questionnaire about your search experiences and complete a small query formulation exercise. Respondents will be screened based on search experiences and how well queries are formulated. Although you may or may not be selected to participate in the study, you will be entered in a drawing for a $25 Best Buy gift card as long as you complete this questionnaire.

If you are selected to participate in the study, you will be contacted via email and asked to select a study session that fits with your schedule. The study will take approximately 1 hour. It will take place in a computer lab at the School of Information and

Library Science (Manning Hall) on the UNC campus. You will be asked to use a search engine of your choice (e.g., Google, Yahoo!, AOL) to search on 6-8 topics. Your interactions as you search will be logged for later analysis, which include everything that happens on the screen and where you look at on the screen, but the logs will not include any identifying information about you. You will also be asked to answer several short questions about your experience and your answers will be audio recorded only for transcription purposes. All study sessions will be conducted individually. You will be offered $10 or some souvenirs (e.g., T-shirts, USB drives) from search engine companies (e.g. Google, Microsoft) as a token of our appreciation of your help.

Please email me at websearchstudy@unc.edu if you have questions about the study.

Appendix B:

Search problems used in Phase I

Topic a: Your friend is coming to visit you next week. You know she really likes Chinese cuisine. Please find a good restaurant that you can take her to dinner during her visit.

(Facets involved: topic – Chinese restaurant; quality – good food; location – where the searcher lives)

Topic b: Your friend visited the Kennedy Space Center recently. When he was there, he watched a movie about the Apollo Project. The video included a segment showing President Kennedy announcing the lunar landing project. Your friend vaguely remembers that President Kennedy said something like the project was undertaken not because it was easy, but because it was difficult. Can you find the exact quote for what President Kennedy actually said and where he made the speech?

(Facets involved: topic – speech, Apollo Project; person – President Kennedy; keywords – easy, difficult; question – quote, location)

Topic c: Your niece was watching TV last weekend and saw a team sport like the one shown in the picture. Your niece was curious what they were doing and what they used the brush for. Can you try to find the answer to her questions?

(Facets involved: topic – sport; location – on ice; tool – brush)

Topic d: I heard that the famous Tar Heel basketball coach Roy Williams used the quote "Don't be pushed by your problems; be led by your dreams" to inspire his player. Can you help me to find if he was the original source for the quote?

(Facets involved: topic – quote; keywords – "Don't be pushed by your problems; be led by your dreams"; question – author)

Topic e: My nephew is doing a school project on the deaf population. He wants to find out how many deaf people in the U.S. speak English and also use the American Sign Language. Can you help him?

(Facets involved: topic – deaf, communication, population; location – United States; question – statistics/usage/census/percentage; keywords – American Sign Language, English, bilingual)

Topic f: My neighbor has a dog. He noticed that his dog sometimes eats June bugs. He wonders if this will cause any problem to his dog. Can you look for some information to answer his question?

(Facets involved: entity – dog, June bug; relationship – eat; question – harmful/hurtful/toxic)

Topic g: You went to NC History Museum over the weekend and saw a picture showing a railroad spur built in Durham which ran directly into the American Tobacco Company. When you came home, however, you realized that you didn't pay attention to the time when it was built. Can you do a search and find out that?

(Facets involved: topic – railroad spur, American Tobacco Company; location – Durham; question – time)

Topic h: The gas price in Chapel Hill as well as in other U.S. cities has been going up crazily since earlier this year. It is costing roughly $3 per gallon now. You are curious about the situation in European countries. Can you do a search to find out what the situation is like there?

(Facets involved: topic – gas price; location – Europe; time – July 2007)

Topic i: You heard that in Russia, people drink a lot of vodka. Can you find out on average how much vodka a Russian drink?

(Facets involved: topic – drink, vodka; question – consumption/statistics/amount; location – Russia)

Topic j: My friend John has a farm in North Central Arkansas. He is interested in knowing if he can use his farm to grow plants for the production of bio fuels. He wants to know what plants he should grow in his farm and whether there is a market to sell the plants.

(Facets involved: entity – biofuel plants; topic – production, market/sell; location – North Central Arkansas)

Topic k: Your neighbor has a boy who is diagnosed to have ADHD (Attention Deficit Hyperactivity Disorder). Can you find some information on how diet/sugar affects ADHD in kids.

(Facets involved: topic – ADHD; entity – kids, diet, sugar; relationship – affect)

Appendix C:

Interview script for Phase I

Pre-search:

On a 7-point scale, with 1 being the least familiar and 7 being the most familiar, how familiar are you with the this topic?

Answer: ( )

[The searcher does the search.]

[The searcher finishes the search and signals the investigator.]

Post-search

Great! So, what did you find? (Question should be customized according to the task)

Now, please think about the search you've done just now. Do you think your initial query clearly stated what you wanted?

Did you learn something in the search process which made you change your search strategy? If so, what are some of the critical instances which triggered the change?

