Supporting Finding and Re-Finding Through Personalization

by

Jaime Teevan

S.M. Computer Science and Engineering, Massachusetts Institute of Technology (2001)
B.S. Computer Science, Yale University (1998)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2007

Author ……...
Department of Electrical Engineering and Computer Science
October 13, 2006

Certified by ……...
David R. Karger
Professor of Computer Science and Engineering
Thesis Supervisor

Accepted by ……..
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Supporting Finding and Re-Finding Through Personalization

by

Jaime Teevan

Submitted to the Department of Electrical Engineering and Computer Science
on October 13, 2006, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

Although one of the most common uses for the Internet to search for information, Web search tools often fail to connect people with what they are looking for. This is because search tools are designed to satisfy people in general, not the searcher in particular. Different individuals with different information needs often type the same search terms into a search box and expect different results. For example, the query "breast cancer" may be used by a student to find information on the disease for a fifth grade science report, and by a cancer patient to find treatment options.

This thesis explores how Web search personalization can help individuals take advantage of their unique past information interactions when searching. Several studies of search behavior are presented and used to inform the design of a personalized search system that significantly improves result quality. Without requiring any extra effort from the user, the system is able to return simple breast cancer tutorials for the fifth grader's "breast cancer" query, and lists of treatment options for the patient's.

While personalization can help identify relevant new information, new information can create problems re-finding when presented in a way that does not account for previous information interactions. Consider the cancer patient who repeats a search for breast cancer treatments: she may want to learn about new treatments while reviewing the information she found earlier about her current treatment. To not interfere with re-finding, repeat search results should be personalized not by ranking the most relevant results first, but rather by ranking them where the user most expects them to be.

This thesis presents a model of what people remember about search results, and shows that it is possible to invisibly merge new information into previously viewed search result lists where information has been forgotten. Personalizing repeat search results in this way enables people to effectively find both new and old information using the same search result list.

Thesis Supervisor: David R. Karger
Title: Professor of Computer Science and Engineering

Thomas Escher

To my father.
James Ripley Teevan
1945 - 2002

Appreciation is a wonderful thing: It makes what is excellent in others belong to us as well.

- Voltaire (1694 - 1778)

Acknowledgements

I would like to acknowledge many people for their support during my doctoral work. I would especially like to thank my advisor, David R. Karger, for his willingness to support my explorations into the evolving research area of personal information management. He consistently provided me with an interesting perspective and insight. I am also grateful to an exceptional doctoral committee, and wish to thank Mark S. Ackerman, Susan T. Dumais and Robert C. Miller for their support and encouragement.

Much of the research in this thesis is the result of collaboration with a number of phenomenal researchers. Christine Alvarado, Mark S. Ackerman and David R. Karger worked with me on the diary study presented in Chapters 3 and 4. Most of the work on personalized search presented in Chapters 4 and 5 was done with Susan T. Dumais and Eric Horvitz while interning at Microsoft Research. The log analysis presented in Chapters 7 and 8 that motivates the Re:Search Engine was done with Eytan Adar, Rosie Jones and Michael Potts. I am grateful to Yahoo for making this analysis possible.

I have also enjoyed collaborating with Diane Kelly on a review of implicit measures in information retrieval, as well as later on evaluation methodology for personal information management (PIM), and with Rob Capra and Manuel Pérez-Quiñones on a review of finding and re-finding literature. These works helped shaped my discussion of related work in Chapters 2 and 6.

A large number of additional people, such as Nick Belkin, William Jones, Ben Bederson, Anita Komlodi, Mary Czerwinski and Ed Cutrell, all provided valuable feedback on my thesis research at some point. Randy Davis encouraged me to keep the big picture in mind as I dove into the details, and Marti Hearst provided valuable suggestions for evaluation. I enjoyed brainstorming about the Re:Search Engine with Stephen Intille, and the paper prototype study described in Chapter 9 was performed in a class taught by him.

My good friend Michael Oltmans has provided, in addition to moral support, valuable technical support and babysitting services. He, Tracy Hammond and Christine Alvarado were always willing pilots for studies and willing sounding boards for difficult ideas, and I am grateful for that, as well as for their friendship. Janina Matuszeski was a big help with some of the statistical analysis presented in Chapter 9.

I appreciate the feedback and advice I have received from members of the Haystack group over the years, including my fabulous officemate Kai Shih, Nick Matsakis, David Huynh, Karun Bakashi, Vineet Sinha, Har Chen, Yuan Shen and Steve Garland. I am also grateful to the hundreds of people who participated in the studies presented here, including my friends from BabyCenter and Pappy.

This thesis is dedicated to my father, Jim Teevan, who always enjoyed helping me figure things out. My family is important to me beyond words, and I owe much to him, my

mother Connie Teevan, and my siblings Conor and Brooks Teevan. My husband, Alex Hehmeyer, has been a fabulous support throughout my graduate school career, and I could not have asked for a better excuse to take breaks during crunches than to play with my son, Griffin Hehmeyer. Thank you also to Cale and Dillon Teevan for keeping me company as I finished my research and wrote my thesis. It is great to finally meet you both.

Contents

List of Figures

List of Tables

A man travels the world over in search of what he needs and returns home to find it.

- George Moore (1874 - 1958)

Chapter 1

Introduction

This thesis explores Web search personalization as a means of helping people find what they are looking for faster and more easily than they would with a generic search system. The studies presented in Part I of the thesis address the finding of information. They demonstrate the importance of personalization by highlighting the different ways people find information and judge its relevance. Based on these studies, a personalized search system is developed and tested that significantly improves result quality by personalizing the ranking of results using a rich user profile.

However, while returning the personalized results that are the most relevant to a user's query at a given time can help the user find new information faster, it can exacerbate problems with re-finding previously encountered information because personalization increases the rate of change to search result lists. Part II of this thesis addresses the re-finding of previously viewed information. Several studies are presented that explore re-finding behavior, demonstrating the prevalence of re-finding behavior and the importance of past experience when re-finding. From these studies, a search system is developed that personalizes search results not by ranking the most relevant results first, but rather by ranking them where the user expects to find them. Such a system is shown to help people re-find previously viewed information as quickly as they would with no new information present, and simultaneously to find new information as quickly as if no adjustments were being made to support re-finding.

The purpose of this chapter is to introduce and motivate the research discussed in this thesis. The chapter begins with an example that makes concrete the issues and problems explored in the thesis. It then gives an overview of the research approach taken and the results to be presented. Finally the thesis outline is presented, and the contributions made in this work are highlighted.

1.1 A Motivating Example

Consider, as an example of the importance of personalization in finding and re-finding, Connie's search for breast cancer treatments. Connie was recently diagnosed with breast cancer. When first diagnosed, she wanted to find a list of breast cancer treatments in

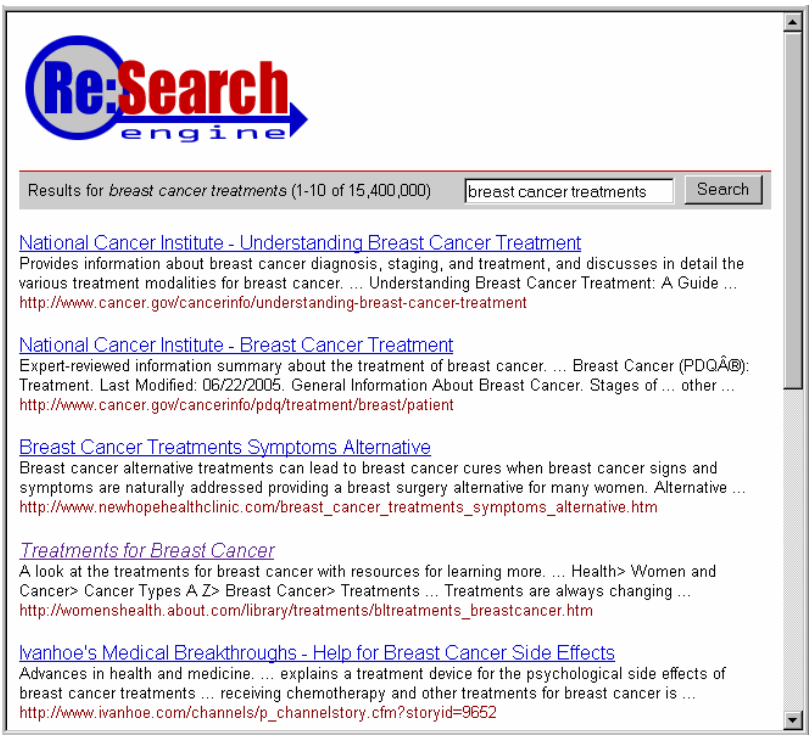


Figure 1-1. Generic search results for Connie's query for "breast cancer treatments". The result that she clicked on is shown in *italics* for emphasis, but is not italicized by the Re:Search Engine.

order to learn more about the medical alternatives suggested by her doctor. For this reason, she ran a Web search for "breast cancer treatments". The result list returned to her for this search is shown in Figure 1-1. Several results from the National Cancer Institute are listed first, followed by a result about alternative treatments, a link to About.com's page on treatments for breast cancer, and so on. The government pages appeared too technical to interest Connie, and she is not generally interested in learning about alternative treatments, so she skipped over the first couple of results in the list. She decided to follow the fourth link, and found the list of treatments she was looking for there.

As time passed, the chemotherapy Connie was taking stopped working effectively, and she and her doctors decided it was time for her to change approaches. Once again, she ran a search for "breast cancer treatments" to review the list of treatments she found before when originally making her decision, and perhaps learn about newly available treatments. The results returned for this subsequent search are shown in Figure 1-2.

The new result list is better and more relevant to Connie's query than the previous result list seen in Figure 1-1 because when Connie ran her second search, the search engine personalized the result list ordering. The personalization is based on her past interactions with the system and other available information about her, such as the Web pages she has visited, the documents she has authored, and the emails she has received. The result she clicked on initially is bumped to the top of the list, since she is likely to find it relevant to this query given that she did last time. There are also a number of potentially valuable new results in the new list. For example, another result from About.com on breast cancer

Figure 1-2. The personalized results Connie received when she repeated her search for "breast cancer treatments". Results she's clicked before have moved up in rank, and results that are likely to interest her are highly ranked. Sites for alternative treatments are no longer listed.

is listed highly, and there is a result about a new hormone therapy treatment for breast cancer that appears promising. Gone are results that were less valuable to Connie, such as a link to a page with information about alternative breast cancer treatments. All of this was done without Connie having to explicitly describe what she was looking for in great detail (e.g., "The list of breast cancer treatments I found before, as well as more detailed information about particular medical treatments. I like results with general and easily accessible information about conventional treatments, but I'm not interested in alternative medicine, with the exception of diet-related information and …").

The chapters in Part I of this dissertation focus on understanding how people like Connie find information and supporting the behavior simply and intuitively. An investigation into they ways different individuals' search strategies vary leads to the development of a system search that personalizes results in the manner seen in Figure 1-2.

However, although such personalization clearly provides some benefit to Connie, naively personalizing the result list to rank those results that were most relevant to Connie's repeat query first is not necessarily the best way to help her find what she was looking for. While it does help her find new information (e.g., information on hormone therapy and About.com's Breast Cancer site), it can interfere with her ability to re-find information she has seen before. Connie developed expectations about what results the search result list for "breast cancer treatments" contains during her first search. Placing the About.com result first because she liked it before did not necessarily make it easier

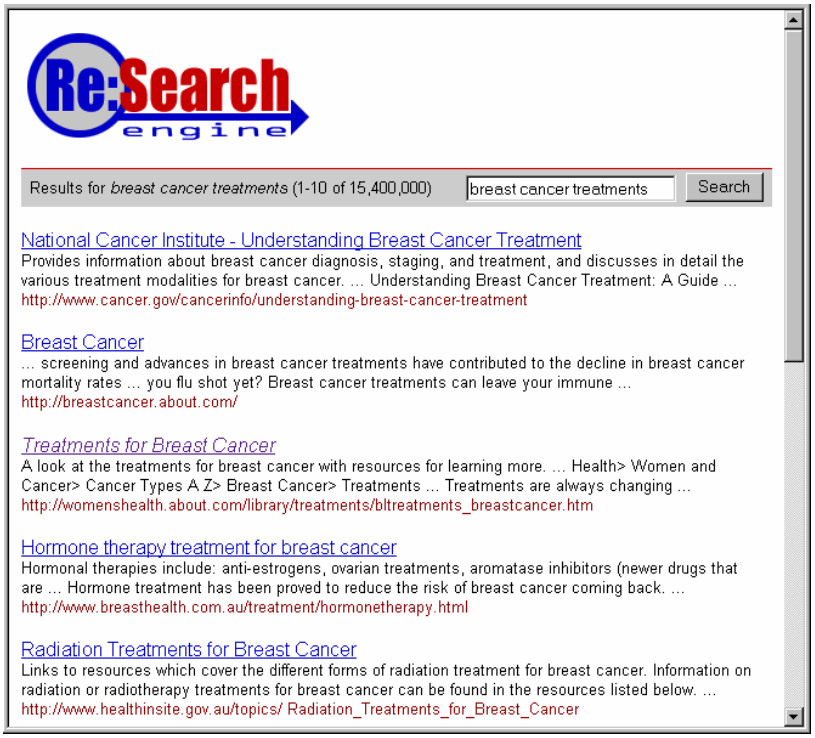


Figure 1-3. A search result list that contains information Connie has seen before where she expects it, while still including the new personalized results shown in Figure 1-2 that she has not seen before but that might be useful.

for her to find it again. She expected the result be a ways down the list, and research in this thesis suggests she is thus likely to have looked for it there first when she repeated her search. Returning the same result list Connie saw initially (shown in Figure 1-1) would have matched her expectations and helped her re-find previously viewed resources like the About.com site. But doing so would also have caused her to miss valuable new – and potentially life-saving – information like the result on hormone therapy.

Part II of this thesis focuses on understanding and supporting re-finding in a way that does not interfere with the finding of new information. The chapters in Part II show that a good personalized search system can support re-finding by placing the results a person remembers where that person expects to find them, while still supporting finding by including new results in locations where the user has not already developed expectations.

An example result list that does both of these things for Connie can be seen in Figure 1-3. The first result from the initial list is preserved, as it is likely she remembered that result even though she did not click on it. Similarly, the result she clicked on is preserved at about the same rank it appeared when she initially ran the query. On the other hand, the second result and the bottom two results point to new information that may be of greater use to her than what was previously there. Unless the results from Figure 1-1 are compared side by side with the results in Figure 1-3, the two lists are likely to look the same despite major differences. Personalizing results in this manner can help Connie find new information quickly while not interfering with her ability to re-find previously viewed information.

1.2 Approach Overview

In the previous example, Connie needed to make an important medical decision based on the information available to her. Like Connie, people regularly need to make important decisions with the support of the hundreds of emails they receive daily, the thousands of documents they store on their personal computers, and the billions of Web pages to which they have access, and large private databases like corporate intranets and digital libraries. The research presented in this thesis aims to enable people to take advantage of all of this information, rather than be overwhelmed by it. To effectively support the complex information interactions people perform daily, it is necessary to understand and model what people really do, and build tools accordingly. To do this, the research focuses on answering the following three questions:

- **How do people search for information?** What strategies do people currently use to find new information? To re-find previously viewed information? How do interactions vary across information type, from email and Web pages to local files and even paper? How do interactions vary across individuals?
- **What tools will best support realistic search interactions?** How should next-generation search tools be built? How can individual behaviors be supported? What new approaches are needed for people to efficiently re-find information?
- **Once built, how effective are the tools?** What testing framework should be used to evaluate them? How do the tools affect a person's information management in the long term?

The research done for as part of this thesis illustrates how answering these questions can lead not only to the discovery of creative solutions to existing problems (e.g., better personalization algorithms), but also to the uncovering of unexpected and important personal information management problems (e.g., problems with re-finding exacerbated by, and ultimately also solved by, personalization).

1.3 Results Overview

This section reviews the results reported in this thesis. The thesis begins with a discussion of finding behavior. A naturalistic study of people performing personally motivated searches suggests that people use contextual knowledge to navigate (or *orienteer*) to their target with small, local steps – even when they know exactly what they are looking for in advance and could jump to their target directly (or *teleport*) using a tool like a keyword search engine. The study reveals several benefits to orienteering over teleporting, including that it is easier to recognize the information being sought than to recall how to uniquely specify it.

While participants did not generally try to give full descriptions of what they were looking for to a keyword search engine, they did commonly use keyword search as a step in the orienteering process. For example, the keyword search Connie performed above for "breast cancer treatments" was only part of a greater search to learn more about two treatment options offered by her doctor. Queries referring to intermediate orienteering

steps are common probably because it is easy to generate a simple query to get partway to the target, but hard to fully name the target up front. The list of treatments Connie found, for example, enabled her to correctly spell the chemotherapy regimes her doctor mentioned and perform additional searches on the two drugs. It would have been difficult for her to generate a query with the two drug names merely from memory.

It is well known that simple queries can be ambiguous. A person searching for "cancer" might be looking for information about the disease, like Connie, or the astrological sign. However, a study of what different people consider relevant to the same query shows that even when people used unambiguous terms (e.g., "malignancy" instead of "cancer"), their information needs often remain underspecified. A cancer researcher, for example, may use the same terms to search for recent journal articles that Connie used to find treatment options.

Because people have difficulty explicitly specifying their search target, the use of implicit information about a user's interests and intent to re-rank Web search results is explored. All of the electronic information a person has seen – including email, files and Web pages – can be used to create a user profile that is incorporated into the search result ranking algorithm using a relevance feedback framework. These algorithms are tested using a test bed of 131 queries with associated personalized ratings. The tests reveal that the personalization algorithms can significantly improved upon current Web search when based on rich user profiles.

Even though personalizing search can lead to objectively better results, the potential volatility of the rankings can make the common behavior of revisiting a result set by repeating a query a challenge. A large scale analysis of re-finding in Yahoo's query logs reveals that when a person repeats a query, it is harder to click on a previously viewed result if the result has changed rank than if it has remained in the same place. Connie, for example, may re-click on the About.com site she found more easily if it remains fourth in the list than if it is listed first. Another study comparing search result quality shows an objectively better result list can appear worse if it appears to have changed from when it was first seen. For example, Connie may think the better personalized result list returned on her subsequent search (Figure 1-2) is worse just because it is different.

Because people value consistency, one way to handle the volatility of personalized search results is to cache all of the results a person sees and always return the cached results when a query is later repeated. Several problems exist with this solution; although it provides consistency, users' needs may evolve and they risk missing important new information. This was seen in the above example where Connie performed a search for treatment options shortly after being diagnosed. When she wanted to revisit her treatment options, it would be frustrating for her not to find information that was initially found useful, but it would similarly be frustrating to miss learning about new treatments.

Instead, the research presented in this thesis reveals that changes to a result list can be made in a way that does not disrupt the user's interaction with previously viewed results by taking advantage of the fact that people have limited memories. Rather than keeping the entire search result list static when a person re-issues a query, only that information that the user remembers need be kept static. New information can be snuck into the holes where results were forgotten. If Connie only remembers the first and fourth result

returned for "breast cancer treatments", the second and third results can be changed to include new information when the query is repeated.

To test this approach, a study was conducted of what 119 people found memorable about search result lists. The results of this study were used to develop a model of which aspects of a search result list are memorable, and thus should be changed with care, and which are not, and can change freely. A follow-up study was then conducted with 165 different people that tested the model by asking participants whether they noticed changes that either occurred in accordance with the model or not. When changes were made according to the model, the result list looked static to participants, while changes made naively were commonly noticed and thought to hurt search result quality. A subsequent study of 30 people performing re-finding and new-finding tasks with result lists that had changed in various ways revealed that people were able to re-find more easily and faster when they did not notice a change to the result list. However, they were able to find new information with the same success as if past interactions were not accounted for.

1.4 Thesis Outline

The research in this thesis is presented in two parts. The first, Part I (Finding), focuses on understanding and supporting the finding information. The second, Part II (Re-Finding), focuses on understanding and supporting re-finding. Each part follows a similar structure and consists of four chapters:

i. **Introduction to (Re-)Finding**
ii. **Understanding (Re-)Finding**
iii. **Why (Re-)Finding Requires Personalization**
iv. **Supporting Finding (and Re-Finding) through Personalization**

The first chapter (*i*) in each part is intended to give an introduction of the area and provide a summary of related research necessary to understand the rest of the part.

The second chapter (*ii*) in each part presents studies conducted as part of this thesis research that give a better understand the behavior. In Part I, finding behavior is explored through an observational study of people searching for information, and Part II, re-finding behavior is explored through log analysis of repeat queries.

The third chapter (*iii*) in each part presents studies that show personalization is important to support the behavior. Personalization is important for finding because individuals have different finding strategies (as demonstrated through an observational study) and use the same queries to find different information (as demonstrated through a study of what results people consider relevant to the same query). Personalization is important for re-finding because individuals develop expectations about information based on previous interactions with information.

The final chapter (*iv*) of each part presents, based on the previous three chapters, a personalized search system designed to support, in the case of Part I, finding, and in the

case of Part II, re-finding. In Part I, search results are personalized to rank the most relevant result for an individual first based on a rich profile of that individual, and this improves the quality of the ranking for new information. In Part II, search results are personalized based on the individual's previous interactions with the search results, aiming to rank information where the searcher expects to find it. Such a ranking is shown to help people re-find previously viewed information quickly while not interfering with the finding of new information.

Following Part II, **Chapter 10** concludes the thesis. It highlights the contributions of this work and discusses a number of interesting implications and related problems that arise in relation to this work for future consideration.

1.5 Contributions

This work explores personalized Web-based information retrieval. It makes five important contributions to the areas of information retrieval and human computer interaction. These contributions are presented in the order they are discussed in this dissertation.

- First, this thesis shows that for directed search people prefer to navigate to their target with small, local steps using their contextual knowledge as a guide (*orienteer*) rather than to jump directly to their target using keywords (*teleport*). It also gives insight into the differences in search behavior between individuals, suggesting that people use different step sizes while orienteering and that people have very different notions of what is relevant to even fairly unambiguous queries.
- Second, this thesis presents a novel application of relevance feedback to the realm of implicit search result personalization. It demonstrates that implicit relevance feedback based on large amounts of information about the searcher can significantly improve search result quality.
- Third, this thesis offers evidence that repeat Web search queries are extremely prevalent. It gives insight into common features of repeat searches and shows that people often search for new information using the same query results they use to find previously viewed information.
- Fourth, this thesis presents a model of what people remember about search result lists based on a large-scale study of human memory. The study shows that what people remember about a result is a function of their interaction with the item and the item's location in the list.
- Fifth, this thesis presents an algorithm that uses the model of what people remember about search result lists to invisibly merge new information into a previously viewed search result list. Doing this allows people to effectively find both new and old information using the same search result list.

PART I

FINDING

To find a fault is easy; to do better may be difficult.

- Plutarch (46 - 120)

Chapter 2

Introduction to Finding

Finding information is one of the most basic tasks of information management, and is a common activity both on the Web and on the desktop. A recent Pew Internet and American Life report showed that Internet searches are a top Internet activity, second only to email (Rainie & Shermak, 2005). Nonetheless, searchers are unable to find what they are looking for over 50% of the time, and knowledge workers are estimated to waste 15% of their time because they cannot find information that already exists (Feldman, 2004).

The research in this thesis works to close this gap by better understanding how people search and building systems that address the problems they encounter. The chapters in Part I address finding in general, while the chapters in Part II focus specifically on the re-finding of information. This initial chapter is intended to place the finding research of Part I in context. The chapter begins with a few basic definitions and a brief overview of the other chapters in this part. It then presents related research that has been conducted to understand finding behavior and support personalization, and discusses the various study methodologies employed throughout the thesis.

2.1 Basic Definitions

Finding is part of a large and active area of study on information seeking.

> **Information seeking** (or just **seeking**). *The purposive seeking of information as a consequence of a need to satisfy some goal. In the course of finding, the individual may interact with manual information systems (such as a newspaper or a library), or with computer-based systems (such as the World Wide Web). (Wilson, 2000).*

Information seeking includes all activities directed towards accessing information to meet an information need. The need can be very specific, like a list of breast cancer treatments, or something broad and less defined, like learning about the disease in general. Finding activities include the specification of a focused query to a search service

(e.g., typing "breast cancer treatments" into a search engine's query box) as well as less directed browsing.

"Information seeking" is often used interchangeably with "finding". Finding is used in this thesis in preference to seeking because, as outlined by Jones (in press), finding tends to involve limited, closed actions (the finding of a list of treatments), as opposed to the more open-ended activity of seeking (the seeking of information related to breast cancer). Such directed tasks fall within a narrow slice of most existing information seeking and searching models.

Because the finding of previously viewed information is an incredibly important piece of finding, re-finding is defined here and is discussed in detail in Part II of this thesis:

> **Re-finding**. *Re-finding is the process of finding information that has been seen before.*

The finding of new information and the re-finding of previously viewed information are often interleaved as part of a larger information-seeking task. For example, Connie may want to both find information about new breast cancer treatments and re-find the list she found initially in the process of deciding the next step in her treatment.