Appendix D:

Verbal overview of the first phase of the study

Welcome to the study! There are a few things we need to go over before we start. If you have a cell phone with you, you might want to turn it off now.

In today's study, I'm going to ask you to look for a few different things on the web. Please do whatever you'd normally do if you were searching at home. I know you might not be interested in all the topics, but please try to pretend that they are something you really want to look for and try your best. It would be most helpful if you can forget that you are in a study and searching for something that I give. Treat it as if you were at home and searching for your own questions.

You don't need to talk to me or tell me what you're doing. I'll be sitting there all the time during the study, and if at any point you're not sure what you're supposed to do, please ask me. Other than that, please ignore me. I will read you the questions one at a time. You can ask me to repeat the questions as many times as you want, but I can't make any clarifications. You have to interpret the questions by yourself.

When you finish a search, please close all additional browser windows that you opened during the search and bring the main browser window to the home page. [Ask which browser the participant uses most often. Show her the Home button if necessary.] Then, please tell me that you are done. I will ask you a few questions about the search before we move to the next question.

We have a microphone here and I'll also be using some software to record our voices, plus everything that happens on the screen. This is so that I don't have to take too many notes during the study and can go back and review things later. This [gesture] is an eye-tracker. It will tell me where you look at on the screen while you search. This will also be recorded. Otherwise, we don't have any hidden video cameras, so your face isn't being recorded anywhere. Also, we will not use your name in connection with the recordings or the results. The study is also described in this consent form, so please read and sign the consent form before we start.

[Participant reads and signs the consent form.]

OK, let’s start by doing some setup for the eye-tracker.

Appendix E:

Queries for Search Topic e

(064)

percentage of deaf Americans “united states” speak english

ASL percentage of deaf Americans “united states” speak english

“ASL and English” percentage of deaf Americans “united states”

“speak english and asl” “ASL and English” percentage of deaf Americans “united states” – no result

“speak english and asl”

“speak english and asl” united states percentage – 1 result

“speak english and asl” united states proportion – 1 result

“english and asl” united states proportion – 1 result

both english and sign language asl

(102)

population of deaf Americans

population of deaf Americans that use english

deaf Americans that use English

deaf Americans that use English sign language

english sign language

english sign language users

“english sign language” users

deafness statistics

(014)

deaf people in us who use both american and english sign language

usage of sign languages

statistics of sign languages

statistics of american sign languages

proportion both english and american sign language – 3 results

proportion both british and american sign language – 3 results

(165)

USA deaf population

percent deaf population communication

percent deaf population communication usa

deaf persons ASL communication usa

deaf using ASL usa

deaf using ASL usa how many

census deaf population communication USA

(137)

sign language

sign language english

sign language english usage

american sign language english usage

use both american sign language and english

h american sign language (typo from the searcher)

(108)

+“american sign language” +deaf +population +speak

+us +deaf +population +“american sign language” +speak

+us +deaf +population +statistic +“american sign language” +speak

+us +“american sign language” +speak +english

+deaf +statistic +us +“american sign language” +speak +english

(029)

“Users of American Sign Language in America” -- no result

“Users of American Sign Language”

“Users of American Sign Language” AND “Deaf People”

“Users of English” AND “Deaf People”

“Users of English” AND “Deaf People in America” -- no result

“English” AND “Deaf People”

gallaudet.edu

(160)

Statistics - ASL and English Users

Statistics - Simultaneous ASL and English Users

Estimates of Simultaneous ASL and English Users

Total ASL and English Speakers in US

(207)

deaf, us, statistics

deaf, american sign language

deaf, us, statistics

american sign language prevalence

american sign language, statistics

american sign language, lipreading

deaf, us, speaking

deaf who can speak

deaf who can speak, statistics

us, deaf who use english

us, deaf who use english, statistics

english speaking deaf, statistics

deaf who speak english

(036)

deaf speaking

deaf speaking and sign language

speech sign language

deaf communication

deaf bilingual

deaf resources

deaf language skills

speech and sign language

deaf

two languages

two languages spoken and visual

sign language

(079)

deaf people who speak english

term for deaf people who speak english and use sign

speak english and use American sign language

Deaf signers who speak english

speak english and use American sign language

Deaf signers who speak English

Appendix F:

Verbal overview of the second phase of the study

Thank you for agreeing to participate in the Web search analysis study.

[For participants in the first group before they used Type A/B stimuli]

Let me first tell you briefly what we are going to do today – basically, I am going to show you 4 recordings of Web searches. They will appear as a series of screen shots of the Web pages that the searcher visited during the search. So imagine you do a search on Google. You first go to Google's homepage, type in a query, get the results list; then you probably click on a result and do some reading there; then you will probably come back to Google and click on more results, or modify the query, so on and so forth. So, I captured a screen shot for each page that the searcher visited during the search, including the Google search results page and other pages that the searcher clicked on.

To put it in a simple way, your goal is to infer what the searcher was looking for. As you see more and more pages that the searcher visited, you will probably make better and better inferences. But remember that I'm most interested in how you make the inference, so I'd like you to think aloud while you watch and also whenever you see a new piece of useful evidence, which either reinforces or challenges your current inference, please let me know. I'll ask you a few questions, such as what you learn, how you learn it, and how confident you are with your inference. Of course you don't have to memorize these questions for now. I'll probe you as we go along. We'll also work together on a warm up

task in just a minute. The only thing you need to remember is to let me know whenever you see something interesting. Please do not wait till you are sure about your inference. I am more interested in how you reach that level of confidence and I am not measuring your ability to make the inference. So, please tell me whatever goes into your mind, no matter how confident you are. I am not doing a speed test, either, so you can spend however long you want on any page and use any evidence that is available to you to make the inference.

[For participants in the first group before they used Type C/D stimuli and participants in the second group who only used Type D stimuli]

Let me first tell you briefly what we are going to do today – basically, I am going to show you 4 recordings of Web searches. They will appear as a series of video segments showing consecutive episodes of the searches. So imagine you do a search on Google. You first go to Google's homepage, type in a query, get the results list; then you probably click on a result and do some reading there; then you will probably come back to Google and click on more results, or modify the query, so on and so forth. So, I record the search session as a video and cut them into smaller segments whenever the searcher clicks on a result or returns to Google results page after viewing a result.

To put it in a simple way, your goal is to infer what the searcher was looking for. As you see more and more search segments, you will probably make better and better inferences. But remember that I'm most interested in how you make the inference, so I'd like you to think aloud while you watch and also whenever you see a new piece of useful

evidence, which either reinforces or challenges your current inference, please let me know and I will pause the video for you. I'll also ask you a few questions, such as what you learn, how you learn it, and how confident you are with your inference. Of course you don't have to memorize these questions for now. I'll probe you as we go along. We'll also work together on a warm up task in just a minute. The only thing you need to remember is to let me know whenever you see something interesting. Please do not wait till you are sure about your inference. I am more interested in how you reach that level of confidence and I am not measuring your ability to make the inference. So, please tell me whatever goes into your mind, no matter how confident you are. I am not doing a speed test, either, so please feel free to pause and rewind the recording at any time you feel necessary to reexamine some part in detail or discuss it with me. Remember you can use any evidence that is available to you to make the inference.

[For all participants]

Is this clear?

I think it will also be helpful if I tell you a little bit about how those searches were collected. To collect them, I recruited about 20 searchers. Many of them were inexperienced Web searchers who claimed to have difficulties with searching Google. I paid them 10 dollars and had them do about 5 to 6 Web searches in an hour in my lab. I read them the questions one by one. Then they worked on their own without telling me what they were doing while they searched. I recorded the screen contents. All the tasks were a little

complicated involving multiple facets. So there is no question like "who is the president of the United States". Instead, they can be something like "My sister bought a Nikon digital camera online from CompUSA last weekend. However, she noticed that the camera was not what she liked. She wants to see if she can return the camera for a refund and how she can do it." As you can see, this topic involves multiple facets. First, the item is a Nikon digital camera, but this searcher was not interested in buying a digital camera; instead she was looking for the return policy. Please notice that I always described the search task in a scenario like the one just now and left it to the searcher to determine what is required and to formulate a query. I never stated the question as "please find the return policy of a Nikon digital camera bought from CompUSA" although that is what the question was asking essentially. Also, most of the questions are pretty close ended, with a best answer. This means, I never asked people to find, for example, recent research on heart failure medicine, which is very broad and open-ended. I allowed the searcher to personalize some of the questions. For example, if I ask them to find a place to buy a TV, then they can personalize the tasks by considering their own budget, preference, etc., but still, the answer should be fairly closed, rather than a set of Web pages. It turned out that most of the searches ended with some level of success.

I hope this has given you some context of what you are working on today. To highlight:

They are paid searchers, mostly less experienced with Web search.

They did the search in my lab and the search tasks were given to them.

The search questions were multifaceted, but were close ended with certain expected answers, although some questions can be personalized.

The 4 searches you are going to work on were done by 4 different searchers, so you should not relate the behavior pattern you may discover from one recording to another recording.

For each search you are going to analyze, I want you to infer what I asked the searcher to do. Remember, you can use all the evidence that is available to you and you don't need to rush. Some of the evidence may not be very obvious, so you sometimes need to pay full attention and think hard.

[For participants in the first group before they used Type A/B stimuli]

Do the training B on ADHD; show the interface for A and point out the difference

[For participants in the first group before they used Type C/D stimuli]

Do the training C on biofuels, point out things to watch; watch 2 segments from D and point out difference between C and D

[For participants in the second group]

Do the training D on gas price, point out things to watch

[For all participants]

Start recording with Camtasia

REFERENCES

AbdulJaleel, N., Corrada-Emmanuel, A., Li, Q., Liu, X., Wade, C., & Allan, J. (2004). UMass at TREC 2003: HARD and QA. In *Proceedings of the Twelfth Text Retrieval Conference (TREC '03)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning user interaction models for predicting web search result preferences. In *Proceedings of SIGIR '06,* Seattle, WA, USA, 3-10.