Note that the focus of re-finding is not on repetition of a previous finding activity, but rather on the retrieval of information previously experienced. The process of re-finding may actually be very different from the process of finding new information. This is because as people remember experiencing the information before, they may think of different ways to find the information again. For example, in Chapter 1 Connie returned to the About.com Web she had seen before site by repeating her initial search, but she might have chosen to re-find the site by browsing to it, using information she remembered about it from her original encounter to guide her search. And if Connie didn't remember ever searching for treatment options, the process of re-finding the site would look like a new search entirely. Re-finding differs from finding when the seeker takes advantage of knowledge remembered from the initial encounter, and how this difference affects re-finding is highlighted in Part II.

2.2 Outline of Part I

Including this chapter, Part I consists of four chapters:

2. **Introduction to Finding**
3. **Understanding Finding**
4. **Why Finding Requires Personalization**
5. **Supporting Finding through Personalization**

The research presented in **Chapter 3 (Understanding Finding)** was conducted to develop a rich understanding of finding behavior. The chapter presents a modified diary study that investigated how people performed personally motivated searches in their email, in their files, and on the Web. The study revealed that instead of jumping directly to their information target using keywords, participants navigated to their target with

small, local steps using their contextual knowledge as a guide, even when they knew exactly what they were looking for in advance. The observed advantages of searching by taking small steps include that it allowed users to specify less of their information need and provided a context in which to understand their results. The personalized search system presented in Chapter 5 exploits these advantages by implicitly personalizing search results and by placing an emphasis on result context.

In **Chapter 4 (Why Finding Requires Personalization)** evidence is given to suggest personalization can improve systems that support finding. The chapter focuses on how individual differences affect finding behavior. The data collected as part of the diary study introduced in Chapter 3 is analyzed to reveal that multi-stepped search behavior was especially common for participants with unstructured information organization (*pilers*, as opposed to *filers*). The chapter then dives deeper into one common step people take when searching for information: keyword search. The diverse goals people have when they issue the same query to a Web search engine are investigated, as is the ability of current search tools to address such diversity, in order to understand the potential value of personalizing search results. Great variance was found in the results different individuals rated as relevant for the same query – even when those individuals expressed their underlying informational goal in the same way. The analysis suggests that while current Web search tools do a good job of retrieving results to satisfy the range of intentions people may associate with a query, they do not do a very good job of discerning an individual's unique search goal.

In light of the findings of these two chapters, in **Chapter 5 (Supporting Finding through Personalization)** personalized search algorithms are formulated and studied. Rather than relying on the unrealistic assumption that people will precisely specify their intent when searching, techniques are pursued that leverage implicit information about the user's interests. The approaches explored consider a user's prior interactions with a wide variety of content to personalize that user's current Web search. This information is used to re-rank Web search results within a relevance feedback framework. Rich models of user interests are explored, built from both search-related information, such as previously issued queries and previously visited Web pages, and other information about the user such as documents and email the user has read and created. The research suggests that rich representations of the user and the corpus are important for personalization, but that it is possible to approximate these representations and provide efficient client-side algorithms for personalizing search. Such personalization algorithms are shown to significantly improve on current Web search.

The rest of this chapter highlights the related work necessary to understand the research described above and presented in greater detail in the following chapters of Part I. First related research that has been conducted to understand finding behavior is discussed, followed by a discussion of related search personalization systems, and a brief overview of study methodology.

2.3 Related Work on Understanding Finding

Information science (IS) research has identified factors that influence people's information seeking and searching behaviors, and modeled such behavior to help guide the design of information seeking and information retrieval systems (Wilson, 1999). However, the small slice of information seeking research that focuses on directed finding is typically a specialized form of information seeking with characteristics that are not yet fully understood. This section shows that like information seeking, finding is a multi-stepped process, and discusses a variety of factors that affect finding, including the information being sought, the task context of the search, and the person doing the finding. Much of this section is based on a review of finding literature conducted by Robert Capra, Manuel Perez-Quinones and me (Teevan, Capra, & Pérez-Quiñones, in press).

2.3.1 Finding is a Multi-Stepped Process

Information seeking is well understood to be a multi-stepped process. For example, Marchionini (1995) detailed the importance of browsing in information seeking and O'Day and Jeffries (1993) characterized the seeking process by outlining common triggers and stop conditions that guide people's search behaviors as their information needs change. Bates (1989), Belkin (1993), and Belkin, Marchetti, and Cool (1993) proposed search interfaces that allow users to modify and refine their queries as their information need evolves, thus modeling search as an information gathering activity rather than a single, static search.

Although finding typically involves simple searches for information that is known in advance, the search behavior follows the broader information seeking pattern. Several studies of finding behaviors (Barreau & Nardi, 1995; Ravasio, Schar, & Krueger, 2004), as well as the study reported in Chapter 3, suggest that people prefer to perform even the most directed searches by *orienteering* via small, local steps using their contextual knowledge as a guide, rather than by *teleporting*, or jumping directly to it using a keyword-search utility. For example, if Connie wanted to find the phone number of her oncologist to discuss her treatment options, she could easily type her oncologist's name into the search box to find it. But she is more likely to follow a complex process like first to typing the URL of her hospital into her Web browser, then navigating to the oncology section, and finally searching for her doctor's name and phone number in the directory listed there – even though such a process takes more steps and possibly more time.

Researchers have identified several reasons why people may choose orienteering over teleporting. One is that tools that support teleporting, and in particular tools for searching one's personal space of information (PSI), don't always work (Ravasio, Schar, & Krueger, 2004). For example, the search engine Connie uses may not be able to find her oncologist's phone number if she enters the query directly into the search box. A number of additional reasons why people may choose to orienteer over teleporting appear in the study discussed in Chapter 3. For example, orienteering can provide both an overview of the information space being searched and context about where the desired information is located in that space. For Connie, she better understands oncologist's role at the hospital by orienteering to the phone number. She may also find it easier to search by recognizing

the information she is searching (e.g., "This looks like the name of my oncologist,") rather than recalling it (e.g., "This is how I think my doctor's name is spelled,").

2.3.2 Information Target

Although it is true that people in general tend to find information by orienteering, aspects of the information target, as well as the task and individual performing a search, have an impact on the resulting behavior. For example, the corpus of information being searched (e.g., the entire Web, a folder of personal email messages, a digital library of documents?), the location of the target, the type of information being sought (e.g., a "nugget" of semi-structured information or a summary of information collected from a variety of sources?) (Lieberman, Nardi, & Wright, 1999), and the addressability of the target (Ramakrishnan, 2005) (e.g., how many different ways are there to locate the target and how easy is the location to describe?) are important to consider when trying to understand how people find information. They influence how people express what they're looking for, the strategies they employ to find it, and how easy it is to recognize when found.

One important aspect of the target to consider is whether it is digital or physical. This thesis focuses on the finding of digital information. But people find and re-find within physical spaces as well. There are both similarities and differences between the physical and digital domains. For example, physical organization, location, and layout play an important reminding function in the finding of physical information. Malone (1983) noted that spatial location helps support finding and visible items on a physical desktop help support reminding. He observed the use of both files and piles to organize physical documents, and noted that some people (*filers*) are more comfortable organizing and finding within rigid organizational structures, while others (*pilers*) are more comfortable with loose structures. Barreau and Nardi (1995) noted a similar reminding function of electronic files on a computer desktop. Whittaker and Sidner (1996) describe a related issue for email – users may be reluctant to file email messages because once they a message is are out of their Inbox, they it may be forgotten and difficult to re-find.

2.3.3 Information Task

Factors related to the task that inspires the search can also play a role in information seeking behaviors. Many models of information seeking incorporate characterizations of task and stage of task (Vakkari, 1999; Belkin, 1993; Kuhlthau, 1991). Researchers studying Web information seeking behaviors have also created taxonomies of task types including dimensions such as the purpose (why?), method (how?), and content (what?) of the search (Morrison, Pirolli, & Card, 2001).

The stage of a task is of special interest. Kuhlthau (1991) discusses stages of initiation, selection, exploration, formulation, collection, and presentation in information seeking. For example, if during her first search for "breast cancer treatments" Connie spent time in the initiation and selection stages of understanding cancer treatments (e.g., by browsing general treatment sites), this invested effort would influence her behavior during the collection and presentation stages when it came time to make a treatment decision – even if she performs the identical keyword search she performed initially.

2.3.4 Individual Factors

While the information target and the information task are both important, another important factor that influences finding and re-finding strategies is the individual performing the search. Different people have different ways of approaching problems, and these different approaches carry over to how they find information. For example, if a person approaches learning in a holistic manner, starting with a big picture and then diving in to understanding the details, they are likely to be more exploratory in their information finding behavior (Ford et al., 2002). People also have different backgrounds, and their knowledge of the Web and searching techniques (Hölscher & Strube, 2000) have been shown to influence search behavior. On the other hand, the searcher's familiarity with the search task or domain has commonly been investigated as an influencing factor (Bhavnani & Bates, 2002; Kelly & Cool, 2002; McDonald & Stevenson, 1998; Shiri & Revie, 2003; Wildemuth, 2004), but the effect on the search process remains unclear (Wildemuth, 2004).

The importance of individual differences in search behavior for search result personalization is discussed further in Chapter 4. As suggested by the research mentioned above, the studies in Chapter 4 suggest Web search tools can be enhanced significantly by considering the variation in relevancy of results for users. The following section discusses systems that have attempted to personalize search results, and compares them with the successful search personalization system presented in Chapter 5.

2.4 Related Work on Personalizing Search Support

There have been a number of prior attempts to personalize Web search (Keenoy & Levene, 2005). One of the studies in Chapter 4 suggests that people rate the results to the same queries differently because they had different intents. One solution to ambiguity is to aid users in better specifying their interests and intents. This can be done by having users describe their long-term general interests. For example, users may be asked to build profiles of themselves by selecting categories of interests. These profiles can then be used to personalize search results by mapping Web pages to the same categories. Many commercial information filtering systems use this approach, and it has been explored as a means to personalize Web search results by Gauch et al. (2004) and Speretta and Gauch (2004). Personal profiles have also been used in the context of the Web search to create a personalized version of PageRank (Jeh & Widom, 2003) for setting the query-independent priors on Web pages. Liu et al. (2002) used a similar technique for mapping user queries to categories based on the user's search history.

In addition to asking users to describe their long-term interests, information about the user's intent can also be collected at query time by means of techniques such as relevance feedback or query refinement. Koenemann and Belkin (1996) examined several different interface techniques that varied in their transparency for allowing users to specify how their queries should be expanded. Anick (2004) and McKeown, Elhadad, and Hatzivassiloglou (2003) explored alternative methods for generating query refinements. Relevance feedback and query refinement harness a very short-term model of a user's interest, and require that a query first be issued then modified.

While it appears people can learn to use these techniques (Anick, 2004; Koenmann & Belkin, 1996), in practice, on the Web they do not appear to improve overall success (Anick, 2004; Eastman & Jansen, 2003), and such features have been found to be used rarely. People are typically unwilling to spend extra effort on specifying their intentions. The findings reported in Chapters 3 and 4 suggest that instead of fully specifying their search goals up front, people often browse to their targets via pages identified by less precise but more easily specified queries. This result resonates with the intuitions of Nielsen (1998), who cautions against requiring users to perform extra work for personalization. Also, even with additional work, it is not clear that users can be sufficiently expressive. Participants in the study in Chapter 4 had trouble fully expressing their intent even when asked explicitly to elaborate on their query. Similarly, participants in the modified diary study presented in Chapter 3 were found to prefer long search paths to expending the effort to fully specify their query.

A promising approach to personalizing search is, instead of requiring users' information goals be stated explicitly, to infer them automatically. This inference can be done in a variety of ways. Kelly and Teevan (2003) review research on the use of implicit measures to improve search, highlighting several approaches in the literature that seek to tailor results for individuals. A wide range of implicit user activities have been proposed as sources of information for enhanced Web search, including the user's query history (Shen & Zhai, 2003; Speretta & Gauch, 2004), browsing history (Morita & Shinoda, 1994; Sugiyama, Hatano, & Yoshikawa,, 2004), and rich client-side interactions (Bharat, 2000; Budzik and Hammond, 1999; Morita and Shinoda, 1994).

The focus of the system developed in Chapter 5 is on the use of implicit representations of a user's long-term and short-term interests. With this approach to personalization, there is no need for the user to specify or maintain a profile of interests. Unlike the systems described above, very rich client models are explored that include both search-related information such as previously issued queries and previously visited Web pages, and other information about the user such as documents and email the user has read or created. The paradigm allows the contribution of different sources of information to be evaluated over different periods of time to the quality of personalization in different contexts.

Regardless of whether implicit or explicit measures are used for the personalization, Pitkow et al. (2002) describe two general approaches to personalizing search results for individual users. In one case, the user's query is modified or augmented before being sent to the search engine. For example, the query "cancer", when issued by breast cancer patient, might be expanded to "breast cancer". In the other case, the same search request is issued for all users, but the results are re-ranked using information about individuals. In this case, while the same results are returned from the underlying search engine, results relating to breast cancer are pushed up in rank before being presented to the breast cancer patient, and results relating to the astrological sign Cancer are pushed down.

The approach presented in this thesis focuses on the later, result re-ranking, using the current Web search results as a starting point for user-centric refinement. This is similar to what is done by the Compass Filter (Kritikopoulos & Sideri, 2003). One benefit of result re-ranking is that the original ranking of results by a Web search engine is a useful source of information. For example, the first several results are particularly likely to be

relevant to all searchers, as is shown in the study in Chapter 4 of results consider relevant by individuals.

The re-ranking is performed in Chapter 5 by taking a relevance-feedback perspective (Sparck Jones, Walker, & Robertson, 1998) on modeling personalization. Relevance feedback has a solid theoretical foundation and a long history of application to information retrieval. The approach differs from standard relevance feedback in that it does not do any query expansion nor require explicit judgments. The methods are distinguished from blind or pseudo-relevance feedback as they operate over a longer time frame than an individual query (Ruthven & Lalmas, 2003).

2.5 Related Work on Study Methodology

There are a number of the different studies included in this thesis. It is interesting to consider the particular study methodologies employed when considering the results presented because different types of studies and different study structure influence the findings in different ways. This section first presents the different ways evaluation components are used in this thesis. It then dives into the different study types that are employed in greater detail. Much of this discussion derives from a review study methodologies for personal information management written by Kelly and me (Kelly & Teevan, in press).

2.5.1 Study Components

This section discusses several important components of the studies included in this thesis – participants, tasks, and measures – in greater detail.

Participants

One of the big challenges to conducting the studies presented here was recruiting participants. In some cases, such as the modified diary study introduced in Chapter 3, participants needed to be willing to grant at least partial access to their personal information collections. Granting someone access to personal information collections and behaviors requires a level of self-disclosure which can make people uncomfortable. Users may be self-conscious about how they have organized and grouped things or of the types of communications in which they've engaged. Users may also be unsure about the contents of their collections and want to avoid embarrassment. The time commitment required for participation and the potential disruption to regular activities can also make participant recruitment difficult. The studies presented in this thesis vary in the level of commitment and disclosure required by participants, and accordingly vary in the depth and quantity of data collected.

Tasks

Identifying appropriate tasks to study also presents some significant challenges, particularly in the realm of personalization. The types of tasks that are relevant to finding are very broad, user-centric and situation-specific. Further, tasks are often identified at varying levels of specificity. For instance, "finding information on breast cancer" is a

task, but one might subdivide this task into searching for a list of treatments, managing and filing breast cancer Web pages, and setting up folder to store relevant information. In natural environments it is difficult to anticipate the number and kinds of tasks that users are interested in accomplishing. While there are many generic classes of tasks that users do, many tasks are idiosyncratic. Tasks also differ according to the length of time they take to accomplish and the frequency with which users work on them. Multitasking is also common; understanding how to capture, document and simulate these activities present even more challenges.

Researchers have dealt with the problem of defining tasks in a number of ways, including the use of natural tasks. For instance, researchers have allowed subjects to self-identify tasks and structure their activities in ways that make sense to them (Kelly & Belkin, 2004; Czerwinski, Horvitz & Wilhite, 2004). While this approach presents its own problems, it does allow for a more user-centered task focus. Other studies have deployed a tool, observed what users did with it, and inferred tasks from these observations (Dumais et al., 2003), while others (Harada et al., 2004) have started with generic tasks and personalized them to individual users based on the content of such collections. Studies where tasks are assigned to users most often occur in laboratory setting and creating task scenarios that are robust enough to allow for comprehensive, valid and reliable evaluations in this setting is a challenge. Creating tasks and task scenarios require a significant time investment, and there is no guarantee that such tasks will maintain their relevancy over time.

The studies presented in this thesis take various approaches to task definition. Generally, those studies that focus on understanding finding behavior are open ended and allow users to self-identify tasks. For example, the modified diary study discussed in Chapters 3 and 4 was structured to merely ask participants if they had looked for something recently, leaving open exactly what was looked for, and even what "looked for" meant. Those studies in this thesis intended to evaluate tool design tend to be more closed with respect to task, often explicitly defining the task, as is done, for example, in Chapter 9.

Measures

In order to evaluate the search tools developed as part of this thesis research, it was necessary to measure their performance. Devising appropriate measures involves the provision of two basic types of definitions: *nominal* and *operational*. Nominal definitions state the meaning of a concept; for example, Nielsen (2006) defines *usability* as "a quality attribute that assesses how easy user interfaces are to use," and divides the concept into five dimensions (learnability, efficiency, memorability, errors and satisfaction). Operational definitions specify precisely how a concept (and its dimensions) will be measured. For instance, an operational definition of *learnability* might include three questions and a 5-point Likert scale for responding. Alternatively, one might operationalize learnability as the length of time it takes a user to learn to use an interface. Without both nominal and operational definitions it is impossible to know exactly what concepts researchers hope to capture with their measures, and it impossible to evaluate the validity of measures and subsequent results.

Measures that are considered in this thesis to evaluate the systems presented are the system's effectiveness, efficiency, satisfaction and ease of use. A tool is *effective* if it

helps users accomplish particular tasks. For example, did the search tool Connie used to search for "breast cancer treatments" help her find what she was looking for? While the way in which effectiveness is operationalized varies according to study purpose, one common way to measure effectiveness is to count the number of tasks a user is able to accomplish successfully.

A tool is *efficient* if it helps users complete their tasks with minimum waste, expense or unnecessary effort. A common way to measure efficiency is to record the time it takes to complete a task. For example, how long did it take Connie to find what she was looking for? Efficiency can also be measured by the number of actions or steps taken to complete a task.

Satisfaction can be understood as the fulfillment of a specified desire or goal. It is often the case that when people discuss satisfaction they speak of the contentment or gratification that comes from fulfilling particular goals. Was Connie content with the search process? Did Connie feel her needs were met by the tool she used? Satisfaction is often operationalized as one or more statements that users assess with Likert-type scales.

Ease of use is related to the amount of effort which users expend executing and/or accomplishing particular tasks. Ease of use is very tightly related to efficiency: if a tool isn't easy to use, then it is likely to result in inefficient use. As with previous measures, it is common for ease of use data to be gathered via Likert-type scales. One might also measure of the number of errors made by users while trying to accomplish particular tasks.

It has been demonstrated that users tend to rate ease of use measure high even when they are unable to successfully use software to complete specific tasks (Czerwinski, Horvitz, & Cutrell, 2001). Czerwinski, Horvitz and Cutrell investigated subjective duration assessment as an implicit measure of ease of use. Subjective duration assessment asks users to estimate the length of time it took them to complete particular tasks. The theory is that users will overestimate time when the task was difficult for them to accomplish and underestimate time when the task was easy for them to accomplish. The value of this measure is that it allows researchers to obtain an implicit measure of ease of use, which can then be compared to more explicit ease of use measures.

2.5.2 Approaches

Just as the components that make up a study can vary, studies can take different methodological approaches that influence their ability to shed light on behavior. Laboratory studies allow for controlled study, but do not necessarily provide a realistic picture. Log analysis shows real life behavior, as do observational studies, but both have limited ability to show the motivation behind the information seeker's actions. Interview or survey studies give insight into the searcher's motivation, but tend to involve the study of a limited number of participants and, as self-reported data can be inaccurate, are not necessarily realistic. All of these different approaches are employed in this thesis.

Observational, Interview, Survey or Questionnaire Studies

Observational studies and studies based on interviews or surveys and questionnaires allow researchers to develop a deep understanding of naturalistic search behavior by

collecting information about users' real-world information activities in familiar environments with familiar data. The results of several such studies are presented here: Chapters 3 and 4 discuss a modified diary study conducted to understand how people find information, and Chapter 8 presents a naturalistic study of Web data that looked at how change affects re-finding.

Previous observational studies of personal information interaction have tended to focus on users' interactions with various different subsets of their personal information, such as paper documents (Lansdale, 1988; Malone, 1983), email (Whittaker & Sidner, 1996), files (Nardi & Barreau, 1995), and the Web (Jones, Dumais, & Bruce, 2002; Sellen, Murphy, & Shaw, 2002). The modified diary study discussed in Chapters 3 and 4 is unique in that it focuses on directed search and looks at behavior across a broad class of electronic types, including email, files and the Web. By focusing on the communalities of interaction across types, it gives a broader understanding of general directed search techniques.

The study presented in Chapter 8 is unusual in that the observations analyzed are collected from Web pages. The Web is an emerging source of data for observational studies. Several studies have analyzed postings collected from specific message boards to understand topics ranging from how people view robotic pets (Friedman, Kahn, & Hagman, 2003) to how they recover from injuries (Preece, 1998). Observations have specifically been collected using search results, as is done in the study in Chapter 8. Good and Krekelberg (2003) constructed KaZaA queries to see if people accidentally exposed personal data. Data collected from the Web can be noisy, but the large quantity that can be cheaply gathered compensates for the noise. Further, data can be collected by mining the Web that might otherwise be unobtainable. It would have been difficult to devise a study such as the one presented that would have permitted naturalistic observations of people having difficulties re-finding during personally motivated searches.

Log Analysis

Like observational studies and studies based on interviews or surveys and questionnaires, log analysis allows researchers to gain a realistic picture of what search is like in the real world. Log analysis allows researchers to observe a greater variety of behavior than laboratory and observational studies, and gives a very realistic picture of people's actions, although it gives no insight into people's underlying motivation. In this thesis, query log analysis is used to understand how common re-finding behavior really is and to look at the affect of search result changes on re-finding. Studies investigating how people re-find information have tended to be small-scale laboratory or interview based studies.

In general, query log analysis (Anick, 2004; Broder, 2002; Jansen, Spink, & Saracevic, 2000; Silverstein et al., 1999; Spink et al., 2001; Spink et al., 2002; Tolle, 1984) provides insight into observable behavior, such as the types of information people search for (e.g., sex). It can also provide a cursory understanding of how people search (e.g., they use very short queries), but even when researchers supplement query log analysis with user surveys (Broder, 2002), the insight in intention is limited. For example, query log analysis is restricted to search activities that involve search engines, omitting many other search activities – including, for example, the 61% of search activity that did not involve

keyword search in the study discussed in the following chapter. Web site log analysis (Jul & Furnas, 1997) addresses a broader class of Web behaviors but conflates undirected browsing behaviors and search, whereas the research in this thesis focuses solely on search.

Web log analysis focusing on re-finding has shown that Web site re-visitation is very common (Cockburn et al., 2003; Tauscher & Greenberg, 1997; Herder, 2006). The percentage of Web page visits that are re-visits is estimated at between 58% (Tauscher & Greenberg, 1997) and 80% (Cockburn et al., 2003). While many of these re-visitations occur shortly after the first visit (e.g., during the same session using the back button), a significant number occur after a considerable amount of time has elapsed. The results of these studies have been used to inform Web browser history and back button functionality (Komlodi, 2004; Komlodi, Soergel, & Marchionini, 2006).

Surprisingly, however, little analysis of re-visitation and re-finding has been done of Web search query logs. Some log analysis studies have looked at queries clustered by topic (Broder, 2002; Ross & Wolfram, 2000) to understand how topics evolve. For example, Beitzel et al. (2004) looked at queries in aggregate over time to understand changes in popularity and uniqueness of query topics at different times of day. Wang, Berry, & Yang (2003), in another study of queries over time, found topics and search behavior in general vary little. A few studies have looked at the queries issued by individuals over time, but focused on short periods of time, also called query sessions (Jones & Fain, 2003; Kamvar & Baluja, 2006; Lau & Horvitz, 1999). The work presented in this thesis is unique in that it looks at query patterns for individuals over long time periods.

Laboratory and Controlled Studies

One of the biggest challenges of conducting naturalistic studies using the above approaches is the lack of control the researcher has over the environment. Studies such as laboratory studies allow researchers to conduct controlled experiments and examine users' thought processes during search by having them think aloud as they search (Capra & Pérez-Quiñones, 2003; Maglio & Barrett, 1997; Muramatsu & Pratt, 2001). Of course, such studies also introduce artificialities that can bias behavior. For example, the search tasks are imposed by the researcher rather than motivated by the user, and task has been shown to affect search performance (Kim & Allen, 2002). As laboratory and controlled studies involve a great deal of reduction, it is important that what is studied be well defined and narrow in scope. For instance, while it may not make sense to study general tool usage in a laboratory setting, evaluating a small subset of features of a tool or targeting a specific behavior, activity or task might be appropriate. Chapter 8 presents several controlled studies that test whether an algorithm developed to support re-finding keeps participants from noticing change and allows participants to conduct re-finding tasks as quickly as a static result list while still enabling the finding of new information.