Allen, B. (1989). Content analysis in library and information science research. *Library and Information Science Research, 12*(3), 251-262.

Anick, P. (2003). Using terminological feedback for web search refinement: a log-based study. In *Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. 88-95.

Arroyo, E., Selker, T., & Wei, W. (2006). Usability tool for analysis of web designs using mouse tracks. In *CHI '06 Extended Abstracts*, 484-489.

Atterer, R., Wnuk, M., & Schmidt, A. (2006). Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proceedings of WWW 2006*, 203-212

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*, 407-424.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly Analysis of a Very Large Topically Categorized Web Query Log. In *Proceedings of SIGIR'04*, Sheffield, South Yorkshire, UK, 321-328.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science, 5, 133-143.

Belkin, N. J. (1993). Interaction with texts: information retrieval as information seeking behavior. In *Information Retrieval 1993: von der Modellierung zur Anwendung. Proceedings of the First Conference of the Gesselschaft fur Infomatik Fachgruppe Information Retrieval,* Regensburg. Konstanz: Universitatsverlag Konstanz, 55-66.

Belkin, N. J., Cool, C., Kelly, D., Lin, S. J., Park, S. Y., Perez-Carballo, J., & Sikora, C. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3), 404-434.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems With Applications, 9*(3), 379-395.

Belkin, N. J., Seeger, T., & Wersig, G. (1983). Distributed expert problem treatment as a model for information system analysis and design. *Journal of Information Science 5*(5), 153-167.

Bennett, J. L. (1972). The user interface in interactive systems. *Annual Review of Information Science & Technology, 7*, 159-196.

Bilal, D. (2000). Children's use of Yahooligans! web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science and Technology, 51*, 646-665.

Boekelheide, K., Brown, E., Fu, X., Marchionini, G., Oh, S, Rogers, G., et al. (2006). Audio Surrogation for Digital Video: A Design Framework. UNC SILS technical report. Retrieved January 29, 2007, from http://sils.unc.edu/research/publications/reports/TR-2006-02.pdf.

Boren, T., & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(33), 261-278.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56*(1), 71-90.

Byrne, M. D., Anderson, J. R., Douglass, S., & Matessa, M. (1999). Eye tracking the visual search of click-down menus. In *Proceedings of CHI '99*, 402-409.

Catledge, L. D., & Pitkow, J. E. (1995). Characterizing browsing strategies in the World Wide Web. *Computer Networks and ISDN Systems, 27*(6), 1065-1073.

Chen, M.-C., Anderson, J. R., & Sohn, M.-H. (2001). What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, Seattle, Washington, USA, 281-282.

Choo, C. W., Detlor, B., & Turnbull, D. (1999). Information Seeking on the Web - An Integrated Model of Browsing and Searching. In *Proceedings of the 62nd Annual Meeting of the American Society of Information Science*, Washington, D.C. Retrieved March 13, 2007, http://choo.fis.utoronto.ca/fis/respub/asis99/

Claypool, M., Le, P., Waseda, M., & Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI '01)*, USA, 33-40.

Cooper, M. D., & Chen, H. (2001). Predicting the relevance of a library catalog search. *Journal of the American Society for Information Science and Technology, 52*, 813-827.

Cox, A. L., & Silva, M. M. (2006). The role of mouse movements in interactive search. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society,* Lawrence Erlbaum, Mahwah, NJ, 1156-1161.

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002) Predicting query performance. In *Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '02)*, Tampere, Finland, 299-306.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*, 1827-1837.

Duchowski, A. (2002). A breadth first survey of eye-tracking applications. *Behavior Research Methods, Instruments and Computers, 34*(4), 455–470.

Dumais, S. T., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*, Seattle, Washington, USA, 277-284.

Eick, S. G., Steffen, J. L., & Sumner, Jr., E. E. (1992). SeeSoft -- A tool for visualizing line oriented software statistics. *IEEE Transactions on Software Engineering, 18*(11), 957-968.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Boston: MIT Press.

Fox, S. (2003). Evaluating implicit measures to improve the search experience. Presented at *SIGIR 2003 Workshop on Implicit Measures of User Interests and Preferences*, Toronto, Canada. Retrieved April 2, 2004, from http://research.microsoft.com/sdumais/SIGIR2003/FinalTalks/Fox-SIGIR2003_Fox_Presented.ppt

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems, 23*(2), 147-168.

Freund, L. & Toms, E. G. (2002). A preliminary contextual analysis of the web query process. In *Proceedings of the 30th Annual Conference of the Canadian Association for Information Science* (Toronto, ON). 72-84.

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B., & Coyle, M. (2007). Collecting community wisdom: Integrating social search & social navigation. *Proceedings of the Intelligent User Interfaces Conference (IUI '07)*, 52-61.

Fu, X., Kelly, D., & Shah, C. (2007). Using collaborative queries to improve retrieval for difficult topics. In *Proceedings of the 30th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '07)*, Amsterdam, the Netherlands, 879-880.

Furnas, G. W. (1981). The FISHEYE View: A New Look at Structured Files. *Bell Laboratories Technical Report*, Murray Hill, New Jersey. Retrieved January 28, 2007, from http://citeseer.ist.psu.edu/furnas81fisheye.html

Furnas, G., Landauer, T., Gomez, L., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM, 30*(11), 964-971.

Geisler, G. (2003). *AgileViews: A Framework for Creating More Effective Information Seeking Interfaces*. Unpublished Ph.D. dissertation, SILS, the University of North Carolina, Chapel Hill, NC.

Goecks, J., & Shavlik, J. (2000). Learning users' interests by unobtrusively observing their normal behavior. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, New Orleans, Louisiana, United States, 129-132.

Goldberg, J. H., & Kotval, X. P. (1999). Computer Interface Evaluation Using Eye Movements: Methods and Constructs. *International Journal of Industrial Ergonomics, 24,* 631-645.

Goldberg, J., Stimson, M., Lewnstein, M., Scott, N., & Wichansky, A. (2002). Eye Tracking in Web Search Tasks: Design Implications. In *Proceedings of the Eye Tracking Research & Applications (ETRA) Symposium*. New Orleans, LA.

Golovchinsky, G., Price, M. N., & Schilit, B. N. (1999). From reading to retrieval: Freeform ink annotations as queries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, USA, 19-25.

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of SIGIR '04*, United Kingdom, 478-479.

Granka, L., & Rodden, K. (2006). Incorporating eyetracking into user studies at Google, position paper for workshop "Getting a Measure of Satisfaction from Eyetracking in Practice". In *Proceedings of ACM CHI 2006*.

Greene, S., Marchionini, G., Plaisant, C., & Shneiderman, B. (2000). Previews and overviews in digital libraries: Designing surrogates to support visual information seeking. *Journal of the American Society for Information Science, 51*(4), 380-393.

Griffiths, J., Hartley, R., Willson, J. (2002). An improved method of studying user-system interaction by combining transaction log analysis and protocol analysis. *Information Research, 7*(4). Retrieved February 8, 2007, from http://InformationR.net/ir/7-4/paper139.html

Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's Web use skills. *Journal of the American Society for Information Science and Technology, 53,* 1239−1244.

Harper, D. J., Koychev, I., Sun, Y., & Pirie, I. (2004). Within-document retrieval: A user-centered evaluation of relevance profiling. *Journal of Information Retrieval, 7,* 265-290.

Hert, C. A., & Marchionini, G.. (1998). Information seeking behavior on statistical websites: Theoretical and design implications. In *Proceedings of the American Society for Information Science Annual Meeting*, Pittsburg, PA, 303-314.

Hijikata, Y. (2004). Implicit user profiling for on demand relevance feedback. In *Proceedings of the 9th International Conference on Intelligent User Interfaces,* Portugal, 198-205.

Hoffman, J. E. (1998). Visual attention and eye movements. In: Paschler, H. (ed.) *Attention.* London: University College London Press, 119-154.

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. *The International Journal of Computer and Telecommunications Networking, 33*(1-6), 337-346. Retrieved March 15, 2007, from http://www9.org/w9cdrom/81/81.html.

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

Ingwersen, P. (1982). Search procedures in the library - analysed from the cognitive point of view. *Journal of Documentation*, 38 (3), p.165-191.

Ingwersen, P. (1992). *Information retrieval interaction*. London: Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*(1), 3-50.

Jacob, R. (1991). The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems, 9*(2), 152-169.

Janes, J. (1991). Relevance judgments and the incremental presentation of document representations. *Information Processing & Management. 27*(6), 629-646.

Jansen, B. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research, 28*, 407–432.

Jansen, B. J., & McNeese, M. D. (2005). Evaluation the effectiveness of and patterns of interactions with automated searching assistance. *Journal of the American Society for Information Science and Technology, 56*(14),1480–1503.

Jansen, B. J., & Pooch, U. (2001). A review of web searching studies and a framework for future research. *Journal of the American Society for Information Science, 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 & Management, 36*, 207-227.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining,* Edmonton, Alberta, Canada, July 23 - 26, 2002, 133-142.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately Interpreting Clickthrough Data as Implicit Feedback. In *Proceedings of SIGIR 2005* (Salvador, Brazil, August 15-19, 2005). 154-161.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. & Gay, G. (2005). Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems (TOIS), 25*(2), 7.

Joho, H., Coverson, C., Sanderson, M., & Beaulieu, M. (2002). Hierarchical presentation of expansion terms. In *Proceedings of the 17th Annual ACM Symposium on Applied Computing (SAC '02)*, Madrid, Spain, 645-649.

Jones R. & Fain, D. (2003). Query word deletion prediction. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*, Toronto, CA, 435-436.