Test Collections

All of the evaluation approaches discussed thus far involve the direct involvement of people finding information. This is a good way to understand finding behaviors and tools, but also makes it difficult to compare findings across studies and test conditions. An evaluation frameworks makes it possible to compare results across tools, replicate

findings, and explore alternative hypotheses. However, as seen above, there are also many challenges to using evaluation frameworks since individual behavior can vary; it makes little sense, for example, to ask someone to execute generic tasks. Thus, the design of test collections for the evaluation of personalized search systems requires some creative thinking since such collections must differ from more traditional shared test collections.

While there are a number of existing sharable test collections that can provide direction for a personalized search test collection, they did not meet the evaluation needs of the systems developed for this thesis. Information management test beds, such as the Distributed Integration Testbed Project, the D-Lib Test Suite, and those created as part of the Text Retrieval Conference (TREC), are intended to represent large databases and generic users – not personal information and individual users. Search engine logs are another good source for studying aggregate information behavior, but do not necessarily lend themselves immediately to testing tools. These collections also do not provide a complete picture of users' finding behaviors. A limited number of test beds and tools have been developed with an eye towards how individuals deal with the information at hand, such as the TREC HARD track (Allan, 2006), the Curious Browser (Claypool, et al., 2001) and the collection developed by Zhang and Callan (2005). A large corpus of email messages is available, and could be useful, for example, in understanding individual's email classification (Klimt & Yang, 2004).

A personalized search test bed needs to contain rich user information and individualized evaluations for finding tasks. For example, Kelly and Belkin's (2004) collection contains thousands of documents which have associated with them user-defined tasks and topics, relevance assessments, and behaviors such as document display time and retention. Unfortunately, there are numerous privacy concerns associated with sharing collections. For this reason, in the analysis of Chapters 4 and 5 it was necessary to create a collection for the study of personalized search. The collection created involves aggregate information about individuals collected via their personal computers (e.g., word occurrence frequencies, document level information such as time of last access, etc.), and example searches run by the user with a listing of which results are relevant to which searches.

The following chapters highlight how test beds and evaluation approaches like those described above can enhance our understanding of how individuals find information and allow the testing of the tools built based on this understanding. The next chapter begins with the discussion of a modified diary study of finding behavior, and is followed with the analysis of the test bed created as part of this thesis research for search result personalization. This test bed is then used to evaluate the system created as a result of the findings of these studies. Many additional studies are further presented in Part II.

Look for the ridiculous in everything, and you will find it.

- Jules Renard (1864 - 1910)

Chapter 3

Understanding Finding

This chapter presents a naturalistic study of people performing personally motivated searches within their own information spaces. The data collected through this study are analyzed to give a deep understanding of finding behavior. As mentioned in the previous chapter, searching for electronic information can be a complex, multi-stage process, where the information need evolves throughout the course of the search. However, the vast majority of search behavior observed via the study involved search targets that were known in advance (e.g., a phone number or address). Such small, directed searches have been assumed to be simpler than large, evolving information seeking activities. But although very few complex search tasks were observed, participants exhibited a range of complex behaviors to find what they were looking for.

Consider Rachel[1], one participant in the study. She attempted to locate a document that she knew existed in her file system. Although she knew exactly what document she was looking for and her information need was not evolving, she could not describe the document, its contents, or its location in advance:

> I don't know how I could have the directory [the document was in] in mind without knowing its name, but I felt sure which it was.

As a result, even though her need was simple she found her target through a series of small steps, using the local context at each stage of her search to inform her next step.

Despite the fact that participants almost always knew their information target up front, they, like Rachel, typically used small steps to find it. Keyword search was employed in only 39% of the observed searches. Instead of trying to jump directly to their information target using keyword search as might be expected, participants performed directed situated navigation, similar to the Micronesian islanders' situated navigation described by Suchman (1987). This behavior is referred to here as *orienteering* because it is similar to the notion of orienteering in O'Day and Jeffries (1993). This chapter explores the observed range of orienteering behavior, and discusses the observed advantages of

[1] All names and identifying details reported have been changed. Minor changes to the transcripts have been made for readability.

orienteering over teleporting. The study presented is the result of a collaboration with Christine Alvarado, Mark S. Ackerman and David R. Karger (Alvarado et al., 2003, Teevan et al., 2004).

3.1 Study Methodology

The study involved 151[2] semi-structured interviews in which 15 participants reported their most recent search activity. Each participant was interviewed twice daily on five consecutive days, interrupting them in their offices at unspecified times. They were asked to describe what they had most recently "looked at" and what they had most recent "looked for" in their email, their files, and on the Web. Each semi-structured interview lasted about five minutes. The method was similar to the diary studies used in many information interaction studies, as well as the Experimental Sampling Method (Palen & Salzman, 2002). These data were supplemented with direct observation and an hour-long semi-structured interviews with each participant about their information patterns.

Participants consisted of 15 graduate students (10 men, 5 women) in Computer Science at MIT. Participants had attended the university from one to seven years; this range allowed for the observation both of those in the process of developing their information organization and those with long standing structure. This group is certainly not representative of the general public (e.g., all were expert computer users). However, participants did reveal some important search issues. In this chapter, the surprising lack of search tool use among this population is discussed extensively. Since participants were familiar with complex information spaces and sophisticated search tools, it seems likely that this lack of tool use is even more prevalent among the general population.

In the interviews, the term "look for" was used instead of "search". This was done so as not to predispose participants to think specifically of keyword search. What precisely was meant by "look for" versus "look at" was defined by the participants themselves based on what they considered effort. Allowing participants to self-categorize when they had to exert effort to find information (as in, for example, Bernard, 1994) revealed what types of information needs required effort and what techniques were relied on in those cases. Participants were encouraged to give as much detail as possible.

Each short interview was examined independently by Christine Alvarado and me, and each search incident was coded as to the type of search performed, with an 85% inter-rater reliability. Because the incidents were not randomly selected (e.g., there are temporal patterns in people's information use), this chapter presents only qualitatively-based findings. The data were analyzed using standard qualitative techniques (e.g., Ackerman & Halverson, 1998; Strauss & Corbin, 1990). The findings are exploratory and observational, and as with many qualitatively-based studies, the intent is only to analyze interesting phenomena, rather than to confirm existing theory. Accordingly, this chapter presents the incidents that emerged as particularly illustrative of the general patterns observed.

[2] One participant was inadvertently interviewed 11 times. This participant is labeled "M" in Figure 4-1.

3.2 Search Strategies

It was observed that when people searched for specific pieces of information, such as phone numbers or addresses, they generally knew exactly what they were looking for at the onset of their search. Participants were expected to take advantage of this advanced knowledge of their target by using keyword search[3] more often than they would when searching for general information, where the information need often evolves. Surprisingly, only 34 of the 81 searches for specific information that were observed (42%) involved keyword search, compared to 23 of the 42 searches for general information (55%). To understand how participants performed directed searches, and why they avoided keyword search in many cases, a qualitative examination of the data was performed, and two differing search strategies were uncovered: *orienteering* and *teleporting*.

3.2.1 Orienteering

Many directed searches were observed, like the following, where a series of small steps were used to narrow in on the target. Here, although Jim is looking for the office number of a math professor, Ellen Brooks, he does not try to find it directly but instead looks for it via her department's page.

> ***Interviewer****:* Have you looked for anything on the Web today?
>
> ***Jim***: I had to look for the office number of the Harvard professor.
>
> ***Interviewer***: So how did you go about doing that?
>
> ***Jim***: I went to the home page of Math Department at Harvard.

Jim went on to explain that he knew there was a specific Web page with her address:

> ***Interviewer***: Did you know it would be there [on a page] or you just hoped it would be there?
>
> ***Jim***: I knew that she had a very small Web page saying, "I'm here at Harvard. Here's my contact information."
>
> […]
>
> ***Interviewer***: So you went to the Math department, and then what did you do over there?
>
> ***Jim***: It had a place where you can find people, a link to the page where you can find people and I went to that page and they had a dropdown list of visiting faculty, and so I went to that link and I looked for her name and there it was.

This search by localized or situated navigation, shown in Figure 3-1, is an illustration of *orienteering.* Orienteering involves using both prior and contextual information to

[3] Keyword search includes the Windows file system "Find", the UNIX `grep` or `find` commands, any Web-based search engine, and any keyword search in an email client. The `grep` command allows a user to search for files containing a given word or set of words; `find` allows the user to search for a file by its name.

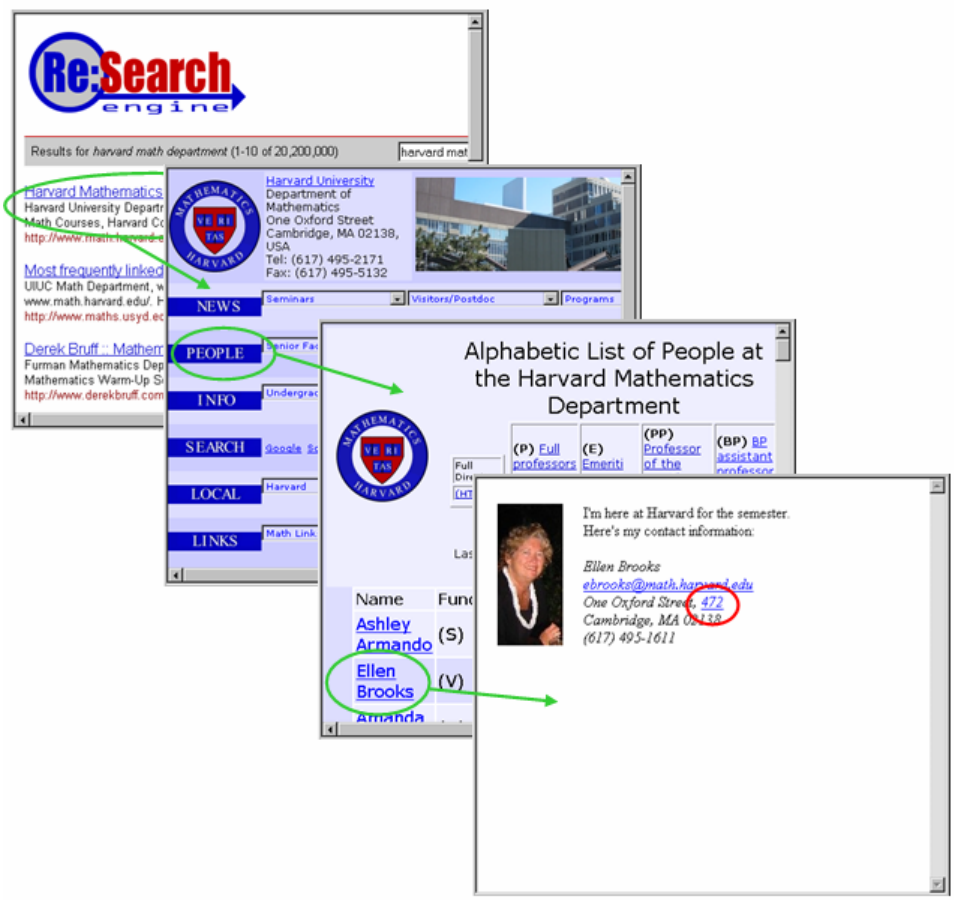


Figure 3-1. Jim's search for something as simple as an office number is a multi-stepped process.

narrow in on the actual information target, often in a series of steps, without specifying the entire information need up front. Orienteering appeared to be heavily relied upon, even in directed search for specific information. Its characteristics are explored further in the following sections.

3.2.2 Teleporting

At the other end of the spectrum from a search strategy that involves many local, situated steps is a strategy called *teleporting*. When a person attempts to teleport, they try to jump directly to their information target. Figure 3-2 illustrates what Jim's search for Professor Brooks' office number would have looked like if he had tried to teleport instead. As might be expected, incidents of people teleporting (or trying to teleport) were observed. For example, to find housing prices in Boston, Alex went to an Internet search engine and typed in "real estate prices Boston". In doing so, he tried to jump directly to that information. Of course, perfect teleporting was rarely observed in practice—even in this example, Alex reported having to "browse through all the different graphs and statistics" that the returned site provided. Regardless, it is worth noting that participants do sometimes attempt to jump directly to their information target, but also that such attempts

Figure 3-2. An example of what the same search shown in Figure 1-1 would look like if Jim had tried to look for Ellen Brooks' office number by directly teleporting to the information.

were surprisingly rare. This chapter addresses why people often chose not to teleport, and what they did instead.

3.3 Exploring Orienteering

Orienteering denotes a search behavior in which people reach a particular information need through a series of small steps. Within this general class of activities, a range of search behaviors were observed, including the size of the steps taken along the way and the methods chosen to take those steps.

Most commonly, the participant knew definitively how to get into the vicinity of the information in question and made a large step to get to the correct area. Once there, the participant used local exploration to find the information target. As an example, Erica was trying to find a piece of information about Quebec. She first typed the URL

"bonjourquebec.com", which she knew to exist, and then she "kept clicking on links from the main page" to get the information she wanted.

Erica's search also illustrates that the participants often associated their information need with a particular information source. Erica associated information on Quebec with the Bonjour Quebec Web site. Participants made this type of association not only on the Web, but also in their email and files. In another incident, Carla performed a search to determine the location of a meeting. She knew this information was contained within a particular email, so instead of searching for the information (e.g., by doing a keyword search for "Tuesday meeting location"), she searched for the email – the source of the information that she needed.

This ability to associate information with a source was critical in helping participants orienteer to their information target, as participants often remembered a lot about the source. During Carla's search for the email containing the meeting location, she didn't know much about where the meeting was, but once she associated this information with a particular email she was able to recall a large amount of meta-information to help guide her search, including the folder the email was located in, the date it arrived, who the sender was, and an idea of where it would be visually.

A person's information target could be associated with a source even when the participant had never seen the target or the source before. This is illustrated in a search Lawrence conducted to determine if a particular company had any job openings. Although he had never been to the company's Web site and did not know the URL, he guessed a URL, typed it in, and successfully reached the company's homepage—the source where he suspected he would find the information he was looking for. There he found a link to a listing of job openings.

These examples of orienteering involved steps made by typing URLs, clicking on links, and navigating through email. A large variety of techniques appeared to be used to make small steps orienteering, including keyword search. Carla used keyword search in orienteering when looking to buy an electric toothbrush. She first performed a keyword Web search to find an online pharmacy site. Then, after navigating through the pharmacy site, she performed a site search for electric toothbrushes. Although most of her activity involved keyword search, the strategy she employed was orienteering, taking relatively small steps to narrow in on a goal. As in Bates (1979), it is worthwhile to draw a distinction between search strategies and tactics: Orienteering and teleporting are strategies; keyword search is a tactic that can be used to achieve either strategy.

Orienteering was not always characterized by a large step followed by local exploration, as in the above incidents. Often it appeared as if the participant was following a path they couldn't quite articulate but believed to exist. In the following incident, Rachel described navigating down her directory hierarchy using cues at each level to remind her which step to take next.

> ***Rachel***: I didn't know necessarily how to type that path name from memory and so I used the path completion. […] I knew what its name was relative to the directory above it. I didn't know the path down the whole tree.

Interviewer: Did you ever make any false completions, start with the wrong letter or something?

Rachel: No.

Compared to the previous incidents, Rachel's steps as she narrowed in on her goal were relatively small. Because her memory of the path and even the target was so vague, these small steps allowed her to reach a target she may not have been able to access using any sort of keyword search.

3.4 The Advantages of Orienteering

One could argue that the reason that people rely on orienteering is that the tools available for teleporting do not work well enough yet. For example, one participant attempted to teleport but failed. She fruitlessly tried to determine how much to tip hairdressers performing various keyword searches using the words "tip", "hairdresser", "percent", and "gratuities". However, a number of cases were also observed where people chose to orienteer even when teleporting might have worked. For example, Neil had difficulty finding the location of a city in Switzerland. He did not know exactly where to find that information, but he had four map sites bookmarked. Rather than relying on a keyword search directly to locate the city (something many Web search engines explicitly support), he used the bookmarks to access the map sites and then clicked around to see whether he could find a map with the information he was looking for.

This incident with the map site was not an isolated case; many cases were noted where people made no attempt to teleport to their information need, even when teleporting appeared to be a viable option. It is likely that orienteering is more than a coping strategy – it appears to hold many advantages even compared to a significantly improved search engine. This section covers three properties of orienteering that appeared, in the study, to be important to the participants: it decreased their cognitive load, allowed them to maintain a sense of location during their search, and gave them a better understanding of their search result.

3.4.1 Cognitive Ease

Orienteering appeared to lessen participants' cognitive burden during their searches. It did this by saving them from having to articulate exactly what they were looking for and by allowing them to rely on established habits for getting within the vicinity of their information need, thus narrowing the space they needed to explore.

In the incident described in the introduction, Rachel looked for a specific file, but could not articulate the properties or location of that file. She relied on cues during the search process to help her narrow in on the file, saving herself the cognitive burden of specifying the exact file she wanted:

> I knew what directory I thought it would be in. I had this mental idea of which directory it was. It is just that I didn't know necessarily how to type that path name from memory and so I used the path completion to get the directory. […] I

> didn't know that path down the whole tree. I didn't know how many levels down it was, even though I knew what the name was at the lowest level of that sub-directory.

In a similar situation, Meredith looked for the location of some documentation. She had no sense of where to find the documentation itself, but she did remember that an email she received contained the path to the documentation. Although she did not remember the path to the email either, she recalled meta-information about the email that she could use to help her orienteer to it:

> The last email I read was an email from Bill describing where to find the documentation on [a project]. […] And I looked for it in the research directory which was where I put things that are sort of done for a research. […] I went and tried to look for the email that looked familiar for being the correct one. The only thing I had to go by was that it was probably from Bill. But I wasn't exactly positive on that. And I wasn't sure where it would be anyway.

In the above cases, the participants orienteered because it helped remind them of exactly what they were looking for and how to find it. It would have been difficult for them to describe their search target at the beginning of their search. In other cases, the participants had a good idea of what they were looking for, but had strong associations between their target and an intermediate location. In these cases, orienteering was an automatic response or habit, where the participant used the first route to their target that came into their mind. In the following instance, Fernando orienteered to a paper posted on the Web through a familiar source:

> ***Fernando***: So Web pages, as a result of getting the… lab memo announcement from Tony, I went to [the lab's homepage] and then clicked on publications and then looked at 2001 publications and looked for something to see if it were up there and how it was listed [and so forth].
>
> ***Interviewer***: So why did you choose to go that route?
>
> ***Fernando***: Because, well I knew it was a [lab] memo and the only thing I know about it was it was with the… Lab and I figured it would be a click or two away from the home page, so I chose to go navigate through the home page and it didn't take me too long to find publications on the lab page. I was just looking at it, it is right there. It is under research, publications.

In other cases, the importance of relying on habit was even more explicit. Here, Meredith had just searched for a restaurant using a path that had been recommended to her, instead of finding it as she normally would:

> ***Interviewer:*** Next time you search for restaurants, how do you think you'll do it?
>
> ***Meredith***: Either way. Whichever way I remember first.

This suggests that orienteering might sometimes be used because it is easier to recall the intermediate steps than to explicitly name the target.

3.4.2 Sense of Location

The relatively small steps taken in orienteering also appeared to allow participants to maintain a sense of where they were, helping them to feel in control, know they were traveling in the right direction with the ability to backtrack, and feel certain that they had fully explored the space when they could not find what they were looking for. Recent literature suggests that people are bad at understanding the models that search engines use (Muramatsu & Pratt, 2001), and this finding could suggest why teleporting, in contrast to orienteering, might feel disorienting and untrustworthy to some people.

In a particularly telling incident, Lawrence performed an extensive search to determine if a company was publicly traded. Throughout his search he seemed to try to keep a sense of place. He began by visiting the company's home page via a URL he was emailed and looking at links there:

> I looked at some links on that page… I didn't actually search, I just looked at the headings.

He was unable to locate the information he needed on the company's Web page, so he found another page he thought was relevant in his browser's history. The page in his history was not the homepage for the second site, so he took advantage of the sense of location the URL provided him and deleted a suffix from it to arrive at the site's homepage. Still not finding whether or not the company was public, he went to his browser's homepage by clicking on the home button and attempted to find the information from the financial links located there. When he failed to find the stock price of the company, he felt he had exhaustively explored the space, and concluded that the company must be private, despite not having found an explicit answer.

Although Lawrence's search was complex, involving several different Web sites and much exploration at each one, he explicitly directed the majority of his search in order to stay in a portion of the Web that he was familiar with. Although he began on an unfamiliar Web site, the company's Web site, even his initial step was not blind because he had received an email saying that it existed. He also used the technique of deleting the suffix of a URL to arrive at a site's homepage in order to avoid a blind step to that page, either through the use of a search engine or by guessing the URL. When he failed to find the information on the company's page, he fell back to two sites to which he had been before, at least one of which (his homepage) was very familiar.

3.4.3 Understanding the Answer

Another observed advantage of orienteering was that it gave people a context for their results. Participants appeared to use the context of the information they found to understand the results and to get a sense of how trustworthy those results were. Context was often essential in helping the participant understand that they had found what they were looking for, as illustrated in the following incidents in which Rachel looked for files:

> ***Rachel***: I listed the directory. I saw it. Let's see, I saw "setup.1.lisp", setup.2.lisp", setup.3.lisp" and "setup.3" was the most recent one. That is the one I took.
>
> ---
>
> ***Rachel***: I was looking for a specific file. But even when I saw its name, I wouldn't have known that that was the file I wanted until I saw all of the other names in the same directory and that made me realize what the naming scheme had been.
>
> ***Interviewer***: So by itself you wouldn't have known?
>
> ***Rachel***: By itself, I probably [would have] thought that [it] wasn't right.

The importance of context in understanding search results has been reported previously (Dumais, Cutrell, & Chen, 2001; Lin et al., 2003). Orienteering was observed to have an added advantage over simply presenting keyword search results with some surrounding context: it allowed participants to arrive at their result along a path they could understand. This allowed them to understand exactly how the search was performed, and consequently to accept negative results. This understanding is what allowed Lawrence to conclude that a company was not public, as described previously. The use of context is also illustrated in the following incident in which Alex looked for a particular image but did not find it:

> ***Interviewer:*** So how'd you go about looking around for the bigger logo?
>
> ***Alex:*** Systematically. I basically clicked on every single button until I was convinced that, in fact, they all used the same style sheet and there was no home for a bigger one… I don't think that it exists on the Web page.

Because Alex controlled the search, he could say that he believed the image couldn't be found.

3.5 Supporting Finding

The reasons people prefer to use keyword search tools as steps in the orienteering process rather than as a way to navigate directly to their information target have broad implications for keyword search tool design. The implications that influence the tools presented in this thesis are discussed here; those implications that are beyond the scope of this thesis are discussed further in the Future Work section of Chapter 10.

A major observed motivation for why people orienteer is that orienteering lessens the cognitive burden of fully specifying the search target. Participants clearly found the effort of navigating through several steps to be easier than generating more detailed queries. A search system that performs some of the query specification on behalf of the user could alleviate some of the cognitive burden of keyword search, and this, along with the findings about individual differences in search behavior presented in Chapter 4, strongly motivates the use of implicit information about the user's interests during a search session as is done in the both the system presented in Chapter 5 and the system presented in Chapter 9. It also suggests that for re-finding, recognizing one's past query

is easier than recalling it, and for this reason the system presented in Chapter 9 supports a sophisticated form of query completion, matching the current query with any similar previous query.

The source of the information target was also seen to be an important navigational clue for people. For example, Erica associated information about Quebec with the Web site bonjourquebec.com, and used that information source as a way to narrow down or refine her search for Quebec information. Similarly, Jim associated with information about Professor Brooks with the Harvard Mathematics Department, and used his association to help him restrict his search and find her office number. Navigating by source provides many of the benefits discussed in Section 3.4. Knowing the source of information helps people understand their answer better because the source can provide context (Section 3.4.3). Searching by source supports a sense of location (Section 3.4.2) because it allows the searcher to work with known entities in a known information space, and it supports cognitive ease (Section 3.4.1) because it is often easier to specify a probable information source for the target and then to perform a more restricted search within that source than it is to fully specify the information target enough to uniquely identify it in a large space. The personalized search system presented in Chapter 5 explores how to take advantage of information sources that the user trusts by biasing result rankings in favor of results from sources the user has seen before.

Knowledge is of two kinds. We know a subject ourselves, or we know where we can find information on it.

- Samuel Johnson (1709 - 1784)

Chapter 4

Why Finding Requires Personalization

The previous chapter discusses orienteering behavior and its variations. While there are many aspects of finding behavior that appear to be common for all people, aspects of the individual performing a search can have an impact on finding behavior. This chapter looks at how individual differences affect finding. Further analysis of the data collected through the study presented in Chapter 3 suggests orienteering was especially common for participants with unstructured information organization. The chapter begins by presenting evidence that *filers*, or people who file their information, are more likely to take larger steps when finding than *pilers*, or people who leave their information in piles.

Although participants appeared to prefer orienteering to teleporting, keyword search was a common and important tactic. Search engines were relied on heavily as a step in the orienteering process, rather than as an instrument for teleporting. Following the discussion of filers and pilers, this chapter dives deeper into the variation across individuals in how they use keyword search. In another study, the diverse goals people have when they issue the same query to a Web search engine are investigated, and the ability of current search tools to address such diversity explored, in order to understand the potential value of personalizing search results.

4.1 Organizational Behavior Affects Finding

Some of the variation in orienteering behavior observed in Chapter 3 appears to be due to differences in behavior between individuals. Some individuals relied more on keyword search as a tactic during their search activities than others. Somewhat ironically, these same individuals tended to put more effort into the organization of their information spaces, and thus were better set up to support orienteering.

When email use was examined, two groups of individuals emerged: those who found previously received messages in their inboxes most of the time and those who found previously received messages in other email folders. Those who found email in their inboxes almost never spoke of interacting with emails that were not in their inboxes and almost always expected to find messages in their inboxes, implying they did not file their messages in general:

> ***Interviewer***: How did you know [the message] was still in your inbox?
>
> ***Susan***: I don't know if I'm weird… [but] I don't move stuff.

Those who didn't find email in their inbox almost always went directly to folders and never expected to find messages in their inboxes, implying they regularly filed their messages. For example, Fernando found a previously received email in his inbox during only one of his six email searches. When asked whether or not he had expected to find the message in his inbox he said:

> No… [but then] I thought, "where would I have put this," [and] I hadn't really had a category for that yet, so I kind of still have it in my inbox and I haven't quite decided where I could put it yet.

His response implies that he ordinarily files his messages except in the unusual circumstances in which he cannot assign a topic to a message.

As discussed in Section 2.3.2, Malone (1983) classified people as *filers* or *pilers*. Filers organize information using a rigid structure, and pilers maintain an unstructured information organization. Using this classification, those participants who regularly filed their email can be called *filers,* and those who piled the email they interacted with in their inbox can be called *pilers*. While similar, finer-grained categorizations of people's email behavior have been suggested (Whittaker & Sidner, 1996), this study was not designed to explore such distinctions. However, the study does allow for insight into how filers' and pilers' search behaviors varied. The analysis expands on work by Ducheneaut and Bellotti (2001) that discussed filers' and pilers' search efficiency by exploring how organization behavior correlates with search behavior.

Participants clustered into filers and pilers on the basis of their email use. Six participants each found email in their inboxes in fewer than 40% of their email search activities (mean=23%), while the remaining seven[4] each found email in their inbox in over 80% of their email search activities (mean=95%). This difference was significant ($p<<0.001$). The difference was not related to one group searching in their email more, as the average number of email searches observed per subject was not significantly different for the two groups (filers=5.0, pilers=4.4).

Filers and pilers tended to rely on different search tactics when looking for things within their files and on the Web. As can be seen in Figure 4-1, filers reported performing more searches for files (or information within those files) than pilers (filers=5.0, pilers=2.4, $p<0.02$). A possible explanation for this result might be that filers, despite their organizational efforts, lose information more often. However, a more plausible explanation is based in how filers and pilers interact with their electronic information. Filers are used to assigning specific locations to their electronic information and going to that specific location to retrieve it. Pilers, on the other hand, maintain a more unstructured organization, and commonly must look around even when directly accessing information. For this reason, pilers would be less likely than filers to report some small amount of looking around as a search activity.

[4] Two participants never reported finding anything in their email, and thus are not included in this discussion.

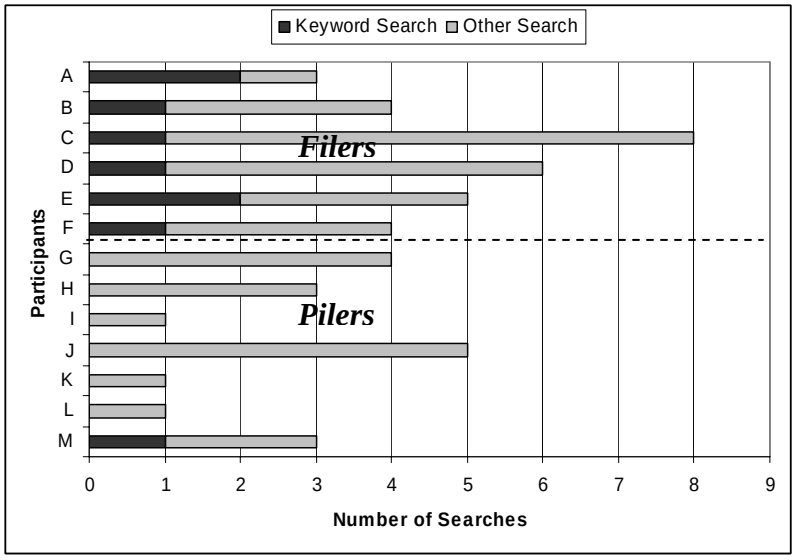


Figure 4-1. The number of times participants used each search tactic in their files. Filers searched more than pilers and use keyword search more often.

A difference in how often filers and pilers used keyword search to find information was also noted. As seen in Figure 4-1, filers relied more on keyword search when searching in their files than pilers (filers=1.3, pilers=0.1, $p<0.001$). Similarly, while there was not a significant difference in how often filers and pilers used keyword search on the Web (filers=4.0, pilers=5.1), there was a significant difference in the keyword search tools they used. Pilers were more likely to use site search (e.g., search the eBay Web page) as opposed to using a global search engine (e.g., Google). Pilers performed on average 1.7 site searches, while filers only averaged 0.3 site searches ($p<0.02$). Although filers were more likely to use keyword search their files and global search on the Web, these results do not necessarily imply they were more likely to teleport to their information target. Keyword search can be used as a tactic in taking small steps toward a target, and both groups were observed to orienteer in their files and on the Web. However, these results do suggest that filers in general tried to take bigger steps when searching for information.

One potential reason for the difference in search behavior may be that pilers are more used to relying on contextual information to find their target because they typically navigate through a relatively unstructured information space. Thus they have developed habits that involve taking more local steps to first arrive at a site before performing a keyword site search to reach their goal. On the other hand, filers are used to assigning meta-data to information in the filing process, and they are more likely to use this meta-data in the retrieval of information through global keyword search. These data suggest that there exist significant individual differences in how people perform directed search.

4.2 Individuals Find Different Results Relevant

Because keyword search tools are the most common tools used to support search behavior, the following section presents a study that investigates more closely the

variation across individuals for keyword search. The study looks at the diverse goals its participants had when they issued the same query to a Web search engine, and the ability of current search tools to address such diversity, in order to understand the potential value of personalizing search results. Great variance is found in the results different individuals rated as relevant for the same query – even when those users expressed their underlying informational goal in the same way. The analysis suggests that while current Web search tools do a good job of retrieving results to satisfy the range of intentions people may associate with a query, they do not do a very good job of discerning an individual's unique search goal.

Understanding relevance is a complex problem (Mizzaro, 1997; Schamber, 1994), and this analysis only addresses only a small portion of it. The analysis is aimed at assessing the relationship between the rank of a search result as returned by a Web search engine and the individual's perceived relevancy of the result. A considerable mismatch is found due to a variation in the informational goals of users issuing similar queries. The study suggests personalization of results via re-ranking would provide significant benefit for users, and this solution is explored further in Chapter 5. This research was conducted with Susan T. Dumais and Eric Horvitz (2005).

4.2.1 Study Methodology

Because the goal of the study was to understand which search results individuals consider relevant to their queries, 15 people were asked to evaluate how personally relevant the top 50 Web search results were for approximately 10 queries of their choosing. The resulting collection of queries, results and relevance judgments served both as a base for exploring differences in query intent, and as an evaluation collection for the personalized search tool developed as a result and presented in Chapter 5. Each participant also provided an index of the information on their personal computer. The indices provided rich, unstructured information about the individual. This data is used to understand the differences in rating individuals gave to the same results, and to implicitly personalize search result ordering. Participants' indices ranged approximately in size from 10,000 to 100,000 items.

Participants were employees of Microsoft. Their job functions included administrators, program managers, software engineers and researchers. All were computer literate and familiar with Web search.

The Web search results that were rated were collected from MSN Search. For each search result, the participant was asked to determine whether they personally found the result *highly relevant*, *relevant*, or *not relevant* to the query. So as not to bias the participants, the results were presented in a random order. The queries evaluated were selected in two different manners, at the participants' discretion.

In one approach (self-selected queries), participants were asked to choose a query that mimicked a search they had performed earlier that day, based on a diary of Web searches they had been asked to keep. It is likely that these queries closely mirrored the searches that the participants conducted in the real world. In another approach (pre-selected queries), participants were asked to select a query from a list formulated to be of general interest (e.g., "cancer", "Bush", "Web search").

For both the self-selected queries and the pre-selected queries, participants were asked to write a more detailed description of the informational goal or intent they had in mind when they issued the query. Because the pre-selected queries were given to the user, the user had to create some intent for these queries. However, by allowing them to decide whether or not they wanted to evaluate a particular query, the goal was to provide them with a query and associated results that would have some meaning for them. By using pre-selected queries, it was possible to explore the consistency with which different individuals evaluated the same results. Such data would have been difficult to collect using only self-selected queries, as it would have required waiting until different participants coincidentally issued the same query on their own. The conclusions drawn from pre-selected queries are validated with data from the self-selected queries.

Ignoring queries with no results, or where no results were marked relevant, a total of 131 queries were collected. Of those, 53 were pre-selected queries and 78 were self-generated queries. The number of people who evaluated the same set of results for the pre-selected query ranged from two to nine. Thus evaluations were collected by different people for the same queries drawn from the pre-selected set of queries, as well as a number of evaluations for the searches that users had defined themselves.

In the analysis that follows of the participants' different ratings, it was useful to quantify two things: 1) the quality of a particular ranking, given a rating, and 2) how similar two rankings are. The methods used to do this are described in greater detail below, and used throughout this chapter.

Measuring the Quality of a Ranking

For scoring the quality of a ranking, *Discounted Cumulative Gain* (DCG) was used. DCG is a measure of the quality of a ranked list of results commonly used in information retrieval research (Järvelin & Kekäläinen, 2000). It measures the result set quality by counting the number of relevant results returned. The idea that highly-ranked documents are worth more than lower-ranked documents is incorporated by weighting the value of a document's occurrence in the list inversely proportional to its rank (i). DCG also makes it possible to incorporate the notion of two relevance levels by giving *highly relevant* documents a different gain value than *relevant* documents.

$$\mathrm{DCG}(i) = \begin{cases} \mathrm{G}(1) & \text{if } i = 1\,, \\ \mathrm{DCG}(i-1) + \mathrm{G}(i)/\log(i) & \text{otherwise.} \end{cases} \qquad \textbf{(1)}$$

For *relevant* results, G(i)=1 was used, and for *highly relevant* results, G(i)=2, reflecting their relative importance. Because queries associated with higher numbers of relevant documents will have a higher DCG, the DCG was normalized to a value between 0 (the worst possible DCG given the ratings) and 1 (the best possible DCG given the ratings) to facilitate averaging over queries. In the analysis, different gain functions were explored (from G(i)=2 to G(i)=100, for highly relevant results), and alternative overall performance measures (e.g., percent of *relevant* or *highly relevant* documents in the top ten results). In almost all cases, the conclusions drawn using the normalized DCG measure with a gain of 2 reported here would be the same using these other measures.

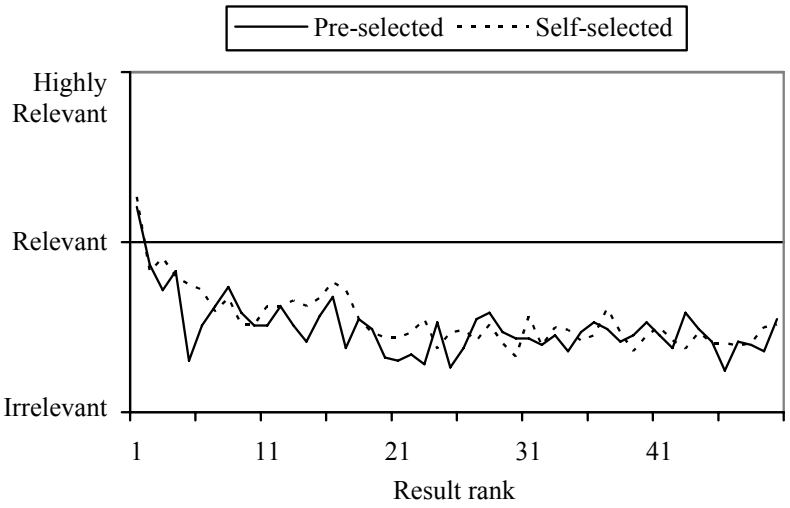


Figure 4-2. Average ratings for Web search engine results as a function of rank. There are many relevant results that do not rank in the top ten.

Measuring the Similarity of Two Rankings

To measure the "closeness" of two result rankings, the Kendall-Tau distance for partially ordered lists (Adler, 1957) was computed. The Kendall-Tau distance counts the number of pair-wise disagreements between two lists, and normalizes by the maximum possible disagreements. When the Kendall-Tau distance is 0, the two lists are exactly the same, and when it is 1, they are in reverse order. Two random lists have, on average, a distance of 0.5.

4.2.2 Rank and Rating

The data collected was used to study how the results that the Web search engine returned matched the participants' search goals. The results were expected to match the goals relatively closely, as current search engines seem to be doing well, and in recent years satisfaction with result quality has climbed.

Figure 4-2 shows the average result's relevancy score as a function of rank. To compute the relevancy score, the rating *irrelevant* was given a score of 0, *relevant* a score of 1, and *highly relevant* a score of 2. Values were averaged across all queries and all users. Separate curves are shown for the pre-selected (solid line) and self-selected (dashed line) queries. Clearly there is some relationship between rank and relevance. Both curves show higher than average relevance for results ranked at the top of the result list. The correlation between rank and relevance is -0.66. This correlation coefficient is significantly different from 0 ($p < 0.01$). However, the slope of the curves flattens out with increasing rank. When considering only ranks 21-50, the correlation coefficient is -0.07, which is not significantly different from 0. Importantly, there are still many relevant results at ranks 11-50, well beyond what users typically see. This suggests the search result ordering could be improved.

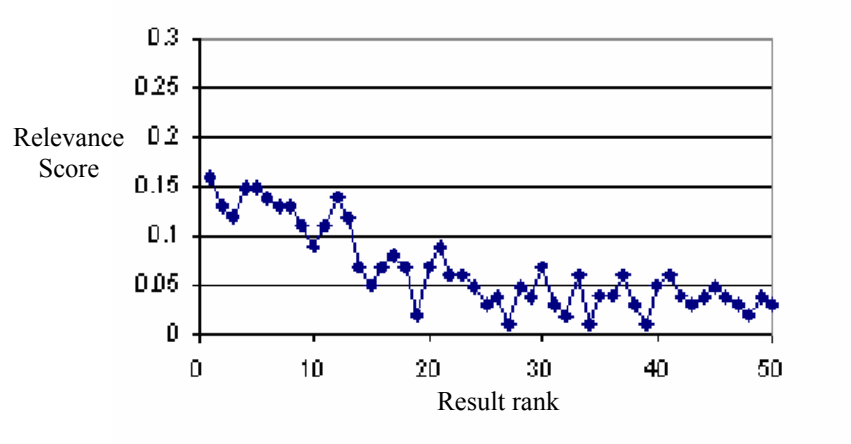


Figure 4-3. Average ratings for the TREC Web track results as a function of rank. The pattern is similar to what is seen in Figure 4-2.

The general pattern of results seen in Figure 4-2 is not unique to the sample of users or queries. A reanalysis of data from the TREC Web track (Hawking, 2000) shows a similar pattern (Figure 4-3). In the TREC-9 Web track, the top 100 results from 50 Web queries were rated using a similar three-valued scale, *highly relevant*, *relevant* and *not relevant*. Results for one top-performing search systems, uwmt9w10g3, yielded an overall correlation between rank and relevance of -0.81, which drops off substantially – to 0.30 for positions 21-50.

4.2.3 Same Query, Different Intents

The above analysis shows that rank and rating were not perfectly correlated. One reason for this is that while Web search engines may do a good job of ranking results to maximize their users' global happiness, they do not do as well for specific individuals. If everyone rated the same currently low-ranked documents as highly relevant, effort should be invested in improving the search engine's algorithm to rank those results more highly, thus making everyone happier. However, despite the many commonalities among the study participants (e.g., all were employees of the same company, lived in the same area, and had similar computer literacy), the study demonstrated a great deal of variation in their rating of results.

As will be discussed in the following sections, different participants were found to have rated the same results differently because they had different information goals or intentions associated with the same queries. This was evidenced by the variation in the explicit intents the participants wrote for their queries. Even when the intents they wrote were very similar, variation in ratings was observed, suggesting that the participants did not describe their intent to the level of detail required to distinguish their different goals.

Individuals Rate the Same Results Differently

Participants did not rate the same documents as relevant. The average inter-rater agreement for queries evaluated by more than one participant was 56%. This disparity in ratings stands in contrast to previous work. Although numbers can't be directly

compared, due to variation in the number of possible ratings and the size of the result set evaluated, inter-rater agreement appears to be substantially higher for TREC (e.g., greater than 94% (Koenmann & Belkin, 1996)) and previous studies of the Web (e.g., 85% (Eastman & Jansen, 2003)). The observed differences are likely a result of this study being focused on understanding personal intentions; instead of instructing participants to select what they thought was "relevant to the query," they were asked to select the results they would want to see personally.

The ratings for some queries agreed more than others, suggesting some queries might be less ambiguous to the study population than others. Similarly, some participants gave ratings that were similar to other participants' ratings. It might be possible to cluster individuals, but even the most highly correlated individuals showed significant differences.

Same Intent, Different Evaluations

Participants appeared to sometimes use the same query to mean very different things. For example, the explicit intents observed for the query *cancer* ranged from "information about cancer treatments" to "information about the astronomical/astrological sign of cancer". This was evident both for the pre-selected, where the user had to come up with an intent based on the query, and self-selected queries, where the query was generated to describe the intent. Although no duplicate self-selected queries were observed, many self-selected queries, like "rice" (described as "information about rice university"), and "rancho seco date" (described as "date rancho seco power plant was opened") were clearly ambiguous.

Interestingly, even when the participants expressed the same intent for the same query, they often rated the query results very differently. For example, for the query *Microsoft*, three participants expressed these similar intents:

- "information about microsoft, the company"
- "Things related to the Microsoft corporation"
- "Information on Microsoft Corp"

Despite the similarity of their intent, only one URL (www.microsoft.com) was given the same rating by all three individuals. Thirty-one of the 50 results were rated *relevant* or *highly relevant* by one of these three people, and for only six of those 31 did more than one rating agree. The average inter-rater agreement among these three users with similar intentions was 62%.

This disparity in rating likely arises because of ambiguity; the detailed intents people wrote were not very descriptive. Searches for a simple query term were often elaborated as "information on *query term*" ("UW" → "information about UW", leaving open whether they meant the University of Washington or the University of Wisconsin, or something else entirely). It appears participants had difficulty stating their intent, not only for the pre-selected queries, where it might be expected they would have some difficulty creating an intent (mitigated by the fact that they only rated pre-selected queries by choice), but also for the self-selected queries.

Although explicit intents generally did not fully explain the query term, they did provide some additional information. For example, "trailblazer" was expanded to "Information about the Chevrolet TrailBlazer", clarifying the participant was interested in the car, as opposed to, for example, the basketball team. Further study is necessary to determine why people did not include this additional information in their original query, but it does suggest that they could perhaps be encouraged to provide more information about their target when searching. However, even if they did this, they would probably still not be able to construct queries that expressed exactly what wanted. For example, the Trailblazer example above did not clarify exactly what kind of information (e.g., pricing or safety ratings) was sought. This suggests searchers either need help communicating their intent or that search systems should try to infer it.

4.2.4 Search Engines are for the Masses

The previous sections showed that participants ranked things very differently, in ways that did not correspond perfectly with the Web search engine ranking. This section describes analyses that show that the Web ranking did a better job of satisfying all of the participants than any individual.

Web Ranking Best for Group

The best possible group ranking that could be constructed based on the relevance assessments collected from different individuals for the same query were investigated, and this ideal ranking was compared with the original Web ranking.

If the Discounted Cumulative Gain value described earlier is taken to be a measure of ranking quality, the best possible ranking for a query given the data collected is the ranking with the highest DCG. For queries where only one participant evaluated the results, this means ranking *highly relevant* documents first, *relevant* documents next, and *irrelevant* documents last. When there are more than one set of ratings for a result list, the best ranking ranks first those results that have the highest collective gain.

The best possible group rankings were compared to the rankings the search engine returned using the Kendall-Tau distance described in Section 4.2.1. Recall that when the Kendall-Tau distance is 0, the two lists are exactly the same, and when it is 1, they are in reverse order. Two random lists have, on average, a distance of 0.5. It was found that for eight of the ten queries where multiple people evaluated the same result set, the Web ranking was more similar to best possible ranking for the group than it was, on average, to the best possible ranking for each individual. The average individual's best ranking was slightly closer to the Web ranking than random (0.5), with a distance of 0.469. The average group ranking was significantly closer ($p < 0.05$) to the Web ranking, with a distance of 0.440. The Web rankings seem to satisfy the group better than they do the individual.

Gains for Personalization via Re-Ranking

Again taking DCG as an approximation of user satisfaction, a sizeable difference was found between participants' satisfaction when given exactly what they wanted rather than the best group ranking for that query. On average, the best group ranking yielded a 23%

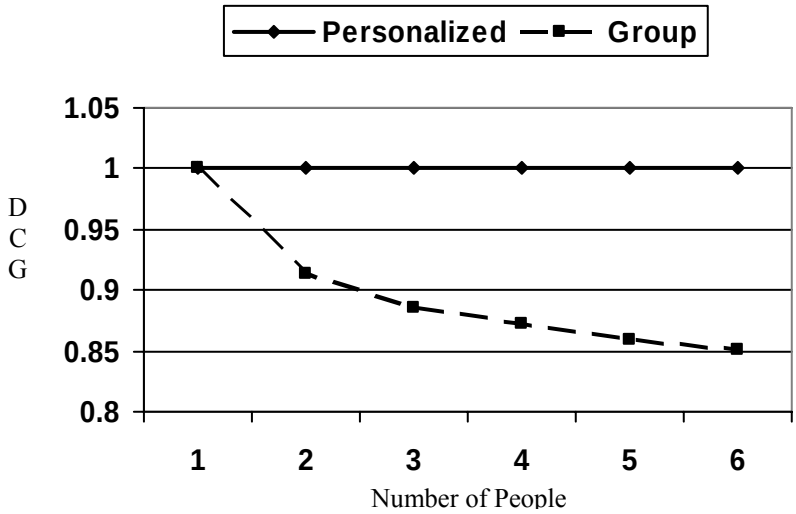


Figure 4-4. As more people are taken into account, the average DCG for each individual drops for the ideal group ranking, but remains constant for the ideal personalized ranking.

improvement in DCG over what the current Web ranking, while the best individual ranking led to a 38% improvement. Although the best group ranking does result in a sizeable improvement, it is likely as more and more people are considered in the group ranking (e.g., all of the users of Web search engines, rather than the fifteen participants in this study), the group ranking will yield a significantly smaller improvement. On the other hand, considering additional people is not likely to decrease the improvement seen for the best individual rankings.

The graph depicted in Figure 4-4 shows the average DCG for group (dashed line) or personalized (solid line) rankings. These data were derived from the five pre-selected queries for which six or more individual evaluations of the results were collected, although the pattern held for other sets of queries. To compute the values shown, for each query, one person was first randomly selected and the DCG found for that individual's best ranking. Additional people were then continually added, at each step re-computing the DCG for each individual's best rankings and for the best group ranking. As can be seen in Figure 4-4, as additional people were added to the analysis, the gap between user satisfaction with the individualized rankings and the group ranking grew. The sample is small, and it is likely that the best group ranking for a larger sample of users would result in even lower DCG values.

These analyses underscore the promise of providing users with better search result quality by personalizing results. Improving core search algorithms has been difficult, with research leading typically to very small improvements. This chapter has shown that rather than improving the generic results to a particular query, significant boosts can be obtained by working to improve results to match the intentions behind it.

If you don't find it in the index, look very carefully through the entire catalogue.
- Sears, Roebuck, & Co. (1897)

Chapter 5

Supporting Finding via Personalization

Based on the previous chapter, it is evident that Web search tools can be enhanced significantly by considering the variation in relevancy of results for users. This chapter presents a system, developed with Susan T. Dumais and Eric Horvitz, that supports finding through the personalization of Web search results (Teevan, Dumais, & Horvitz, 2005). Because, as seen in Chapter 3, people do not like to exert the cognitive effort to fully specify detailed information goals, and, as seen Chapter 4, they are not good at it even when they do, the personalized search tool presented here uses information about the searcher that can be gleaned in an automated manner to infer an implicit goal or intent.

This chapter explores the use of a very rich user profile, based both on search-related information such as previously issued queries and previously visited Web pages, and on other information such as documents and email the user has read and created. The research suggests that by treating the implicitly constructed user profile as a form of relevance feedback, it is possible to obtain better performance than explicit relevance feedback and improve on Web search. Although the most successful personalization algorithms rely both on a rich user profile and a rich corpus representation, it is possible to approximate the corpus and the text of the top-ranking documents based on the results returned by the Web search engine, making efficient client-side computation possible

5.1 Description of the Personalized Search System

Web search personalization was explored by modifying BM25, a well known probabilistic weighting scheme. BM25 ranks documents based on their probability of relevance given a query. The algorithm easily incorporates relevance feedback. Relevance feedback can be considered a very simple and short-term user profile, based on documents the user has selected as relevant to the particular query. More complex profiles are incorporated in the same manner that relevance feedback operates on the few documents identified by users as relevant.