Jones, R., Rey, B., Madani, O., & Greiner, W (2006). Generating query substitutions. In *Proceedings of the 15th International Conference on World Wide Web (WWW '06)*, Edinburgh, Scotland, 387-396.

Jung, S., Herlocker, J. L., & Webster, J. (2007). Click data as implicit relevance feedback in web search. *Information Processing & Management, 43*(3), 791-807.

Just, M. & Carpenter, P. (1980). A theory of reading: from eye fixation to comprehension. *Psychological Review, 87*, 329–354.

Kantor, P. B., Boros, E., Melamed, B., Meñkov, V., Shapira, B.,& Neu, D. J. (2000). Capturing human intelligence in the net. *Communications of the ACM, 43*(8), 112-115.

Kelly, D. (2004). *Understanding implicit feedback and document preference: A naturalistic user study.* Unpublished Ph.D. dissertation, Rutgers University, New Brunswick, NJ.

Kelly, D. (2005). Implicit feedback: Using behavior to infer relevance. In A. Spink and C. Cole (Ed.) *New Directions in Cognitive Information Retrieval*. Netherlands: Springer Publishing. 169-186.

Kelly, D. & Belkin, N. J. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preference for relevance feedback. In *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '01)*, New Orleans, LA., 408-409.

Kelly, D., & Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '04)*, Sheffield, United Kingdom, 377-384.

Kelly, D. & Fu, X. (2006). Elicitation of Term Relevance Feedback: An Investigation of Term Source and Context. In *Proceedings of the 29th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '06)*, Seattle, WA, 453-460.

Kelly, D., Harper, D. J., & Landau, B. (2008). Questionnaire mode effects in interactive information retrieval experiments. *Information Processing & Management, 44*(1), 122-141.

Kelly, D. & Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum, 37*(2), 18-28.

Kim, J., Oard, D. W., & Romanik, K. (2000). User modeling for information access based on implicit feedback. Technical Report: HCIL-TR-2000-11/UMIACS-TR-2000-29/CS-TR-4136, University of Maryland, College Park. Retrieved March 25, 2007, from ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2000-11html/2000-11.html

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet news. *Communication of the ACM, 40*(3), 77-87.

Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Ablex.

Kurth M. (1993). The limits and limitations of transaction log analysis. *Library Hi Tech, 11*(2), 98-104.

Lazonder, A., Biemans, H., & Wopereis, I. (2000). Differences between novice and experiences users in searching information on the World Wide Web. *Journal of the American Society for Information Science, 51*(6), 576-581.

Li D., Babcock, J. & Parkhurst, D. (2006). OpenEyes: a low-cost head-mounted eye-tracking solution. In *Proceedings of the Eye Tracking Research & Application Symposium,* San Diego, California, USA, 95-100.

Liberman, V. & Tversky, A. (1993). On the evaluation of probability judgments: calibration, resolution, and monotonicity. *Psychological Bulletin, 114*(1), 162-173.

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40-54.

Lucas, W., & Topi, H. (2002). Form and function: The impact of query term and operator usage on web search results. *Journal of the American Society for Information Science, 53*(2), 95-108.

Lv, Y., Sun, L., Zhang, J., Nie, J., Chen, W., & Zhang, W. (2006). An iterative implicit feedback approach to personalized search. In *Proceedings of the 21st international Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, Sydney, Australia, 585-592.

Maglio, P. P., Barrett, R., Campbell, C. S., & Selker, T. (2000). SUITOR: an attentive information system. In *Proceedings of the 5th International Conference on Intelligent User Interfaces (IUI'00)*, New Orleans, LA, United States, 169-176.

Maglio, P. P., & Campbell, C. S. (2003). Attentive agents. *Communications of the ACM, 46*(3), 47–51.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science, 29*(3), 165–176.

Marchionini, G. (1995). Information seeking in electronic environments. New York, NY: Cambridge University Press.

Marchionini G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM, 49*(4), 41-46.

Marchionini, G., & Mu, X. (2003). User studies informing E-Table interfaces. *Information Processing & Management, 39*(4), 561-579.

Marshall, C.C. and Bly, S. (2005). Turning the Page on Navigation. In *Proceedings of JCDL '05*, Denver, CO, 225-234.

Mitra. M., Singhal, A., & Buckley, C. (1998). Improving Automatic Query Expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 206-214.

Monk A., Wright P., Haver J., & Davenport L. (1993). *Improving your human-computer interface: A practical technique.* New York: Prentice-Hall.

Morita, M., & Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 272-281.

Mueller, F., & Lockerd, A. (2001). Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *CHI '01 Extended Abstracts,* 279-280.

Nichols, D. M. (1997) Implicit ratings and filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapaest, Hungary 10-12, ERCIM. http://www.ercim.org/publication/ws-proceedings/DELOS5/index.html

Nielsen, J. (1993). *Usability engineering*. Boston, MA: AP Professional.

Oard, D. W., & Kim, J. (2001). Modeling information content using observable behavior. In *Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology,* USA, 38-45.