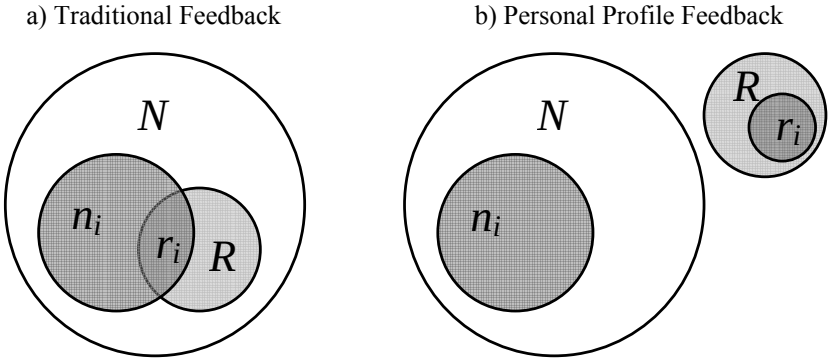


Figure 5-1. In traditional relevance feedback (a) relevance information (R, r_i) comes from the corpus. In the approach presented here to user profiling (b), profiles are derived from a personal store, so $N' = (N+R)$ and $n_i' = (n_i + r_i)$ is used to represent the corpus instead.

5.1.1 Corpus Representation

For a Web search engine to incorporate information about a user, a user profile must either be communicated to the server where the Web corpus resides or information about the results must be downloaded to the client machine where a user profile is stored. This approach focuses on the latter case, on re-ranking the top search results locally, for several reasons. For one, such a methodology ensures privacy; users may be uncomfortable with having personal information broadcast across the Internet to a search engine, or other uncertain destinations. Second, in the re-ranking paradigm, it is feasible to include computationally-intensive procedures because only a relatively small set of documents are worked with at any time. Third, re-ranking methods facilitate straightforward evaluation. To explore re-ranking, only ratings for the top-k returned documents need be collected, instead of undertaking the infeasible task of collecting evaluations for all documents on the Web. Within the re-ranking framework, lightweight user models that could be collected on the server side or sent to the server as query expansions are also explored.

Web search personalization is explored by modifying BM25 (Sparck Jones, Walker, & Robertson, 1998), a well known probabilistic weighting scheme. BM25 ranks documents based on their probability of relevance given a query. In use, the method essentially sums over query terms the log odds of the query terms occurring in relevant and non-relevant documents. The algorithm easily incorporates relevance feedback. Relevance feedback can be considered a very simple and short-term user profile, based on documents the user has selected as relevant to the particular query. More complex profiles are incorporated in the same manner that relevance feedback operates on the few documents identified by users as relevant.

This section focuses briefly on additional details of BM25. The method ranks documents by summing over terms of interest the product of the term weight (w_i) and the frequency with which that term appears in the document (tf_i). When no relevance information is available, the term weight for term i is:

$$w_i = \log \frac{N}{n_i} \quad \textbf{(2)}$$

where N is the number of documents in the corpus, and n_i is the number of documents in the corpus that contain the term i.

When relevance information is available, two additional parameters are used to calculate the weight for each term. R is the number of documents for which relevance feedback has been provided, and r_i is the number of these documents that contain the term. As shown graphically in Figure 5-1 (a), the term weight in traditional feedback is modified to:

$$w_i = \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)} \quad \textbf{(3)}$$

In the approach presented here, the explicit relevance judgments on the documents returned for each query are not available. Instead, relevance is inferred based on a local store of information that is considered to be implicitly relevant. (How the user profile is collected is described in more detail below.) The local user profile contains information that is not in the Web corpus, and this requires an extension to traditional relevance feedback models.

Figure 5-1 (b) shows graphically how one can conceptualize using information outside of the Web corpus for relevance feedback as pulling the relevant documents outside of the document space. Thus the notion of corpus must be extended for the purpose of BM25 weighting to include the outside documents. The variables $N' = (N+R)$ and $n_i' = (n_i + r_i)$ are used to represent the corpus instead. Substituting these values into the previous equation and simplifying results in the following equation for the term weights:

$$w_i = \frac{(r_i + 0.5)(N - n_i + 0.5)}{(n_i + 0.5)(R - r_i + 0.5)} \quad \textbf{(4)}$$

To personalize search results, the similarity of a query to a document is computed, summing these term weights over all terms in the query (or expanded query).

There are a number of different ways to represent the corpus, the user profile, and different approaches to selecting the query terms that are summed over. This chapter explores several different approaches, summarized in Table 5-1, in greater detail below. Three important components of the model are: Corpus Representation (how to obtain estimates for N and n_i); User Representation (how to obtain estimates for R and r_i), and Document/Query Representation (what terms are summed over).

5.1.2 User Representation

To represent a user a rich index of personal content is employed that captured a user's interests and computational activities. Such a representation could be obtained from a desktop index such as that described in Stuff I've Seen (Dumais et al., 2003) or available in desktop indices such as Copernic, Google Desktop Search, Mac Tiger, Windows

Table 5-1. Summary of differences between personalization variables. Significant differences (p<0.01) are marked with <, weakly significant differences (p<0.05) with '≤', and non-significant differences are marked as equal.

Corpus Representation (N, n_i)
Full text of documents in result set < Web < Snippet text in result set Query focused = Based on all documents
User Representation (R, r_i)
No user model = Query history ≤ Indexed Web documents < Recently indexed < Full index Query focused = Based on all documents
Document Representation (terms *i* summed over)
Snippet text < Full document text Words near query terms < All words in document

Desktop Search, Yahoo Desktop Search or X1. The system used indexed all of the information created, copied, or viewed by a user. Indexed content includes Web pages that the user viewed, email messages that were viewed or sent, calendar items, and documents stored on the client machine. All of this information can be used to create a rich but unstructured profile of the user. The most straightforward way to use this index is to treat every document in it as a source of evidence about the user's interests, independent of the query. Thus, R is the number of documents in the index, and r_i is the number of documents in the index that contain term i. As in the case of the corpus representation, the user profile can also be query focused, with R representing instead the number of documents in the user's index that match the user's query, and r_i, the subset that also contains term i.

Several techniques were experimented with for using subsets of a user's index (each which could either be query focused or query independent) to compute R and r_i. For example, the value of considering all document types (e.g., email messages, office documents, and Web pages) versus restricting the document type to only Web pages was explored. The motivation for exploring such a restriction is that the statistical properties of the terms might be significantly different in the user's full index than on the Web because of inherent differences in word frequencies associated with different types of information. As another class of restriction along a different dimension, documents were limited to the most recent ones. Because a user's interests may change over time, documents created or viewed more recently may give a better indication of a user's current interests than older documents. In the general case, the time sensitivity of representations of a user's interests may differ by document type, and it may be possible to draw from a user's index combinations of different types of documents, each restricted to different time horizons. In the studies presented here, the only analysis is of the value of considering documents indexed in the last month versus the full index of documents.

Beyond analysis of the user's personal index, two lighter-weight representations of the user's interests were considered. For one, the query terms that the user had issued in the past were used. For the other, motivated by the importance of the information source shown in Chapter 3, the search results with URLs from domains that the user had visited in past were boosted. Results associated with URLs where the last three components of

the URL's domain name (e.g., http://www.***csail.mit.edu***) matched a previously visited URL were boosted to the top, followed by those where the last two components matched (e.g., http://www.csail.***mit.edu***). Both of these methods for representing a user's interests could easily be collected on servers hosting search services.

5.1.3 Document and Query Representation

The document representation is important in determining both what terms (i) are included and how often they occur (tf_i). Using the full text of documents in the results set is a natural starting place. However, accessing the full text of each document takes considerable time. Thus, using only the title and the snippet of the document returned by the Web search engine was also experimented with. Because the Web search engine used derived its snippets based on the query terms, the snippet is inherently query focused.

In the absence of any information other than the user's query, a document's score is calculated by summing over the query terms, the product of the query term weight (w_i) and the query term occurrence in the document (tf_i). However, when relevance feedback is used, it is very common to use some form of query expansion. A straightforward approach to query expansion that was experimented with is the inclusion of all of the terms occurring in the relevant documents. This is a kind of blind or *pseudo-relevance* feedback in which the top-k documents are considered relevant (Ruthven & Lalmas, 2003). Thus, for the query "cancer", if a document contained the following words,

> The American Cancer Society is dedicated to eliminating cancer as a major health problem by preventing cancer, saving lives, and diminishing suffering through...

each word would affect the document score. To maintain a degree of emphasis on the query, a subset of terms that were relevant to the query were also selected from the documents. This was done in a simple manner, by including the words that occurred near the query term. For example, the following underlined terms would be selected from the previous snippet:

> The American Cancer Society is dedicated to eliminating cancer as a major health problem by preventing cancer, saving lives, and diminishing suffering through...

To summarize, several different techniques were explored for representing the corpus, the user, and the documents. These include *Corpus Representation*: Counts derived from the Web or from the returned set for estimating N and n_i; *User Representation*: Counts from the full index, temporal subsets or type subsets for estimating R and r_i; and *Document/Query Representation*: Words obtained from the full text or snippets of documents, and words at different distances from the query terms.

5.2 Performance of Personalized Search

Sixty-seven different combinations were looked at of how the corpus, users, and documents could be represented, as discussed above, and used these combinations to re-rank Web search results. Several different baselines were also explored.

This section first present the results of ranking the top fifty documents for a query based purely on their textual features, ignoring the ranking returned by the Web search engine. The best parameter settings for personalizing the search results are compared with several baselines. Then augmentations of the personalized content-based rankings that incorporate the Web ranking are reported on. Web rankings take into account many factors, including textual content, anchor text, and query-independent importance factors such as PageRank.

5.2.1 Alternative Representations

The many different combinations of corpus, user, and document representations explored resulted in a complex experimental design. For ease of presentation, this section first summarizes one-way effects in which all but one variable are held constant, and explores the effects of varying that variable (e.g., User Representation – No model, Queries, Web pages, Recent index, Full index). This approach does not examine interaction effects, so at the end of this section, the findings from the best combination of variables are summarized.

Results of the one-way analyses are shown in Figure 5-2. The scores reported in Figure 5-2 are the normalized DCG (recall Equation **(1)**) for the 131 queries in the test set presented in Chapter 4, averaged across levels of the other variables. Analyses for statistical significance were performed using two-tailed paired *t*-tests. The key effects are summarized in Table 5-1, along with their statistical significance levels.

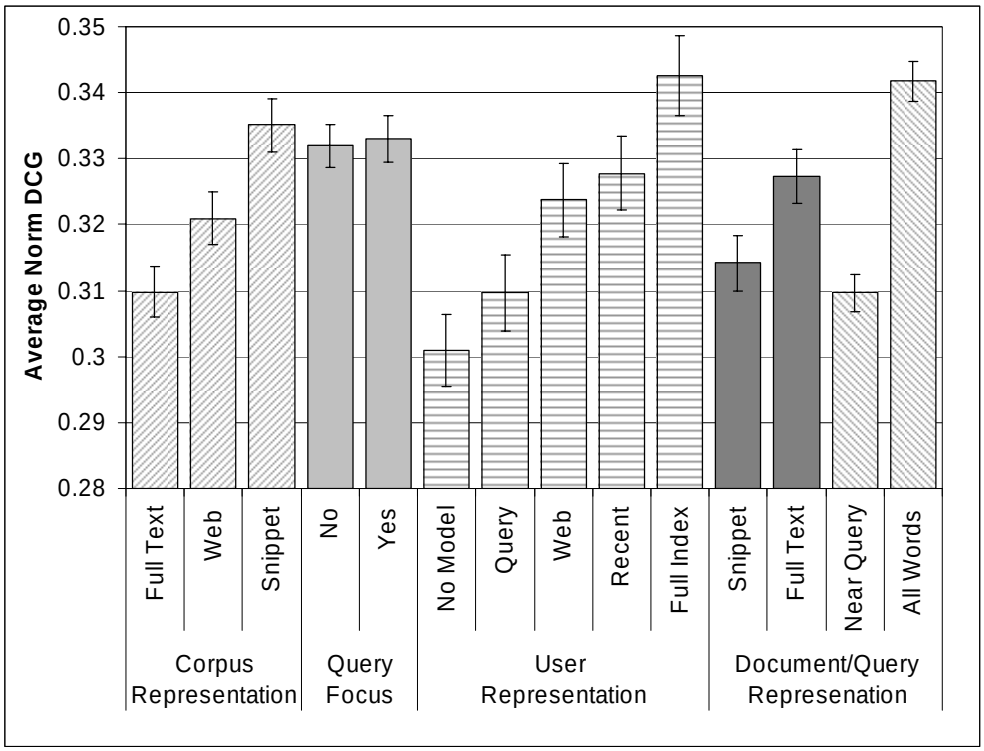


Figure 5-2. Average normalized DCG for different variables, shown with error bars representing the standard error about the mean. Richer representations tend to perform better.

The one-way sensitivity analysis showed that a rich representation of both the user and the corpus was important. The more data used to represent the user (*User Representation* in Figure 5-2), the better. Performance with the user's entire desktop index was best (*Full Index*), followed by representations based on subsets of the index (*Recent*ly indexed content, and *Web* pages only). Using only the user's query history (*Query*) or no user-specific representation (*No Model*) did not perform as well. Similarly, the richer the document and query representation (*Document/Query Representation*), the better the performance. It appeared that the best personalization occurred when using the full document text to represent the document (*Full Tex*t), rather than its snippet (*Snippet*), and performance was better when using all words in the document (*All Words*) than when using only the words immediately surrounding the query terms (*Near Query*).

The only place that more information did not improve the ranking was in the corpus representation (*Corpus Representation*). Representing the corpus based only on the title and snippets of the documents in the result set (*Snippet*) performed the best. An explanation for this finding is that when using only documents related to the query to represent the corpus, the term weights represent how different the user is from the average person who submits the query. It was interesting to find that using the result set statistics performs both better (when the title and snippet are used) and worse (when the *Full Text* is used) than the Web (*Web*). One potential advantage to using the title and snippets to represent the corpus instead of using the full text is that the snippets are relatively uniform in length relative to the full text of the documents. Thus, one or two long documents do not dominate the corpus statistics. Another possible advantage of using the snippets is that they extract query-relevant portions of the document, which is important for documents that cover more than one topic.

Because of the query expansion performed during the result re-ranking, the user's query does not necessarily play a significant role in re-ranking. However, emphasizing the query during re-ranking does not appear to be necessary. Using all terms for query expansion was significantly better than using only the terms immediately surrounding the user's query (*Document/Query Representation*, *All Words* vs. *Near Query*). Using query focused corpus and user representations (*Query Focus*, *Yes*) showed no significant difference from a non-query focused representations (*Query Focus*, *No*). It could be that a query focus provides some benefit, but that the tradeoff between having a query focus and using more information in the ranking is relatively balanced. Alternatively, the lack of importance of the query could be because all of the documents being ranked are more or less relevant to the query, making the primary job of the personalization to match the user.

These results indicate that, although the corpus can be approximated, a rich document representation and a rich user representation are both important. In practice, a system must choose between performing personalization on the client, where the rich user representation resides, or on the server side, where rich document representations reside. Thus, the interaction between parameters was also looked at. It appears it was more important to have a rich user profile than to have a rich document representation.

Because the parameters interact, the relationship is somewhat more complex than suggested by the results of the one-way analyses reported in Figure 5-2. The best combination of parameters found was:

Corpus Representation: Approximated by the result set title and snippets, which is inherently query focused.

User Representation: Built from the user's entire personal index, query focused.

Document and Query Representation: Documents represented by the title and snippet returned by the search engine, with query expansion based on words that occur near the query term.

This parameter combination received a normalized DCG of 0.46, and was the best combination selected for each query using leave-one-out cross-validation.

The corpus and user representations for the best parameter combination are consistent with what was found in the one-way sensitivity analyses. However, the document representation differs. The best combination calls for documents to be represented by their titles and snippets, rather than their full text. This makes sense given that the corpus representation is based on the documents titles and snippets as well. Corpus statistics are not available for terms that appear in the full text but not the snippet.

In addition to performing well, this combination is easy to implement entirely on the client's machine, requiring only the download of the search engine results. Thus, it is investigated further in comparison with several non-personalized baseline conditions.

5.2.2 Baseline Comparisons

To assess how well personalization performed, the results were also compared with several key baselines. These comparisons can be seen in Figure 5-3. The scores reported in Figure 5-3 are the normalized DCG for the 131 queries in the test set, and statistical analyses were performed using two-tailed paired *t*-tests with 130 degrees of freedom. The results of the best personalized search algorithm are shown in the bar labeled *PS*. The baseline conditions include: random ordering of the top-50 results (*Rand*), no user model (*No*), and an idealized version of relevance feedback (*RF*). For the cases of no user model and relevance feedback, a BM25 ranking based on the same content as the personalized search algorithms was used, with the best corpus and document/query representation selected for each. Only the user representation (R, r_i) differed. In the no user model case, R and r_i were equal to zero, and for the relevance feedback case, they were based on the documents in the evaluation test set that the user marked as *highly relevant* or *relevant*.

Not surprisingly, personalized search re-ranking (*PS*) significantly outperformed a random ordering of search results (*Rand,* $p<0.01$), and search with no user model (*No,* $p<0.01$). It was somewhat surprising that Web search personalization also performed somewhat better than ideal relevance feedback (*RF,* $p<0.05$). While this result may seem counterintuitive, it is important to note that in relevance feedback, the relevant documents are used to expand the terms considered and to modify the term weights. This does not guarantee that the documents used for augmentation will be at the top of the re-ranked list. In addition, the rich user profile used in *PS* may contain useful discriminating terms that are not present in the relevant documents in the top-50 results.

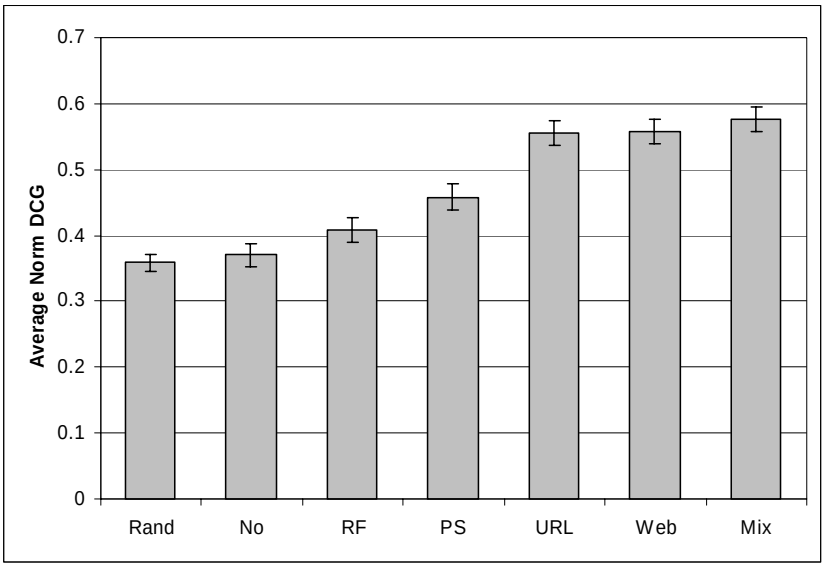


Figure 5-3. Personalized search (PS) compared with a random ranking (Rand), no user model (No), relevance feedback (RF), URL boost (URL), the Web (Web), and personalized search combined with the Web (Mix).

Figure 5-3 also shows a comparison of the best personalized content-based rankings with the Web ranking (*Web*). Personalized search performed significantly worse ($p<0.01$) than the Web rank, which had a normalized DCG of 0.56. This is probably because Web search engines use information about the documents they index in addition to the text properties, most notably linkage information, and this has been shown to improve results for many search tasks, including those of finding homepages finding and identifying resources (Hawking & Craswell, 2001). For this reason, incorporating the ranking information returned by the search engine into the personalization algorithm was explored.

5.2.3 Combining Rankings

To understand whether there was potential value in combining Web rankings and personalized results, the similarity between these two rankings and between these rankings and the user's ideal ranking was explored. To construct the user's ideal ranking, the documents the user considered *highly relevant* were ranked first, *relevant* next, and *not relevant* last. Because such a ranking does not yield a completely ordered list, the Kendall-Tau distance for partially ordered lists (Adler, 1957) was computed to measure the similarity of rankings. As mentioned in Chapter 4, the Kendall-Tau distance counts the number of pair-wise disagreements between two lists, and normalizes by the maximum possible disagreements. When the Kendall-Tau distance is 0, the two lists are exactly the same, and when it is 1, the lists are in reverse order. Two random lists have, on average, a distance of 0.5. The personalized ranking was significantly closer to the ideal than it was to the Web ranking ($\tau = 0.45$ vs. $\tau = 0.49$, $p<0.01$). Similarly, the Web ranking was significantly closer to the ideal than to the personalized ranking ($\tau = 0.45$ vs. $\tau = 0.49$, $p<0.05$). The finding that the Web ranking and the personalized ranking were

both closer to the ideal than to each other suggests that what is good about each list is different.

In an attempt to take advantage of the best of both lists, the Web ranking was merged with the personalized text-based ranking. For the personalized results, the BM25 match scores could be used. For the Web results, only the rank information was available. For this reason, only rank was considered in the merge. To merge the two lists, each position was weighted in accordance with the probability that a result at that position from the Web is relevant. This probability was computed based on all queries in the test set except the one being evaluated. Because the probability curve typically drops quickly after the first couple of results, merging results in this way has the effect of keeping the first couple of results similar to the Web, while more heavily personalizing the results further down the list. The combination of the Web ranking and the personalized ranking (*Mix* in Figure 5-3) yielded an average normalized DCG of 0.58, a small but significant ($p<0.05$) improvement over the Web's average normalized DCG of 0.56. In contrast, boosting previously visited URLs by merging them with the Web results (*URL*) yielded no significant change to the Web's average normalized DCG.

Note that the personalization algorithm used in this analysis was not selected because it produced the best results when merged with the Web ranking, but because it performed well on its own and is feasible to implement. Greater improvement might be obtained by selecting the parameter setting that produces the best results when merged with the Web ranking. It is also likely that the advantages of personalization could further be improved by combining server and client profiles in richer ways. For example, in the same way that content-based matching was personalized, link-based computations could be personalized as proposed by Jeh and Widom (2003). Richer application programming interfaces (APIs) to Web indices could also provide richer corpus or document statistics, further improving the ability to personalize both content-based and query-independent factors in ranking.

5.3 Conclusion

The chapters of Part I have explored finding behavior. A naturalistic study of the search strategies people use demonstrated that orienteering is common, and suggested a number of potential benefits of orienteering over teleporting. Although search engines were not used to teleport, they were commonly used as a step in orienteering, and the range of information goals associated with queries were characterized. Because participants understood diverse intents for the same queries, there appeared to be much potential value to obtain for individual search results by personalizing search results.

This chapter investigated the feasibility of personalizing Web search by using an automatically constructed user profile as relevance feedback in the ranking algorithm. The research suggests that the most successful text-based personalization algorithms perform significantly better than explicit relevance feedback where the user has fully specified the relevant documents, and that combining this algorithm with the Web ranking yields a small but statistically significant improvement over the default Web ranking.

PART II

RE-FINDING

Nothing endures but change.
- Heraclitus (540 BCE - 480 BCE)

Chapter 6

Introduction to Re-Finding

Up until now, the focus of the research presented in this dissertation has been on the finding of information in general. However, re-finding previously viewed information is a particularly important and common aspect of finding. Most of the information people interact with on a regular basis has been seen before, and yet people still have problems getting back to it. In a study of Web users (GVU, 1998), 17% of those surveyed reported "Not being able to return to a page I once visited," as one of "the biggest problems in using the Web."

While many search engines have begun to address the issue of re-finding by, for example, caching query history, these efforts are just a beginning. Most search tools focus solely on supporting the finding of new information. Tools that explicitly account for the finding of previously viewed information as well as new information are likely to significantly improve people's ability to find in general. Such tools may also positively impact people's organizational habits. One of the primary reasons time is invested in information organization is to make the information easy to re-find and reuse. Successful re-finding support may lead to a reduction in the difficulty and effort required to archive and organize information in the first place.

The purpose of the chapters in Part II is to give an understanding of the factors that influence re-finding activities and present a search tool designed to support both the finding of new information and the re-finding of previously viewed information. The research in Part I, through an exploration of finding behavior, presented a way for search tools to identify via personalization the most relevant information to a person's query. The analysis of re-finding presented in this part reveals, however, that presenting the most relevant information without accounting for the individual's previous information interactions can interfere with re-finding. For example, when, in Chapter 2, Connie repeated her search for "breast cancer treatments", valuable new information was available – new treatments had become available and search result personalization identified more relevant results for Connie. Naively presenting this valuable new information without accounting for Connie's past interactions made it difficult for her to re-find the list of treatment options she found during her initial search because the result no longer appeared where she expected it. Through the research in this part, a solution to

the apparent conflict between finding new information and old information is presented and tested.

This initial chapter is intended to place the research presented in Part II in context. The chapter begins with a brief overview of the other chapters included in this part. It then presents related research on understanding re-finding behavior and dynamic information interaction, and discusses existing systems that are intended to support both finding and re-finding. This review of related work is important to properly place the research presented in the rest of Part II in context.

6.1 Outline of Part II

Including this chapter, Part II consists of four chapters:

6. **Introduction to Re-Finding**
7. **Understanding Re-Finding**
8. **Why Re-Finding Requires Personalization**
9. **Supporting Finding and Re-Finding through Personalization**

Thus far much attention has been paid to the finding of new content, while re-finding has not been significantly addressed. In **Chapter 6 (Understanding Re-Finding)**, the focus shifts to understanding re-finding behavior, to give an idea of how prevalent the behavior is and what it looks like. Some of the differences between finding and re-finding that appeared in the modified diary study presented in Chapters 3 are highlighted. Analysis then narrows in on the use of keyword search, and search engine queries are analyzed with the goal of understanding re-finding. The queries studied include those issued to Yahoo by 114 users over the course of a year, and those collected via a controlled study of searches issued 119 users. Repeat searches are revealed to be very common, with 33% of all queries issued by an individual occurring in the logs more than once, and 40% of all queries involving a click on a search result that was clicked during another search by the same user. Characteristics of repeat searches are discussed, and it is shown to often be possible to predict whether a person is searching for new information or old information based on the query and past query interactions.

Although finding and re-finding tasks may require different strategies, tools will need to seamlessly support both activities. However, finding and re-finding can be in conflict – finding new information means retrieving the best new information, while re-finding previously viewed information means retrieving the previously viewed information being sought. In **Chapter 7 (Why Re-Finding Requires Personalization)**, this conflict is explored and it is shown that an individual's previous information interactions are important to account for to properly support re-finding. Further analysis of the query log data introduced in Chapter 6 suggests changes to results lists between interactions can significantly interfere with re-finding. A change in rank significantly reduced the likelihood of a repeat click and slowed repeat clicks even when they did happen. However, based both on the log analysis and on ten queries tracked over several years, it

is apparent that search results change regularly – bringing people better new information, but making the common behavior of returning more difficult.

Chapter 7 further explores the difficulties created for re-finding by information change by analyzing instances where people reported encountering such problems. Web pages, collected via a Web search, are analyzed where the phrase, "Where'd it go?" was used. A number of observations arise from the analysis, including that the path originally used to locate the information target appeared very memorable, whereas the temporal aspects of when the information had been seen before were only important when relevant to other events. People expressed a lot of frustration when problems arose, and often sought an explanation of why they could not find their target, even in the absence of a solution.

Chapter 8 (Supporting Finding and Re-Finding through Personalization) presents a system, the Re:Search Engine, that addresses the conflicting goals of providing new information while maintaining an environment that matches user expectation. The engine consists of a Web browser toolbar plug-in that interfaces with a preexisting search engine, such as Google, Yahoo, or a personalized search service. When a person issues a query that they have issued before, the Re:Search Engine fetches the current results for that query from the underlying search engine. The newly available information is then seamlessly merged with what the user remembers about the previously returned search results. The algorithm underlying the Re:Search Engine is based on a study of what 119 people found memorable about search result lists. This study was used to build a model of which aspects of a result list are memorable, and thus should be changed with care, and which are not, and can change freely.

Whether the Re:Search Engine can successfully incorporate new information into a previously viewed search result list is tested via a study of 165 people. In this study, participants were asked whether they noticed changes that either occurred in accordance with the model or not. Changes made according to the model were less likely to be noticed than other changes. Further, the study reinforced that apparent consistency is important. Even when the new search results were of higher quality, if the participants noticed a change, they viewed the changed result quality as worse than the original quality. A subsequent study of 30 people suggests the Re:Search Engine makes re-finding virtually as easy as if the results had not changed, without interfering with the individual's ability to find new information.

The rest of this chapter highlights the related work necessary to understand the research described above and presented in greater depth in the following chapters of Part II. First related research that has been conducted to understand re-finding behavior is discussed, followed by a discussion of dynamic information interaction. The chapter concludes with an overview of related search personalization systems.

6.2 Related Work on Understanding Re-Finding

This section discusses related research that that has been conducted to understand re-finding behavior. Many studies of finding consider re-finding of some form. For example, most of the searches a person runs within their email or file system are for documents that have been seen before. This section emphasizes research that focuses on

the unique aspects of re-finding. It first discusses what makes re-finding different from finding and highlights the prevalence of re-finding on the Web. It then focuses on several factors that affect re-finding in particular, and shows that how information is kept and organized affects re-finding. It concludes with a discussion of how it is hard to structure information to support future re-finding because it is hard to predict the future value of information at the time it is encountered.

6.2.1 Re-Finding is Different from Finding New Information

It is difficult to identify pure re-finding behavior, as opposed to behavior intended to find new information. There exists a continuum ranging from true finding tasks where little is known about the information target to well known, frequent, re-finding tasks. As a result, searches for information that has been seen before are often similar to searches for new information, but can also can differ greatly.

Despite the fact that some re-finding behavior falls very close to the finding of new information, there are many factors that could affect a person's ability to relocate and reuse the information that make re-finding in general different from the finding of new information. One distinguishing feature of re-finding is that the searcher often knows a lot of meta-information about the target, such as its author, title, date created, URL, color, or style of text. For example, when Connie, in the example in the introduction, wanted to re-find the list of treatment options she found during her first Web search, she knew the information she was looking for could be found on an About.com Web page and that the page was found via a Web search. Several types of meta-information seem particularly important for re-finding: the people associated with the target (Dumais et al., 2003), the path taken to find the information (Capra & Pérez-Quiñones, 2005a; Teevan, 2004), and temporal aspects of the information (Lansdale & Edmonds, 1992; Ringel et al., 2003). Lansdale and Edmonds (1992) argue time is so important that the default ranking for information retrieval should be chronological, and chronology does indeed turn out to be a dominant ranking factor for the *Stuff I've Seen* search engine (Dumais et al., 2003). Some of the meta-information used in re-finding is self generated, and thus particularly easy to remember. If Connie had emailed herself the URL of the Web page she liked, for example, she might have a good idea as what she titled that email. Search failure during re-finding appears to be particularly frustrating in part because the information sought has been seen before and is known to exist.

6.2.2 Re-Finding is Common

The differences between finding and re-finding are important to understand because people commonly return to information they have seen before (Byrne et al., 1999; Cockburn et al., 2003; Tauscher & Greenberg, 1997). Tauscher and Greenberg analyzed six weeks of detailed Web usage logs and found that 58% of all Web page visits were re-visits (Tauscher & Greenberg, 1997). In a more recent study by Cockburn et al. (2003), users were found to revisit Web pages for as many as four out of every five page visits. And while many of the Web page re-visitations occur shortly after a Web page is first visited (e.g., during the same session by using the back button), a significant number are visited after considerable time has elapsed. In their study, Tauscher and Greenberg report, "Users revisit pages that have not been accessed recently. For example, 15% of

recurrences are not covered by a list of the last 10 URLs visited. Still, doubling or tripling the size of this list does not increase its coverage much (Tauscher & Greenberg, 1997)." Indeed, they found 6% of all Web page visits take place to pages that haven't been visited in over fifty visits. Chapter 7 shows that search engines are commonly used to support re-finding and re-visitation.

6.2.3 Factors that Affect Re-Finding

Because people develop knowledge about a piece of information during their original encounter with that information, there are influencing factors that are important to consider for re-finding in addition to the factors that influenced general finding. These factors include as the amount of time that has elapsed from when the information was initially found, the perceived versus actual future value of information when found, the similarity of the initial finding and re-finding tasks, whether the location or context of the information changed, and the fungibility of the information source needed (i.e., is the exact same source necessary?).

Capra and Pérez-Quiñones (2005a) conducted a controlled laboratory study to examine many of the factors that affect re-finding behavior, including information type, task type, and task familiarity. In the study, 17 individual participants completed two experimental sessions held approximately one week apart. In the first session, participants found information on the Web for a set of 18 tasks. The tasks were primarily well-defined, directed information seeking tasks (e.g., "Find the phone number of the Kroger grocery store on Main Street," or "Find a Web page that describes how to solve the Rubik's cube,"), but included several less defined tasks as well (e.g., "Find two sweatshirts that you would like to buy for a friend,"). In the second session, participants were asked to perform tasks that involved re-finding the same or similar information that was found in the initial session (e.g., "Find phone number of the Kroger grocery store on University Boulevard," or "Find a Web page that describes how to solve the Rubik's cube,").

Results indicated that users have strong patterns for information access and that they are likely to use these patterns when re-finding. This result is supported by research conducted by Aula, Jhaveri, and Kaki (2005), and by Capra and Pérez-Quiñones (2005b). For example, users often approached re-finding tasks using the same starting Web page that they used to originally find the information. The study by Capra and Pérez-Quiñones (2005b) revealed that the frequency with which a user previously performed similar tasks had significant effects: high frequency tasks were completed more quickly, involved fewer URLs, and involved less use of Web search engines. However, search engine use did not differ significantly between the two sessions, suggesting that the use of search engines may be strongly linked to specific tasks and not influenced by whether the searcher is finding or re-finding. The results also suggest that, as for finding, keyword search is not a universal solution for re-finding. Instead, participants used a variety of methods to re-find information, including the use of waypoints and path retracing.

6.2.4 How Information is Kept and Organized Affects Re-finding

Another particularly important influencing factor in re-finding is how information is kept and organized. In a study where participants were cued to return to selected Web sites

after a delay of three to six months, Bruce, Jones, and Dumais (Bruce, Jones & Dumais, 2004; Jones, Bruce, & Dumais, 2003; Jones, Dumais, & Bruce, 2002) report a strong preference for methods of returning to Web sites that require no overt prior keeping action. These methods include searching again, following hyperlinks from another Web site, or accepting the Web browser's suggested completion to a partially entered Web address. Nevertheless, for Web sites and for other many other forms of information, people often expend considerable effort making their information available for future use, and have different strategies for doing so. These strategies can affect people's search strategies. For example, Chapter 4 shows that pilers prefer to orienteer with small steps, while filers are more likely to teleport or orienteer with large steps using keyword search.

Organizing information for re-finding is hard because the future value of information is hard to predict, and people regularly misjudge the difficulty they'll have returning to information. Often, the value of encountered information is not realized until well after it is originally encountered – a phenomenon referred to as post-valued recall (Wen, 2003). For example, Connie could have made re-finding the list of treatment options she researched initially easy by bookmarking it, but she did not consider the page she found to be of high enough value at the time she first encountered it. Some tools have been developed to help bring potentially relevant personal information to the user's attention (Dumais et al., 2004), but additional study is necessary to understand when people forget important information and when and how they want to be reminded of it.

Chapter 2 discussed how people often structure their personal information stores to support reminding and recognition. Their fear of forgetting what they believe to be important information can even lead to behaviors such as emailing information to oneself (Jones, Bruce, & Dumais, 2001) or not filing email messages (Whittaker & Sidner, 1996) as a way to support re-finding it later. However, just as it is hard to decide what information is important to keep, it can be difficult to organize and classify information believed to be important because the future value and role of the information is not fully understood. People's difficulty classifying can cause problems retrieving. Lansdale (1988) noted that people had difficulty retrieving information when they were forced to group their information into categories that were not necessarily relevant for retrieval. Some tools have been developed to help bring potentially relevant personal information to the user's attention (Dumais et al., 2004), but additional study needs to be done understand when people forget about important information and when and how they want to be reminded of it.

6.3 Related Work on Dynamic Information Interaction

Confounding many of the factors that affect re-finding is that fact that a person's electronic information often changes. Search results change as new Web resources are discovered (Ntoulas, Choo, & Olston, 2004). The online news stories an individual has read change when new stories are written, and the list of emails in a person's Inbox changes as new emails arrive. Changes to electronic information help people find better search results, new emails and recent news stories; Chapter 7, for example, shows that people are likely to click on new results, as well as old results, when they repeat queries. But, as shown in Chapter 8, changes can also interfere with re-finding.

People have developed coping strategies to deal with information change (Jones, Bruce, & Dumais, 2003; Marshall & Bly, 2005), but such strategies are cumbersome and have many flaws. Rather than requiring a person to cope, change should be introduced in a manner that does not interrupt that person's interactions with previously encountered information. This section presents related work that suggests Web information and search results change, and that such changes can cause problems re-finding.

6.3.1 Web Information and Search Results Change

The Web is tremendously dynamic. Examples of particularly dynamic Web content include message boards, personal Web pages, stock quotes, and Internet auctions. One reason Web information changes is because it is controlled by other people. For example, a moderator could delete an inappropriate message from a message board, or someone could edit their personal Web page. Since any one individual only controls an extremely small portion of the Web, a majority of the Web is controlled by other agents and is likely to change outside of any one individual's control. Additionally, the Web makes a considerable amount of time dependent information, such as stock prices and Internet actions, available.

How the Web changes has been looked at through many studies (Dalal et al, 2004; Davis, Masloy, & Phillips, 2005; Koehler, 2002). Fetterly et al. (2003) have done several studies looking at how the Web evolves over time through a series of large-scale Web craws that track the evolution of 150 million Web pages over almost three months, with the interest of informing search result crawls. They found that fewer Web pages change significantly less than previously believed, and that past change is highly predictive of future change.

Chapter 8 explores how common changes to search results are. Search results change as the Web changes, because search engines update their indices to reflect updates on the Web. As new information becomes available, this new information may be returned for old queries, or may influence the engines judgment of what old information is the most relevant to the query relevant. Results can also change as search engine algorithms change – as a result of global improvements, improvements, collaborative filtering, personalization, or relevance feedback.

Similar to the work presented in Chapter 8, Selberg and Etzioni (2000) studied the rate of change of search result lists specifically to get an idea of how stable search result lists are. They found that within a month, the top ten results for 25 queries had changed by more than 50%. Their study focused on the queries used by Lawrence and Giles (1998) in a study of search engine coverage, and included queries such as "zili liu" and "abelson amorphous computing". These queries were designed to return relatively small result sets (600 URLs or fewer). They found that search results changed regularly, more than might be expected as a result of growth and change in the Web, and hypothesize that the change observed is both due to growth and change, to search engine algorithm tuning, and to trade-offs made between quality and response rate during peak traffic times. The study presented in Chapter 8 confirms a high rate of change of search results Selberg and Etzioni found through a more recent tracking (2005 v. 1999) and over a longer period of time (over a year versus one month) for more popular queries.

It is likely that the rate of change to search engine result lists will increase as personalization, relevance feedback, and collaborative filtering approaches become more sophisticated and common. For example, the results returned by the personalized search system described in Chapter 5 change as the user received new emails, visits new Web sites, and authors new documents.

6.3.2 Change Interferes with Re-Finding

In discussing search result change, Selberg and Etzioni (2000) note that, "Unstable search engine results are counter-intuitive for the average user, leading to potential confusion and frustration when trying to reproduce the results of previous searches." Intille (2002) argues that any change to information erodes the perception of calm. Time and time again, changes to electronic information that should help the user instead get in the way.

This has been observed both in the studies presented in this dissertation and in related work. For example, dynamic menus were developed to help people access menu items faster by bubbling common items to the top of the menu. Rather than decreasing access time, research revealed dynamic menus actually slow their users down because commonly sought items no longer appear where expected (Mitchell & Shneiderman, 1989; Somberg, 1986). As another example, White, Ruthven, and Jose (2002) tried to help people search by giving them lists of relevant sentences that were dynamically re-ranked based on implicit feedback gathered during the search. To the authors' surprise, people did not enjoy the search experience as much or perform as well as they did when the sentence list was static. Similarly, although changes to result lists tend to represent a general quality improvement (e.g., Patil et al., 2005), Marshall and Bly (2005) observed important previously encountered results sometimes become unreachable as a result of the changes.

There is evidence that people have trouble interacting with the Web because they are trying to return to information that has changed. A study of Web usage by the GVU Center at Georgia Tech (1998) surveyed people on their biggest problems in using the Web, and found that "Not being able to find a page I know is out there" and "Not being able to return to a page I once visited" are significant problems. Cockburn et al. (2003) found that 25% of all people's bookmarks no longer worked. The broken URLs once pointed to information that the user indicated, through the process of bookmarking, as worth returning to. However, that information is no longer available where expected, making returning difficult.

People do not trust the Web as a repository for information. Whittaker and Hirschberg (2001) found that people kept paper copies of documents they had found on the Web for archival purposes, even when keeping the documents incurred some cost to the keeper. People have good reason to keep paper copies. Cockburn et al. (2003) analyzed the bookmark files of 17 people over 119 days, and found that two months after data collection, 25% of the bookmarks were invalid. Weiss (2003) notes that scientific papers that cite information found on the Web risk losing context.

For this reason, there has been considerable effort in keeping links from breaking and fixing them when they do (e.g., Ingham, Caughey, & Little, 1996; Park & Pennock, 2002;

Reich & Rosenthal, 2002). There are also efforts to archive Web content (e.g., Alexa, http://www.alexa.com; Kahle, 1997; Lyman, 2003).

6.4 Related Work on Systems that Support Finding and Re-Finding

Chapter 9 presents the Re:Search Engine, a system that works to keep the changes that support the finding of new information from interfering with re-finding. Although there are relatively few systems that have been developed to support both the finding of new information and the re-finding of old information, there are a number of approaches such systems can take. These are reviewed in this section.

The section begins with a discussion of systems that allow people to actively declare whether they are finding or re-finding. Such an approach requires users to choose to interact either only with old information or only with new information. However, as is shown in Chapter 7, people often want to interact simultaneously with both. Systems that present old and new information together, but that highlight interesting aspects of the information, are presented next. While such systems can draw attention to pertinent information, they do not inherently preserve old information or the context in which it was presented, and thus can still result in difficulty re-finding. The approach taken by the Re:Search Engine is to preserve the context of the original information, and present new information in the holes where old information has been forgotten. While this approach is unique, there are several systems from other domains that are relevant in understanding this thesis work, and the section concludes with a discussion of them.

6.4.1 Allow the User to Declare if Finding or Re-Finding

Although Web search engines have traditionally sought to return the search results that are the most relevant to a query without consideration of past user context, some recent search systems have begun to include basic re-searching capability. One simple way to support both finding and re-finding is to ask the user what their intention is. For example, a search engine could have two search buttons instead of one: one that says, "Search" and allows the user to find new information in response to a query, and the other that says "Search Again", and returns a cached version of the results that the user has seen previously for the query.

No search engines offer a "Search Again" button, but several, such as A9 (http://www.a9.com), allow users to mark pages of interest to return to later. Others remember results and search context over time, allowing users to use that context to re-find information (Raghavan & Sever, 1995; Bharat, 2000; Komlodi, 2004; Komlodi, Soergel, & Marchionini, 2006). Users declare their re-finding intent by visiting their search history. Pitkow et al. (2002) suggest allowing users to declare such an intent after the query is issued by adding a, "Have Seen/Have Not Seen" feature for sorting the documents returned by a search engine by either criteria. While such solutions preserve pertinent information, they do not preserve consistency in the interaction. Users are required to take a different path to the same information. They cannot necessarily just

repeat a query and expect to find the result they recall ranked in the same position they last saw it. Chapter 8 shows that such changes to presentation can cause problems re-finding.

Information management systems that preserve consistency in dynamic environments function like the hypothetical "Search Again" button; they permit their users to choose to interact with a cached version of their information space (Hayashi et al., 1998; Rekimoto, 1999). Employing similar methods to keep the results for repeat queries static would make re-finding simpler, but would also deny users the opportunity to discover new information. With such a system, a person could not, for example, revisit previously found information on breast cancer treatments while still learning about newly available treatments, as Connie did in Chapter 1.

The Re:Search Engine allows preserves consistency while not forcing its users to interact with static information by maintaining the information that is important to a person and changing that information that is not to reflect the most recent and relevant information. In this way, what it does is similar to a version control system (Hicks et al., 1998; Østerbye, 1992; Tichy, 1985). Version control systems try to maintain the important edits that a person has made to a set of documents, while including new information that has been created by other people editing the same information.

Search history systems, systems that cache previously viewed information, and version control systems all require their users to actively declare whether they are finding or re-finding information – by clicking a different search button, using a different application, or sorting results in a different way. As seen in Chapter 3, people like to exert the minimum effort possible when searching. They are unlikely to employ keeping strategies that require active involvement (Jones, Bruce, & Dumais, 2001), and even something as simple as declaring whether a task is a finding task or a re-finding task may require too much effort.

6.4.2 Highlight Interesting Information

Systems that present old and new information concurrently do not require their users to actively declare whether they are finding or re-finding, nor to interact only with new or old information. Such systems can help people quickly find information of interest by drawing the user's attention to it. This section briefly reviews systems that highlight new information to support the finding of new content, and that highlight information the user has seen before to support the re-finding of content the user has seen before.

Highlight New Information

Systems designed to support interaction with dynamic information often do so by highlighting new information that has become available. Message boards, for example, often highlight threads to which new messages have been posted. Francisco-Revilla et al. (2001a) developed a system to help people manage change on the Web that highlights the portions of Web sites that have changed. While doing this provides awareness of what has changed and helps people find new information quickly, it does nothing to support re-finding, and can, in fact, get in the way by being distracting.

Highlight Old Information

To actively support re-finding, other systems highlight previously viewed information. For example, a number of commercial search engines (e.g., A9, http://www.a9.com, and Google, http://www.google.com) put a timestamp next to previously visited results. This draws the user's attention to those results. However, results can still be difficult to re-find with such systems because they may no longer appear in the result list, or appear in unexpected locations.

6.4.3 Preserve Context while Changing Information

Chapter 8 suggests it is important that search results appear where expected during re-finding. The Re:Search Engine makes it possible for this to happen while still allowing new information to be displayed by taking advantage of the fact that people forget much of what they have previously seen. New information can be hidden in the holes in the user's memory. This section discusses other systems that take advantage of human memory to present new information in imperceptible ways.

When large changes of visual information pass unnoticed, it is called "change blindness". The section begins by reviewing change blindness literature and discussing computer systems that make use of change blindness to present new information. For the Re:Search Engine to make large changes to search result lists pass unnoticed, it is necessary to understand what people remember about search results. So after the discussion of change blindness, this section reviews studies of list memory. While it is not obvious that invisibly adding new information to search result lists will allow the searcher to effectively use the new information, related literature suggests that changes do not need to be noticed to provide benefit.

Change Blindness

Change blindness is a visual phenomenon where large changes to a scene occur without the viewer's notice. This can happen because the change occurs when the viewer is distracted (by a cut in scene, a flash of the image, or some other visual stimulus), or because the change happens gradually. An example of a large but imperceptible change is shown in Figure 6-1. The crosswalk in the first picture does not appear in the second. This difference is obvious when comparing the two pictures side by side. When the pictures are flashed one on top of the other, the crosswalk looks as if it is appearing and disappearing. But if a small gap is allowed between when each picture is displayed, the change becomes very hard to notice – even when the viewer is actively looks for a difference.

Rensink (2002) gives an overview of change blindness and discusses some of its implications for information displays. For example, some graphics systems have explored using change blindness to reduce computational overhead (Yee, Pattanaik, & Greenberg, 2001; Carter, Chalmers, & Dalton, 2003). These systems use highly computationally intense algorithms to render information where the user will notice at the expense of how information is rendered elsewhere.

Intille (2002), in a study of ubiquitous computing environments, looked at taking advantage of people's change blindness to pro-actively present new information in ways

Figure 6-1. An example of a fairly large difference that can go unnoticed due to change blindness. The lines of the crosswalk are present in one picture, and not the other.

that are unlikely to be disruptive. For example, information may be communicated to one individual in a manner that other people cannot see (e.g., through changes to a picture that cannot be seen unless the viewer is aware they are happening) or in locations that other individuals cannot see (e.g., projected onto a blind spot or shadow).

Several researchers in human-computer interaction have expressed interest in how change blindness might affect users ability to interact with computer based information (Nowell, Hetzler, & Tanasse, 2001; Varakin, Levin, & Fidler, 2004; Durlach, 2004). This research, however, has focused on the fact that people may miss important change due to change blindness, and the solutions presented try to draw users' attention to changes, rather than trying to take advantage of such holes in memory to present useful new information in an unnoticeable manner.

Previous Studies of List Memory

In order to preserve the memorable aspects of search result lists, it is important to study what is memorable about search results lists. Chapter 9 presents the findings of such a study. How lists of information items, and in particular lists of words, are recalled is well studied by cognitive psychologists (Asch, 1946; Henson, 1998; Hunt, 1995; Murdock, 1962; Terry, 2005). Several main effects in human memory of lists have been observed. These effects are described below and illustrated with an example of how a person might remember the following list of characters: *JBT2AMH*

Primacy effect Those items in a list that are presented first tend to be particularly memorable. For example, from the list of characters above, the *J* is particularly likely to be remembered.

Recency effect Those items in a list that are presented last tend to be more memorable than items presented in the middle. The *H* in the above list is more likely to be remembered than the *M*.

von Restorff effect Distinctive items are memorable. The *2* in the above list is memorable even though it is located in the middle of the list, because it is a number in the midst of letters.

Studies of list recall tend to take place in highly controlled environments and often do not involve information items that are either of particular interest to the participants or that are actively interacted with in the way one might interact with a search result. Studies of more complex information items than words, such as television commercials (Terry, 2005), and of more complex information uses, such as the forming of impressions of personality (Asch, 1946), have found the effects described above to be true with some variation, and these effects are useful guides in determining what about search result lists is likely to be recalled.