Oh, K.-T., & Lee, K.-P. (2005). Real-time Logging Method for Interaction Observation - with emphasis on the software requirement. In *Proceedings of the 11th HCI International Conference*, Las Vegas, Nevada, USA.

O'Keefe, K. (2000). The quality of medical students' confidence judgments when using external information resources: the effects of different media formats, source of questions, and question formats. Unpublished Ph.D. dissertation. The University of North Carolina, Chapel Hill, NC.

Pennanen, M., & Vakkari, P. (2003). Students' conceptual structure, search process, and outcome while preparing a research proposal: A longitudinal case study. *Journal of the American Society for Information Science & Technology, 54*(8), 759-770.

Penzo M. (2005). Introduction to eyetracking: seeing through your users' eyes. Retrieved February 22, 2007, from http://www.uxmatters.com/MT/archives/000040.php

Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech, 11*(2), 41-66.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3-25.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T. (1994). *Human-Computer Interaction.* Harlow, England: Addison-Wesley.

Puolämaki, Salojärvi, Savia, Simola, & Kaski (2005). Combining eye movements and collaborative filtering for proactive information retrieval. In *Proceedings of SIGIR 2005,* Salvador, Brazil, 146-153.

Rafter, R., & Smyth, B. (2001). Passive profiling from server logs in an online recruitment environment. In *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP 2001)*, USA, 35-41. Retrieved March 26, 2007, from http://www.changingworlds.com/content/documents/rafterijcai.pdf

Rappoport, A. (2003). Search Query Spellchecking. Retrieved April 25, 2005, from: http://home.earthlink.net/~searchworkshop/docs/avi-chi-search-spelling.doc

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372-422.

Rele. R., & Duchowski A. (2005). Using eye-tracking to evaluate alternative search results interfaces. In *Proceedings of the Human Factors and Ergonomics Society 2005 Annual Meetiing,* Orlando, FL.

Rodden, K. & Fu, X. (2007). Exploring How Mouse Movements Relate to Eye Movements on Web Search Results Pages. In *Proceedings of SIGIR 2007 Workshop on Web Information-Seeking and Interaction*, Amsterdam, the Netherlands, 29-32.

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual ACM International Conference on Research and Development in Information Retrieval (SIGIR '03)*, Toronto, CA, 213-220.

Salojarvi, J., Kojo, I., Simola, J., & Kaski, S. (2003). Can relevance be inferred from eye movements in information retrieval? In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*, Hibikino, Kitakyushu, Japan, 261-266.

Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288-297.

Sandore, B. (1993). Applying the results of transaction log analysis: Transaction log analysis. *Library Hi Tech, 11*(2), 87-97.

Saracevic, T. (1996). Interactive models in information retrieval: A review and proposal. In *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, *33,* 3-9.

Saracevic, T. (1997). Extensions and application of the stratified model of information retrieval interaction. In *Proceedings of the Annual Meeting of the American Society for Information Science*, *34,* 314-327.

Schaefer A., Jordan M., Klas C., & Fuhr N. (2005). Active Support for Query Formulation in Virtual Digital Libraries: A case study with DAFFODIL. In Rauber, A., Christodoulakis, S., & Tjoa, A. M. (Ed.), *Research and Advanced Technology for Digital Libraries: Proceedings of the 9th European Conference, ECDL 2005* (Vienna, Austria, September 18-23, 2005). 414-425. Retrieved March 31, 2007 from http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Schaefer_etal:05.pdf.

Schiessl, M., Duda, S., Thölke, A., & Fischer, R. (2003). Eye-tracking and its application in usability and media research. *MMI interaktiv,* 6. Retrieved February 20, 2007, from http://useworld.net/ausgaben/3-2003/MMI-Interaktiv0303_SchiesslDudaThoelkeFischer.pdf

Schroeder, W. (1998). Testing Web sites with eye-tracking. *User Interface Engineering.* Retrieved March 20, 2007, from http://www.uie.com/articles/eye_tracking/

Shen, X., & Zhai, C. (2004). Active feedback - UIUC TREC-2003 HARD experiments. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, Washington, DC. U.S. Government Printing Office. NIST Special Publication 500-255.

Shen, X., & Zhai, C. (2005). Active feedback in ad hoc information retrieval. In *Proceedings of 2005 ACM Conference on Research and Development on Information Retrieval (SIGIR '05)*, 59-66.

Shen, X., Tan, B., & Zhai, C. (2005a). Implicit user modeling for personalized search. In *Proceedings of the 14th ACM international Conference on information and Knowledge Management, CIKM 2005* (Bremen, Germany, October 31 - November 05, 2005). New York: ACM Press. 824-831.

Shen, X., Tan, B., & Zhai, C. (2005b).UCAIR: Capturing and exploiting context for personalized search. In *Proceedings of 2005 ACM Conference on Research and Development on Information Retrieval - Information Retrieval in Context Workshop (IRiX'2005)*.