Improvements Do Not Need to be Noticed to be Helpful

Even though the Re:Search Engine works by providing new results in a list that looks the same as the originally viewed result list, the inclusion of new and better results can satisfy the user's information needs sooner. Usability improvements do not need to be noticed to benefit the user. A classic example is the Macintosh design for cascading submenus, where some flexibility in navigating to menu items is built into the menu design. The tolerance for small errors in navigation goes unnoticed by almost all users, but leads to fewer errors overall (Tognazzini, 1999). Similarly, a study of an improvement to cascading submenus showed all users performed better even though only three out of the 18 participants actually noticed the change (Ahlström, 2005).

The following chapters highlight this point by providing evidence that changes to the Re:Search Engine's result lists both go unnoticed and provide a noticeable improvement to performance. The next chapter presents, among other things, a study of the re-finding behavior observed in query logs. The results of this study suggest that noticeable changes are actually likely to hinder, rather than improve, performance.

If you want truly to understand something,
try to change it.
- Kurt Lewin (1890 - 1947)

Chapter 7

Understanding Re-Finding

The purpose of this chapter is to provide an understanding of re-finding behavior and how it differs from the finding of new information. The attributes of re-finding are explored via analysis of the data collected during two studies of search behavior: 1) the modified diary study described in Chapter 3, and 2) a log study of queries issued over the course of a year to the Yahoo search engine by 114 individuals. Following this is a discussion of how these findings suggest re-finding behavior can best be supported. The challenges of supporting both finding behavior and re-finding behavior at the same time with the same system are highlighted.

7.1 Observations of Re-Finding

Much of the finding behavior observed in the modified diary study presented in Chapter 3 involved re-finding. A large majority of the searches participants conducted within their email and file systems were searches for emails and documents that they had seen before. Many of the Web searches involved re-finding, too. Even those searches where the primary target was not something that the participant had seen before often involved sub-searches for previously viewed information. For example, one participant wanted to find new sneakers to buy. While she had not viewed the specific sneakers she ended up buying before, she found them by first re-finding the Web site of a sneaker vendor that she knew and liked. This section analyzes these instances from the modified diary study that give insight into to re-finding behavior. It looks at the value of meta-information in re-finding, discusses strategies of typical re-finding targets, and highlights problems information fragmentation can cause.

As mentioned in the previous chapter, one distinguishing feature of re-finding is that the searcher often knows a lot of meta-information about the target. Because more is known about targets that have been seen before than targets that have not, participants were particularly like to take advantage of meta-information when re-finding to orienteer rather than teleport. For example, Fernando reported that he wanted to find a research paper he had read before. Because he knew the author's name and affiliation, he was able to navigate to the author's research group's Web site, follow the links there to the

author's Web site and then to their publications page, and eventually find the paper. The use of meta-information for re-finding was also observed during Jim's search for Ellen Brooks's office number (recall Figure 3-1) – Jim had seen Professor Brooks' office number on her Web page, and used information he knew about her Web page (e.g., that it was associated with the Harvard Math Department) to re-find it.

In Chapter 6, several types of meta-information (people, path, time) were highlighted as particularly important for re-finding, and each of these types were observed to be used by participants in the study. As an example of the importance of the people associated with a re-finding target, Fernando used the author of the paper he was looking for in his search. The importance of the path taken to originally find the information is demonstrated when Jim chooses to return to Ellen Brooks' Web page via the same path he found it initially. And the temporal aspects of the information target were mentioned often, particularly during email searches, where participants used their memory of when the emails being sought was received to find them in folders sorted by date.

In several cases, participants wanted to find information that was similar to what they had seen before, but not exactly the same; they specifically wanted not to re-find information. The behavior in these cases looked particularly like teleporting, and involved large jumps in an attempt get outside of their known space. For example, Rachel wanted to find a different version of a paper than the one she had found on the Web because she believed the copy she had was corrupted. In her search, she used a Web keyword search engine to find other online copies and did very little browsing, except through the result list.

In contrast, re-finding was particularly likely to involve multi-stepped orienteering strategies. This is despite the fact that the information targets of re-finding searches were more likely to be specific pieces of information (e.g., Ellen Brooks' office number) and less likely to be general information (e.g., information on breast cancer) than unseen targets were. Only 39% of all of the Web searches for new information observed were for specific pieces of information, while 64% of the re-finding Web searches were for specific pieces of information.

Information fragmentation, where information is separated or spread among several devices, software programs, or locations, can be a source of problems for both finding and re-finding, but it appeared particularly likely to complicate re-finding searches. For example, when a person looks for something they have seen before, they may not know where to look if they saved copies of the information locally (e.g., in their browser cache or in an archive). They could try to locate the information where they originally found it, or find their local copy. Occasionally, for re-finding searches, the information being sought may be fragmented so as to be inaccessible from the user's current location. Here's an example where Alex tried to find something that did not reside on the machine he was using:

> Note what I did to find the email that I had just sent. I looked first on the wrong machine in my *sent-mail* folder. I realized that I had sent it from my ACM account so I needed to look in the *sent-ACM* folder. I tried that. There was no *sent-ACM* folder here because I don't send mail from ACM on this machine.

Alex's search was further fragmented because he needed the email to find some search terms to use in a Web search. In this way, the information he needed was fragmented across applications as well.

7.2 Query Log Study of Re-Finding

Analysis of the data collected via the modified diary study gives a broad picture of re-finding behavior in general. This section now focuses specifically on how keyword search is used to re-find. Thanks to the ubiquity of the Internet search engine search box, users have come to depend on search engines both to find and re-find information. However, the use of keyword search engines for re-finding has not been significantly addressed. The study presented here analyzes the re-finding queries issued to Yahoo by 114 people over the course of a year. Re-finding appears to be surprisingly common, and a number of characteristics of re-finding are found that suggest ways search engines can better support the behavior.

Log studies are a valuable tool because they give a large-scale, realistic picture of users' actions. However, unlike the modified diary study, they give no insight into underlying motivation. To study re-finding behavior through log analysis, it was necessary to try to glean from the data which queries were intended to re-find information rather than find new information. Re-finding intent was approximated by looking for repeated clicks on the same search result in response to queries issued by the same user at different times. The query used to find the same result may or may not be the same as the original query used to find it. For example, if a person searched with the query "KHTS" and clicked on the result http://www.channel933.com, and then later clicked on the same result while searching for "channel 933", the behavior was considered re-finding.

Because of the limited ability to truly distinguish re-finding behavior from finding behavior in the query logs, the results presented here were supplemented with an additional controlled experiment of 119 individuals.

No matter how it is approximated, re-finding behavior via keyword search engines is clearly very common. Forty percent of all observed queries led to a click on a result that was also clicked during another query session by the same user, and nearly 30% of all URLs clicked in the dataset were clicked by the same user more than once.

This section focuses on understanding how keyword search is used to re-find. First, the study methodology is discussed. Then the queries that were used to re-find information are compared with the queries used to initially find the same information to understand how re-finding queries can be identified. This is followed by a study conducted to predict which results will be clicked and finally investigations into the differences observed between individuals are presented. The research presented is the result of a collaboration with Eytan Adar, Rosie Jones and Michael Potts (Teevan et al., 2006).

7.2.1 Study Methodology

Analysis includes all queries issued to the Yahoo search engine over a period of 365 days (August 1, 2004 to July 31, 2005) by the anonymous users of 114 Web browsers.

Associating queries with a Web browser is an imperfect way to identify a user, since a single user might use multiple browsers (e.g., at home and at work), and multiple users may use a single browser (family members sharing a computer, users on public library terminals, etc.). Throughout this chapter the queries issued via the same Web browser will be referred to as being from the same user, as this represents the most common case. As few instances of repeat queries across users were observed, but many were observed from individual users, it is likely the repeat query behavior observed came from the same user. However, in some cases a complete picture of a given user's search history may not be available, or the records may in fact be the combined records of multiple users.

For the data described here, the focus is primarily on the large majority of queries for which there was a click on a result page. Next-page clicks, spell correction and related search clicks, and instances where there was no click at all are excluded. The data were not filtered to remove search spam or robot/mechanical searches. Some of the strongest repeated-search repeated-click traffic may come from robots and how those may be detected based on re-finding behavior is briefly discussed.

Very short term query repetitions were not of interest for understanding re-finding behavior. In the logs, many queries were repeated with short intervals, likely as a result of page refreshes or back-button clicks. To remove such repeat queries from the data, all instances of a query that occurred within thirty minutes of an identical query were considered to be a single query. The threshold was chosen because there was a clear distinction in the data between the frequency at which searches were repeated before and after this point, suggesting the behavior observed is different.

The following information is analyzed: The query terms issued, an anonymous key identifying the browser which issued the query, the time the query was issued, what results were clicked, and their position in the result list. Users were selected for inclusion in the study if they issued searches on at least four of the last 10 days of the year period. In total, the dataset contains 13,060 queries (an average of 115 per user) and 21,942 clicks. The basic statistics are comparable to those in other recently published studies. The average query length is 2.7 words, similar to what is reported elsewhere (Spink et al., 2001; Spink et al., 2002; Xue et al., 2004). The average number of results clicked per result page is comparable to that observed by others (Xue et al., 2004).

7.2.2 Identifying Re-Finding Queries

To successfully identify repeat queries in this data, it was necessary to associate queries by inferring the intent of the user, rather than rely on the exact query to be repeated. Many users repeated past queries perfectly (e.g., "bbc arabic"). Of the 13,060 query instances, 4256, or 33%, were exactly the same query as what the user issued at another time. In contrast, only 860, or 7%, of all queries were issued by more than one user.

Often when identical queries were issued by the same user, the searcher clicked on the same results each time. However, different queries were nonetheless found that led to the same click (e.g., "colorado lottery" and "colorado powerball"), and similar queries that led to very different clicks. This section proposes a taxonomy of repeat queries, based on various combinations of query and click comparisons, and discusses their probably underlying intent. The taxonomy can be seen visually in Figure 7-1.

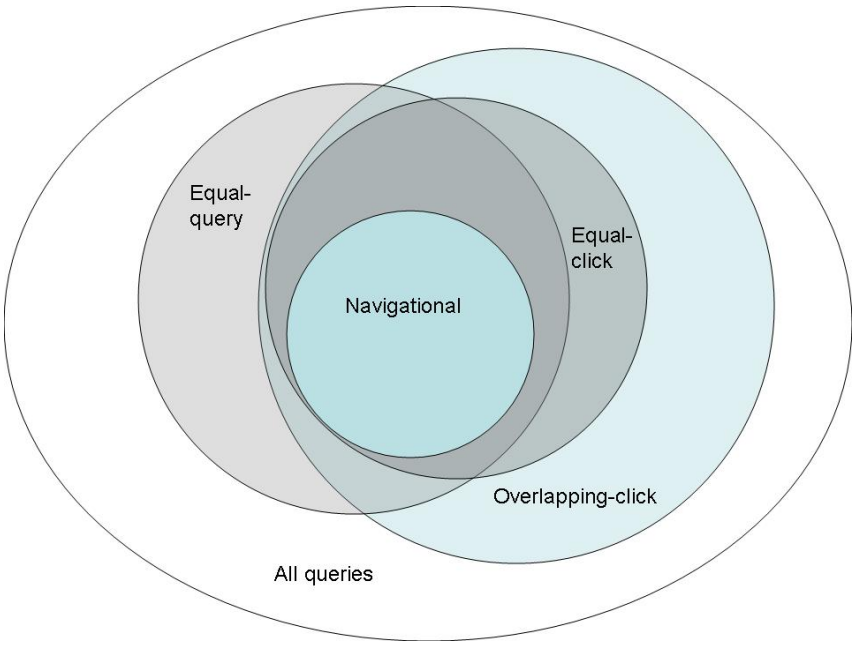


Figure 7-1. Venn diagram of the different types of queries.

As mentioned earlier, clicks were used as a proxy for re-finding intent. For this reason, the cases where users clicked on exactly the same set of results during two different query instances were of interest:

1. *Equal-click queries* – The user clicks on the same results in the two queries, but the queries may not be the same. This type may contain equal-query queries, and is a superset of bookmark queries. Given two click through sets (C_1 and C_2) corresponding to two queries (q_1 and q_2), $C_1 = C_2$.

Of course, equal-click queries do not necessarily represent re-finding intent (it could just be a coincidence that the user clicked on the same result), nor do they necessarily represent all queries with a re-finding intent (it could be the user did not successfully re-find a previously found result, and thus no repeat click appears). A slightly broader analysis of queries that include re-finding intent could include those queries with some overlap in result clicks.

2. *Overlapping-click queries* – Queries that have some common clicks. This type captures related intent and is the loosest form of repeated query. It is a superset of equal-click queries. Given two click-through sets (C_1 and C_2) corresponding to two queries (q_1 and q_2), $C_1 \cap C_2 \neq \varnothing$.

While looking at click patterns is likely to give a relatively accurate picture of whether or not the user is re-finding, search engines do not know what their users are going to click on at the time a query is issued. For this reason, it is useful to identify queries that looked particularly likely to be re-finding queries.

3. *Equal-query queries* – The user issues the same query but visits a potentially disjoint set of Web pages. This type is a superset of bookmark queries. Given two queries, $q_1 = q_2$.

4. *Navigational queries* – Queries where the user makes the same query and always goes to one and only one page. Given two queries, q_1 and q_2, and two corresponding click through sets C_1 and C_2 (each containing unique URLs), a bookmark query is one in which $q_1 = q_2$, $C_1 = C_2$ and $|C_1| = |C_2| = 1$ (in practice, when $C_1 = C_2$ the length of both is nearly always 1).

It is interesting, and important for identifying re-finding on-the-fly, to understand how often equal-query queries are also overlapping-click queries, and whether any query normalization can help better identify re-finding intent.

How Queries Can Differ

Query strings used to re-find information can differ from their original forms in many ways, and these are enumerated in Table 7-1. Most of the differences listed are trivial to identify automatically, but some are not. Those that are starred – including abbreviations, synonyms, and reformulations – are not considered in the analysis for this reason.

While many different changes are possible in a repeat query sequence, it is important to identify which were the most common. To do this, the notion of a *clustering ratio* is introduced. A clustering ratio allowed the differences to be determined quantitatively in

Table 7-1. Ways that queries resulting in repeat clicks can differ. Differences that are starred are not considered in the analysis presented here.

Difference	**Example**
Exact	"california secretary of state" and "california secretary of state"
Capitalization	"Air France" and "air france"
Extra Whitespace	"nick drake" and "nick drake"
Word order	"new york department of state" and "department of state new york"
Stop words	"atlas missouri" and "atlas of missouri"
Non-alphanumerics	"sub-urban" and "sub urban"
Duplicate words	"wild animal" and "wild wild animal"
Word merge	"wal mart" and "walmart"
Domain	"hotmail.com" and "hotmail"
Stemming and Pluralization	"island for sale" and "islands for sale"
Words swaps	"American embassay london" and "american consulate london"
Add/Remove Word	"orange county venues" and "orange county music venues"
*Add/Remove Noun Phrase or Location**	"Wild Adventures in Valdosta Ga" and "Wild Adventures"
*Abbreviations**	"ba" and "British Airways"
*Synonyms**	"Practical Jokes" and "Pranks"
*Misspellings**	"google" and "ggole"
*Reformulations**	"texas cheerleader mom" and "Wanda Holloway"

repeated queries. More formally, the clustering ratio is equal to the number of unique queries issued by a user when looking for a particular target divided by the total number of queries issued by the user when looking for that target. For example, the bag of queries {"apples oranges", "apples and oranges"} has a clustering ratio of 2/2 = 1, while the bag {"apples oranges", "apples oranges"} has a clustering ratio of 1/2. *Perfect clustering* means that given a set of queries of size *n*, the clustering ratio is 1/*n* (i.e., it is possible to create one cluster for all queries). It was interesting to determine the minimum set of transformations that can bring about this perfect clustering. For example, to achieve maximal clustering on the query set {"Weapons-of-Mass-Destruction", "weapons of mass destruction"} the data is transformed using two normalizations by removing capitalization and the dashes (non-alphanumerics).

Note that queries that look similar also represent searches for new information, and it is possible to cluster queries with different intents. It has been shown (Raghavan & Server, 1995) that traditional vector space measures of similarity are unsuitable for finding query similarity. As an example, one user searched for "first commonwealth pittsburgh pa" and "first night pittsburgh pa", and was almost certainly not interested in the same results for the two queries despite some overlap in words. Given that 19% of equal-query searches did not involve overlapping clicks, it is likely that even using an exact match in query string to identify a repeat query sometimes falsely identifies repeat queries.

How People Remember Past Queries

To better understand how people remember past queries, data was collected through a small-scale study of repeat queries. Participants were asked to issue a self-selected query and interact with the search results as they normally would. After an interval of a half hour to an hour, participants were emailed a brief survey that asked them to remember the query they issued without referring back to it. The results of this study give insight into how easy it is to remember past queries and how likely people are to remember them.

One hundred and nineteen people participated in the study. Fifty-two percent were male, and 45% female (numbers do not add to 100% because some participants declined to answer demographic questions). Most (64%) were between the ages of 25 and 39, but 18% were over 40, and 15% under 25. Ninety-seven percent reported using a computer daily. Typically, the follow-up survey was completed within a couple of hours of the initial search. Sixty-nine percent of all responses were received within three hours of the initial search, and all but five were received within a day. The average initial query length was 3.2 words, comparable to what has been found by others through Web query log analysis (Spink et al., 2001; Spink et al., 2002; Xue et al., 2004). The data were collected as part of a study used directly in the construction of the Re:Search Engine, presented in Chapter 9, and the study methodology is discussed further there.

Even though the time that elapsed between when participants in the study initially entered a query and when they were asked to remember it was relatively short, the original query was misremembered by 28% of the participants. To determine clustering ratios the transformations described in Table 7-1 were coded as independent normalization steps (2048 valid transformations).

Table 7-2. Clustering rates for most effective normalizations of repeat queries collected via a controlled study.

Minimal Scheme	Instances (% of total)
Stemming	4 (3.3%)
Capitalization	4 (3.3%)
Word Swap	3 (2.5%)
Capitalization and Word Swap	3 (2.5%)
11 Other combinations at < 2% each	13 (10.9%)

Each of the 119 query pairs were tested with these normalizations. Of the 119, 110 (or 92%), were clusterable using these normalizations. The 9 remaining query pairs appear to be users who were summarizing their previous query instead of repeating it (e.g., "whats the best pricing available for a Honda Pilot or Accura MDX ?" → "best pricing for Accura MDX"). Of the 119 pairs, 81 (or 68%) were exact matches (i.e., no normalization was required). Table 7-2 shows a few of the other normalization schemes that worked well for the dataset. Notably, no significant instances of duplicate words or word ordering changes were found such as were seen in the longer traces.

All queries included at least some overlap in terms. However, the information collected in the study about repeated queries was based on how people remember queries repeated within the same day, and often within the same hour. It is likely that with a longer elapsed time between initial query and follow-up query more drastic changes will occur than observed.

Identifying Re-Finding in the Logs

Given that 28% of participants in the above study misremembered their original query, it seemed likely many repeat searches would not appear in the logs as identical query strings where longer inter-query times dominate. A similar analysis as above was applied to find differences in query strings for searches with overlapping clicks. This is similar to the clustering of queries based on aggregate click-through data done by Wan et al. (2002).

Of the 21,942 clicks observed in the dataset, 6145, or 28%, of them were clicks on URLs that were clicked by the same user more than once. In contrast, only 1435, or 7%, were clicks on URLs clicked by multiple users. People were clearly much more likely to click on results they themselves had seen before. Forty percent of all observed queries (5216/13,060) were overlapping-click queries. Only 3692 of those involved exactly the same query string; similar to what was observed in the controlled study, 29% of overlapping-click queries involved different query strings.

Table 7-3 shows the most effective normalizations for the overlapping-click queries. It was not possible to cluster 12% of the query groups as they were of a form that could not be normalized in this analysis. Because of the large number of query sets where all queries were equivalent (600 or 46%), and the effectiveness of swap and add/remove normalizations, there was no significant trending in other combinations for specific users.

Table 7-3. Clustering rates for most effective normalizations for overlapping-click queries.

Minimal Scheme	Instances (% of total)
Single word addition/removal	172 (13%)
Single word swap	62 (4.7%)
Capitalization	59 (4.5%)
Word merge	29 (2.2%)
Non-alphanumeric and word add/remove	21 (1.6%)
Capitalization and word swap	20 (1.5%)
Non-alphanumeric removal and word swap	18 (1.4%)
35 Other combinations each at < 1%	165 (12.7%)

Temporal Aspects of the Query

There was a significant difference in elapsed time between repeat queries for the two studies. The average time between overlapping-click queries in the logs is over 12.2 days (292 hours) days with a median of 30 hours. This longer interval presents many more opportunities for users to forget or change the query. In fact, given two subsequent queries in a repeated query chain, a slightly shorter interval was found for exact repeats (median of 29.2 hours) than changed queries (median of 31.7 hours). Though not significant (p=0.15), the result is suggestive of a trend.

Deeper analysis of the effect of elapsed time, however, revealed something more subtle. Grouping all repeated query pairs into quartiles based on the inter-query time, the number of queries were counted that were exact matches, as were those that changed. The number of changed queries was 23%, 16%, 19%, and 21% in each bucket ranging from shortest inter-query time to the longest (median of 0.45, 20.7, 75.67, and 526.3 hours respectively). These results may indicate that queries issued near each other temporally are still in flux and are being refined. Queries repeated with a large inter-arrival time may be misremembered. Those queries repeated on a daily or every few day basis have been refined and used with a frequency that reduces changes.

Because the logs contained a year's worth of data it was possible to test whether certain repeated queries occurred at the same time every day. For example, users may make work related queries during particular hours of the day and other entertainment, or lifestyle queries at other times. This seemed likely because time of day has been shown to play an important role in query topic (Beitzel et al., 2004). To test how time affected repeat queries, all time stamps were normalized to represent only the time of day. Thus all queries appeared to have occurred during one "virtual" day (the offset of each day relative to GMT is unknown, but irrelevant). For each user the histogram of total queries during this virtual day was found, as well as for specific queries. A sample of these histograms is shown in Figure 7-2.

In the 272 user/query pairs where queries were repeated more than five times there were a number of daily spikes. The most significant spikes appeared to be for entertainment and sports searches (e.g., "whitesox.com", adult Web sites, and music groups), as well as for searches for news and email sites (e.g., "hotmail"). Interestingly, the "whitesox.com"

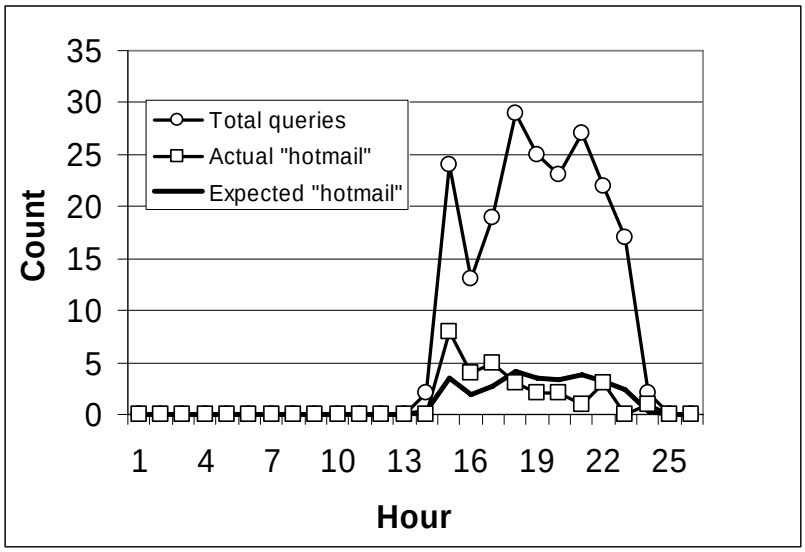


Figure 7-2. Temporal clustering for the query "hotmail" for one anonymous user.

query from one particular user spiked between 5pm and 7pm, which coincides with the televising of games. There were only two users with spikes in (likely) work related searches (e.g., searches for specific company Website and organization). The results for individual users are consistent with those of Beitzel et al. (2004) where an aggregate query stream was analyzed.

7.2.3 Predicting the Query Target

Even with only very basic normalization of the queries, it appears that repeat query strings occurred commonly for individual users. When queries are repeated, it would be useful for a search tool to be able to predict when the user intends to re-find previously viewed information, because this could affect the best results to display or the best manner in which to display them. This section looks at predicting whether a previously viewed result will be clicked based on the query string and past clicks.

Searchers may be looking for new information, or they may be looking for information that they have seen before. It was most common to look for the same information; approximately 87% (3692/4256) of equal-query queries were also overlapping-click queries. Fewer searches (1632 or 38%) resulted in at least one unique click. It was not always the case that the searcher only wanted old information or new information when they issued an equal-query query, as 25% of the searches, or 1070, involved both a repeat click and a unique click.