Shen, X., Tan, B., & Zhai, C. (2005c). Context-sensitive information retrieval using implicit feedback. In *Proceedings of 2005 ACM Conference on Research and Development on Information Retrieval (SIGIR '05)*, 43-50.

Shneiderman, B., & Plaisant, C. (2004). Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th edition). Reading, MA: Addison-Wesley.

Sibert, L., & Jacob R. (2000). Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, Hague, the Netherlands, 281-288.

Smyth, B., Balfe, E., Freyne, J., Briggs, P., Coyle, M., & Boydell, O. (2004). Exploiting query repetition and regularity in an adaptive community-based Web search Engine. *User Modeling and User-Adapted Interaction, 14*(5), 382-423.

Spink, A., & Jansen, B. J. (2005). *Web search: public searching of the web.* New York: Kluwer.

Stevens, C. (1993). *Knowledge-based assistance for accessing large, poorly structured information spaces.* Unpublished Ph.D. dissertation, University of Colorado, Boulder, CO. Retrieved March 22, 2007, from http://www.holodeck.com/curt/mypapers/Thesis.pdf

Rucker, J., & Polanco, M. J. (1997). Siteseer: personalized navigation for the Web *Communications of the ACM, 40*(3): 73-76.

Ruthven, I. & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review, 18*(2), 95-145.

Tanin, E., Lotem, A., Haddadin, I., Shneiderman, B., Plaisant, C., & Slaughter, L. (2000). Facilitating Network Data Exploration with Query Previews: A Study of User Performance and Preference. *Behaviour & Information Technology 19*(6), 393-403.

Tauscher, L., & Greenberg, S. (1997). Revisitation patterns in World Wide Web navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*, Atlanta, Georgia, USA, 399-406.

Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College & Research Libraries, 29*(3), 178-194.

Teevan, J., Dumais, S. T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual international*

ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '05), Salvador, Brazil, 449-456.

Toms E., & Bartlett J. (2001). An Approach to Search for the Digital Library. In *Proceedings of JCDL 2001* (Roanoke, Virginia, USA, June 24-28, 2001). 341-342.

Vakkari, P. (2000). Cognition and changes of search terms and tactics during task performance: A longitudinal case study. *Proceedings of the RIAO 2000 Conference*, 894-907.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology, 37,* 413-464.

Volokh, E. (2000). Personalization and privacy. *Communications of the ACM, 43*(8), 84–88.

Voorhees, E. M. (2005). Overview of the TREC 2004 robust retrieval track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD, 70-79.

Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal web queries: trends and patterns. *Journal of the American Society for Information Science and Technology, 54*(8), 743–758.

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. *Information Processing & Management, 36,* 229-251.

Wang, P., & Pouchard, L. (1997). End-user searching of Web resources: Problems and implications. In *Advances in Classification Research: Proceedings of the 8th ASIS SIG/CR Classification Research Workshop* (November 2, 1996, Washington, DC). 73-85.

White, R. W. (2004). *Implicit Feedback for Interactive Information Retrieval.* Unpublished Ph.D. dissertation. University of Glasgow, Glasgow, U.K.

White, R. W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *Proceedings of the 15th ACM international Conference on information and Knowledge Management (CIKM '06)*, Arlington, Virginia, USA, 297-306.

White, R. W., Kules, B., Drucker S., & Schraefel, M. (2006). Introduction (Supporting Exploratory Search: A Special Section of the Communications of the ACM). *Communications of the ACM, 49*(4), 36-39.

White, R. W., & Marchionini, G. (2007). Examining the effectiveness of real-time Query Expansion. *Information Processing and Management, 43*(3), 685-704.

White, R. W., Ruthven, I., & Jose, J. M. (2002a). The use of implicit evidence for relevance feedback in web retrieval. In *Proceedings of 24th BCS-IRSG European Colloquium on IR Research, Lecture notes in Computer Science 2291*, 93-109.

White, R. W., Ruthven, I., & Jose, J. M. (2002b). Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*, Finland, 57-64.

White, R. W., Ruthven, I., & Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '05)*, Salvador, Brazil, 35-42.

Wildemuth, B. M. (2002). Introduction and overview: effective methods for studying information seeking and use. *Journal of the American Society for Information Science and Technology, 53*(14): 1218-1222.

Wolf, C. G., Carroll, J. M., Landauer, T. K., John, B. E., & Whiteside, J. (1989). The role of laboratory experiments in HCI: help, hindrance, or ho-hum?. In *Proceedings of SIGCHI '89*, 265-268.

Zamir, O. & Etzioni, O. (1999). Grouper: A dynamic clustering interface to web search results. In *Proceedings of the Eighth International World Wide Web Conference (WWW '99)*, Toronto, Canada, 8.

Zhang, B., & Seo, Y. (2001). Personalized web-document filtering using reinforcement learning. *Applied Artificial Intelligence, 15*(7), 665-685.

Zhang, Y., & Callan, J. (2005). Combining multiple forms of evidence while filtering. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, British Columbia, Canada, 587-595.