This section begins by looking at the effect that the elapsed time and number of previous clicks have on repeat queries. A large number of repeat queries are navigational queries, or queries where the searcher uses the query to navigate to a specific Web page (e.g., types the query "yahoo movies" into the search box to navigate to the URL http://movies.yahoo.com). These queries are particularly easy to predict, and their

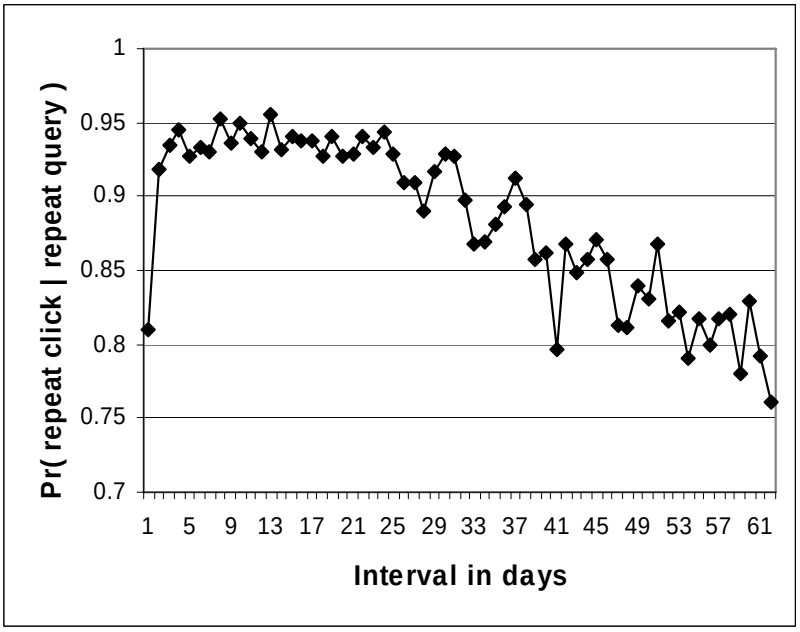


Figure 7-3. Probability of a repeat click for queries where the query string has been seen before as a function of time.

characteristics are discussed in greater detail, as well as the characteristics of other types of queries.

The Effect of Elapsed Time

This section looks at how the elapsed time between equal-query queries affected the likelihood of observing a repeat click. The probability of a repeat click as a function of elapsed time between identical queries can be seen in Figure 7-3. Repeat queries that were issued very close together in time (e.g., within several hours) had a relatively low probability of resulting in a repeat click. The probability of a repeat click for queries issued within an hour is 64%, compared with the earlier reported overall average of 87%. Queries repeated very quickly probably occur as part of the same search session, and represent instances where the user is looking for something new.

The probability of repeat clicks climbs quickly, however, for intervals longer than a day or two. Once it reaches a peak, the probability of a repeat click between identical queries slowly declines for a period of time. This may represent a trend to need to re-access previously seen information over time. It also may be related to the fact that many browsers no longer highlight previously visited links after a default of 20 days.

Navigational Queries

In some instances, it was possible to very accurately predict the likelihood of a repeat click using clicked results from past queries. Navigational queries appeared to be particularly easy to predict. Recall that navigational queries are equal-query queries where the user clicked on the same result for each query instance and did not click on any

other results. Using this definition, 507 (or 47%) of all unique equal-query queries issued were labeled navigational queries.

Navigational queries tended to be somewhat shorter in length than other queries (13.6 characters, compared with 16.4 characters for non-navigational equal-query queries and 16.7 characters for overlapping-click queries). This makes sense because navigational queries are intended to be an easy way to return to a Web page, and thus should be short and easy to remember. Navigational queries were also were more likely to include an indication they were a search for a URL – 12% of all navigational queries contained ".com", ".edu", or ".net", while only 5% of the other equal-query queries did.

Navigational queries were also repeated more often than other repeat queries (4.0 times, compared with 3.8 for equal-query queries and 3.3 for overlapping-click queries) and the interval between bookmark queries was longer (22 days, compared with 20 days and 16 for equal-query and overlapping-click queries respectively). It is likely that navigational queries occur more times because they are more of an access strategy than a search, and people tend to access more than search. The longer intervals are probably because the queries are probably chosen to be particularly memorable even across long periods of time.

It was easy to predict whether or not a query was navigational given two previous instances of the same query as training data. Doing this permitted the automatic, on the fly classification of 1841, or 12%, of all observed searches as navigational. For these searches, it was possible to predict with 96% accuracy on of the URLs that was clicked. When restricted to predicting the first URL clicked, accuracy only dropped slightly, to 95%, and if the prediction was that only that URL was clicked, accuracy dropped slightly more, to 94%.

It was less easy to identify a navigational query using only one previous query. While doing so covered more of the data (2955, or 23%, of the searches), the prediction was right only 87% of the time. Given 87% of all equal-query queries involve overlapping clicks, it is not at all surprising that it is possible to predict exactly which result will be clicked 87% of the time given the user is known to have only clicked one result before.

Other Types of Repeat Queries

It was also explored whether it was possible to predict whether a person was going to click new results or repeat results for equal-query queries that were not navigational. Using a number of features suggested by the earlier analysis presented in this chapter, such as elapsed time, query length, and number of results clicked previously, an SVM (SVM-Light, Joachims, 1999) was trained to predict two things: 1) whether or not a new result would be clicked, and 2) whether or not a repeat result would be clicked.

The strongest predictors for a click on a new result included the number of times the query was issued previously (and if it was issued more than once before), whether any previously viewed result was clicked more than once, and several features based on the number of previous clicks that were the same for queries that were repeated only twice:

- Number of clicks the first time the query was issued
- Number of clicks the previous time the query was issued

- Number of unique clicks the previous time

While there did not appear to be a correlation between the number of clicks and the likelihood of a repeat click, given the value of these features in predicting new clicks it seems it is indicative of a new click.

The strongest predictors for a repeat click were the fact that only one result was clicked during the previous search and the fact that the query had been issued more than once. Interestingly, these features are also useful for identifying bookmark queries, which do have a high incident of repeat query (although queries identified as bookmark queries were excluded from this analysis).

Using the features described above and leave-one-out cross-validation, the ability of the SVM to predict whether a new result or a unique result would be clicked was studied. The baseline accuracy was what could be expected if people were always assumed not to click on something new (61.4% accuracy) and to click on something they clicked before (74.7% accuracy). In both cases, the SVM was able to make a significantly ($p<0.01$) better prediction – getting it right 79.3% of the time for new clicks (an increase of 30%), and 78.1% of the time for repeat clicks (an increase of 5%). The SVM probably does a better job predicting new clicks than old because the bookmark data, which was the most easily identifiable repeat click data, was excluded.

For both cases, the user was also explored as a feature. While including the user led to a slight improvement over not including the user in both cases (80.1% accuracy in predicting new clicks and 79.4% accuracy in predicting repeat clicks), the was no significant improvement in doing so. However, it seems likely that users do exhibit different repeat and new click behavior, and probably need to accumulate additional features that well reflect the user.

7.2.4 Individual Behavior

To get a better understanding of user behavior, the differences between individuals were studied. Table 7-4 shows the variance in how often each different query type was to be observed for each individual user studied. While there were general trends in user behavior with various repeat queries, individual differences did exist. Of the 114 users, 102 issued at least one equal-query query, and 87 performed at least one navigational query. However, 15 users had equal-query queries but no navigational queries, possibly indication an exploratory mode.

Analysis of individual behavior may lend itself to detecting robots and search engine optimizers. For example, users with many regularly spaced bookmark queries are possibly using an automated system. One user, for example had 50 bookmark queries in 52 visits (96.2%). Another had 334 bookmark queries in 417 total queries (80.1%).

Figure 7-4 shows the rank ordered users (by percent of bookmark queries) and displays their percent of bookmark, equal-click, and equal-query queries. Clearly, while there is some trending, there is also a large degree of variability with some users issuing many users issuing significantly more equal-query and equal-click queries relative to their bookmark queries. Figure 7-5 shows a cumulative histogram of the various query types.

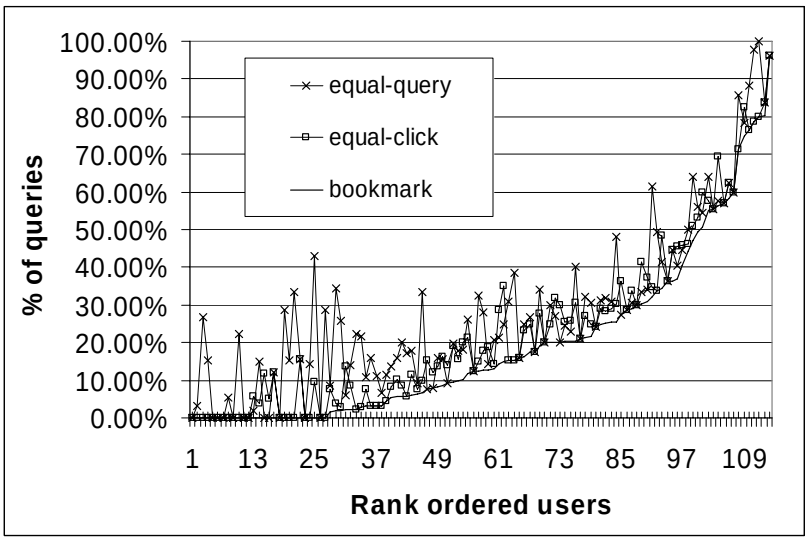


Figure 7-4. Percentage of different types of repeat queries for different users.

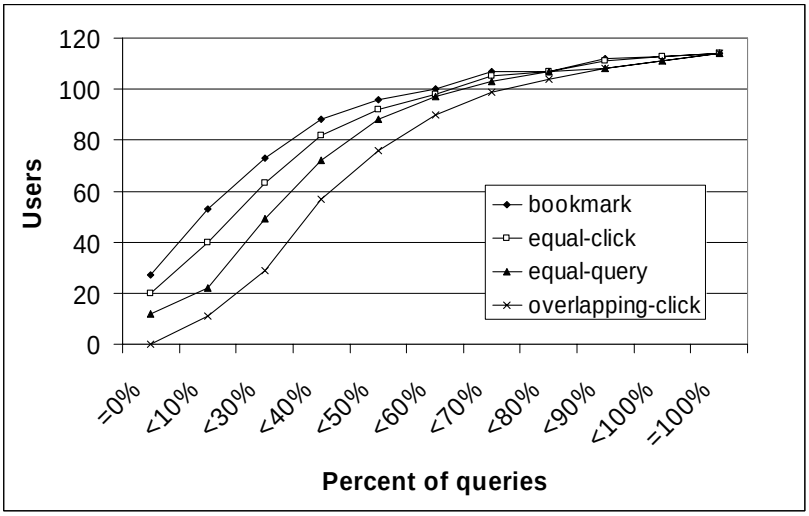


Figure 7-5. The cumulative distribution of query types for all users.

Table 7-4. Repeated query statistics (as % of all queries).

Query type	Mean	Median	Min	Max	Variance
Equal-query	28.2%	24.2%	0%	100%	5%
Equal-click	23.3%	16.1%	0%	96.2%	5.2%
Overlapping-click	35.6%	30.5%	4.5%	100%	4.7%
Navigational	19.4%	12.6%	0%	96.2%	5%

7.3 Supporting Re-Finding

The findings presented in this chapter have many ramifications for search engine design that are applied to the system presented later in Part II. Traditionally, search engines have focused on returning search results without consideration of the user's past query history, but the results of the log study suggest it most likely a good idea for them to do otherwise. For example, for navigational queries it was possible to predict with very high accuracy that the user will click on a particular result and this knowledge can be used to improve search engine performance. Chapter 9 presents the Re:Search Engine, a system that uses people's past query interactions to ensure results that have been clicked on before and are likely to be looked for again remain ranked where the searcher expects them to be. The Re:Search further assists in re-finding by helping its users phrase repeat queries easily, encouraging equal-query queries and making it easier to identify re-finding behavior.

Although finding and re-finding tasks may require different types of support, tools will need to seamlessly support both activities. As shown in the log analysis, people often clicked on both old and new results during the same search. However, finding and re-finding can be in conflict – finding new information means getting the best new information, while re-finding previously viewed information means getting the previously viewed information being sought. The next chapter shows that when previously viewed search results changed to present new information, the searcher's ability to return was hampered, and the following chapter looks at how finding and re-finding can be reconciled, suggesting a way to allow people to interact with the best new information while still finding what they have seen before.

Things do not change; we change.
- Henry David Thoreau (1817 - 1862)

Chapter 8

Why Re-Finding Requires Personalization

The purpose of the previous chapter was to understand re-finding. The analysis revealed that people commonly use search engines to repeat searches and re-find previously viewed information. This chapter demonstrates that an individual's previous information interactions are important for re-finding, and suggests that search tools should personalize their user experience to account for differences in interaction history.

People rely on what they remember about their previous searches to re-find. For example, in Chapter 1 Connie found a list of breast cancer treatments, and then used the many things she learned about that list during her initial encounter (e.g., that the list was hosted by About.com and that it could be accessed via a search for "breast cancer treatments") to re-find it. However, Web based information often changes, and in that example the search results for "breast cancer treatments" changed between searches. Changes can be useful, because they can provide the searcher with new and valuable information. In Connie's case, several new results about new treatments appeared in the new result list when she tried to re-find the list of treatments. However, this chapter shows that changes can also interfere with re-finding. For Connie, the change to result ordering caused her list of treatments to no longer appear where she expected it.

The chapter begins with additional analysis of the query log discussed in the previous chapter. The analysis demonstrates that changes to search results interfere with re-finding. Evidence is then provided that Web search engine results change regularly, even in the absence of change-inducing functionality such as personalization and relevance feedback. To further explore the affect that change has on re-finding, a naturalistic study of the difficulties people said they encountered when returning to information on the Web is presented. These findings influence the design of the Re:Search Engine, a search engine designed to support both the finding of new information and the re-finding of previously found information that is presented in the following chapter.

8.1 Change Interferes with Re-Finding

To understand how changes to search result ranking affected a user's ability to re-find information, the Yahoo query logs presented in Chapter 6 were further analyzed.

Because logs contain a year's worth of queries, they include many instances of changes to the search result rankings associated with repeat queries. It appears that changes to result ranking reduce the likelihood of a repeat click. This suggests that changes to result orderings cause people to re-find less information and view more new information. Such a behavior change could represent a positive outcome for users; the new information may be better than the information that was found previously. However, it could also represent a negative outcome if users do not repeat clicks because they are unable to re-find the old information. Our analysis suggests that the later is the case. Queries where information was clearly re-found were examined to reveal that repeat clicks on the same result are slower when the clicked result changes rank. This analysis, described in greater detail below, was performed with Eytan Adar, Rosie Jones, and Michael Potts (Teevan et al., 2006).

8.1.1 Rank Change Reduces the Chance of Click

The probability that any given click would be a repeat click for overlapping-click searches was compared under two conditions: 1) when a change in rank was observed among one of the common clicks, and 2) where no rank change was observed. Because the dataset did not contain results that were not clicked, it was only possible to identify result lists that had changed when rank changes among clicked results were observed for queries with overlapping-clicks. A better understanding could be derived from a knowledge of which results were displayed, even if not clicked.

Repeat clicks were significantly more likely to occur when there was no observed change (case 2). Eighty-eight percent of the clicks were repeat clicks if there was no change in rank, while only 53% of the clicks were repeat clicks if there was a change in rank. These estimates were obtained by averaging over all consecutive pairs of overlapping-click searches.

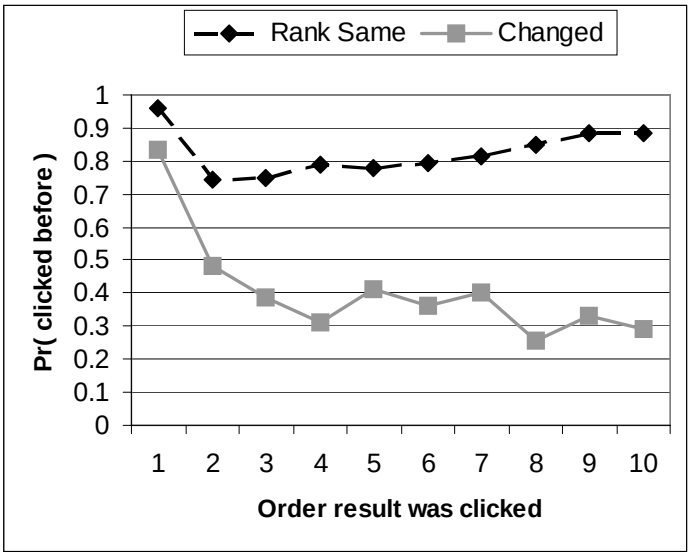


Figure 8-1. Probability of a result being clicked again as a function of the order the result was clicked. Results were significantly less likely to be clicked again if they changed rank.

Figure 8-1 shows the probability that a clicked result was a repeat click as a function of the order in which the click occurred following a repeat overlapping-click query. The dashed curve corresponds to the probability averaged over those searches where no rank change was observed; the solid curve corresponds to an average where at least one result changed rank. Comparing the two curves illustrates that a change in rank between queries makes it substantially less likely that a given result will be clicked on again during a follow-up search.

Also in Figure 8-1 is a sharp drop in the probability of a repeat click between the first result and the second. Given a finite number of results were clicked initially, it seems reasonable if the user first clicked on repeat results that the probability of a repeat click would tend to drop with increasing numbers of clicks as the user exhausts the set of previously-clicked results. The drop continues past click two when restricted to clicks on results with rank changes, which would seem to indicate that users are more likely to click on new results as they continue to interact with the result list than they are to click on previously clicked results which have changed rank.

As mentioned earlier, it is not immediately obvious from this analysis whether a decreased likelihood of re-finding reflects a positive or negative influence of result list changes on user experience. It could be that the changes interfered with re-finding, or it could be that the searcher found new and better information in the new result set.

8.1.2 Rank Change Slows Re-Finding

To get a better idea of whether changes interfered with re-finding, repeat queries where it was clear that information was being re-found (as evidenced by a repeat click) were analyzed. Because easy searches are likely to take less time than harder searches, the time interval between a search and a click on a result that was seen before was used as a proxy for ease. For this reason, the time from when a query was issued until the common URL was clicked was measured for non-equal-query, overlapping click queries.

Table 8-1 shows the average number of seconds it took to click a URL that was clicked during the initial session when that URL was 1) shown at the same rank it originally appeared, and 2) shown at a different rank. If the rank of the result was unchanged, the second click occurred relatively quickly, while if the rank had changed, it took significantly ($p<0.01$) longer. Changes to result ordering appear to make re-finding harder.

Table 8-1. Time to click as a function of rank change.

Query type	Mean	Median	StdDev
Rank the same (case 1)	94	6	234
Rank changed (case 2)	192	26	365

8.2 Search Results Change

It is a problem that changes to result lists interfere with re-finding, because search result lists can change due to personalization, relevance feedback, or improvements made to the search engine's underlying index and algorithms. Analysis of the Yahoo search logs suggests such change is common; 26.8% of the results that were clicked more than once were not actually in the same rank the second time they were clicked as they were the first time they were. It is likely that even more results changed rank than observed through click data, because, for example, a result could have not been re-clicked because it disappeared from the list entirely.

In order to better understand how search results change, ten queries were tracked weekly over the course of a year, and their patterns of change explored. This section discusses how the queries were selected and tracked, and then presents what analysis of the tracked queries' result lists revealed.

8.2.1 Study Methodology

The ten tracked queries, summarized in Table 8-2, were chosen in several ways. Five were selected to represent popular queries, and five were selected to represent realistic queries. All queries were generated in October 2003.

Popular queries were generated in two ways. Four were randomly selected from the Lycos Top 50 Elite queries. The Lycos Top 50 queries represent consistently popular queries issued to Lycos. Each of these four queries remains in the top 50 Elite today, nearly three years later. The query "Harry Potter" has been a top 50 query on Lycos since July, 2000, "Final Fantasy" since September, 1999, and "marijuana" has been in and out of the top 50 since January, 2000. The query "Las Vegas" is particularly popular, and has been a top 50 queries since Lycos began tracking them in August of 1999. The fifth query, "India", was selected as the most popular country search issued to Google.

Table 8-2. Query source for query results tracked over the course of a year.

Query	Source	Type	Number of results
Harry Potter	Lycos Top 50 Elite	Popular query	120,000,000
Final Fantasy	Lycos Top 50 Elite	Popular query	133,000,000
marijuana	Lycos Top 50 Elite	Popular query	38,500,000
Las Vegas	Lycos Top 50 Elite	Popular query	273,000,000
India	Google's top country search	Popular query	1,130,000,000
neon signs	AltaVista Real Search	Realistic query	7,090,000
artificial flowers	AltaVista Real Search	Realistic query	10,900,000
movies	AltaVista Real Search	Realistic query	956,000,000
credit reports	AltaVista Real Search	Realistic query	263,000,000
Teevan	Vanity query	Realistic query	83,100

Realistic queries were also selected two ways. Four were generated using the AltaVista Real Searches functionality that displays real searches being issued to the AltaVista search engine. A fifth realistic query, "Teevan", represents a vanity search.

Table 8-2 also shows the number of results for each query currently returned by Google. This serves as some indication of how popular the query is at the moment. As can be seen, two of the queries gathered through AltaVista Real Search ("movies" and "credit reports") have many results and are likely popular queries in their own right. It is not surprising, of course, that some of the realistic queries are popular, since they were generated by randomly sampling real queries that were being issued.

For each of the ten queries, the top 100 results from Google were downloaded every week. The analysis presented is of the differences between results recorded from April 13, 2005 through April 5, 2006. This represents a total of 45 search result lists per query (the number is not 52 because the data contains three 2-week gaps and two 3-week gaps).

8.2.2 How Results Changed

To compare the similarity between two search results, Selberg and Etzioni (2000) measure the percentage difference between the two result sets, set A and set B.

$$\text{difference}(A, B) = \frac{|\ (A-B) \cup (B-A)\ |}{|\ A \cup B\ |} \qquad \textbf{(5)}$$

This same measure was used to compare the percentage difference between result lists. The difference between each 100 URL query result list from the initial result recorded on Aprli 13, 2005, is shown in Figure 8-2. Clearly there are large changes in the result lists for the each query, and the amount of change increases over time. The percentage difference of a list ranges from on average percentage about 30% when there is only a week or two between queries, to over 70% when a year has elapsed.

Note also that while the rate of change to the result lists appears reasonably consistent across queries, some queries experience more rapid change than others. A difference that arises from the data is between popular and realistic queries. Figure 8-3 shows the percentage difference of the top 100 results between weeks, grouped by type (popular or realistic), and highlights the greater rate of change for realistic queries than popular queries. The average weekly percentage difference for realistic queries is 30%, and 22% for popular queries. This difference is significant ($p<0.01$). It is not surprising that popular queries might have more consistent result sets, as people invest considerable long term effort in placing and maintaining the sites in the top results for popular queries.

Like Selberg and Etzioni (2000), the change to the top 10 results for each query was compared with the change to the top 100 results. The analysis confirmed their findings that there is less stability among the top 10 results, as can be seen in Figure 8-4. This difference is less striking than the difference between popular and realistic queries, but significant ($p<0.01$). Results in the top 10 differed by 21% on average, while results in the top 100 differed by 26%. The greater rate of change among the top 10 results might be a result of these positions being particularly coveted.

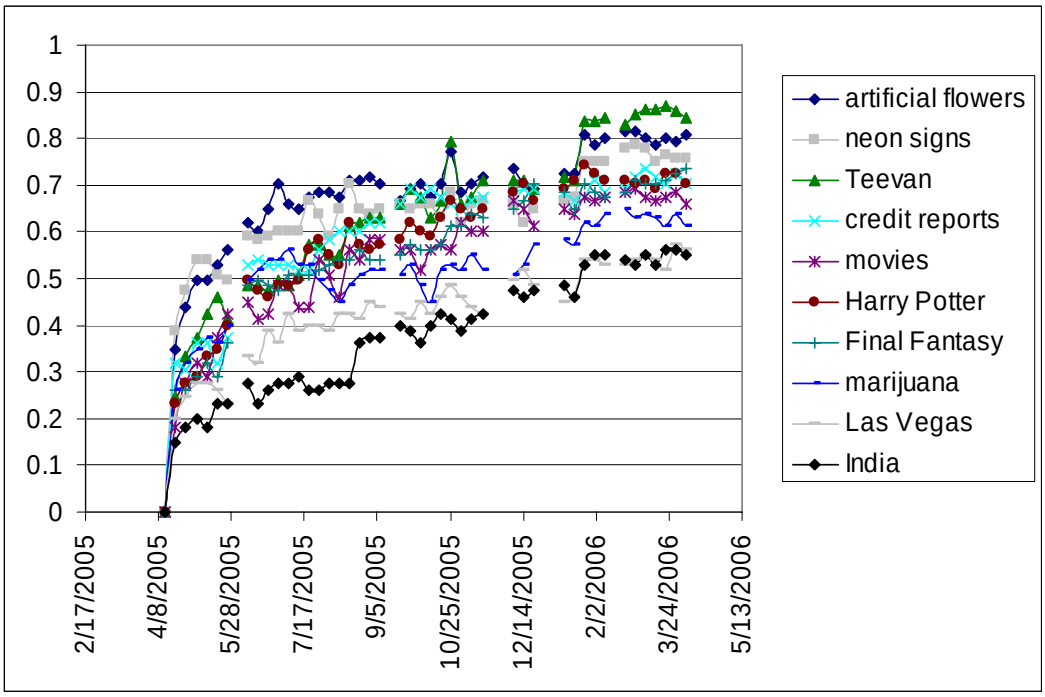


Figure 8-2. The percentage difference for top 100 query results for a query from the initial result set returned for that query on April 13, 2005.

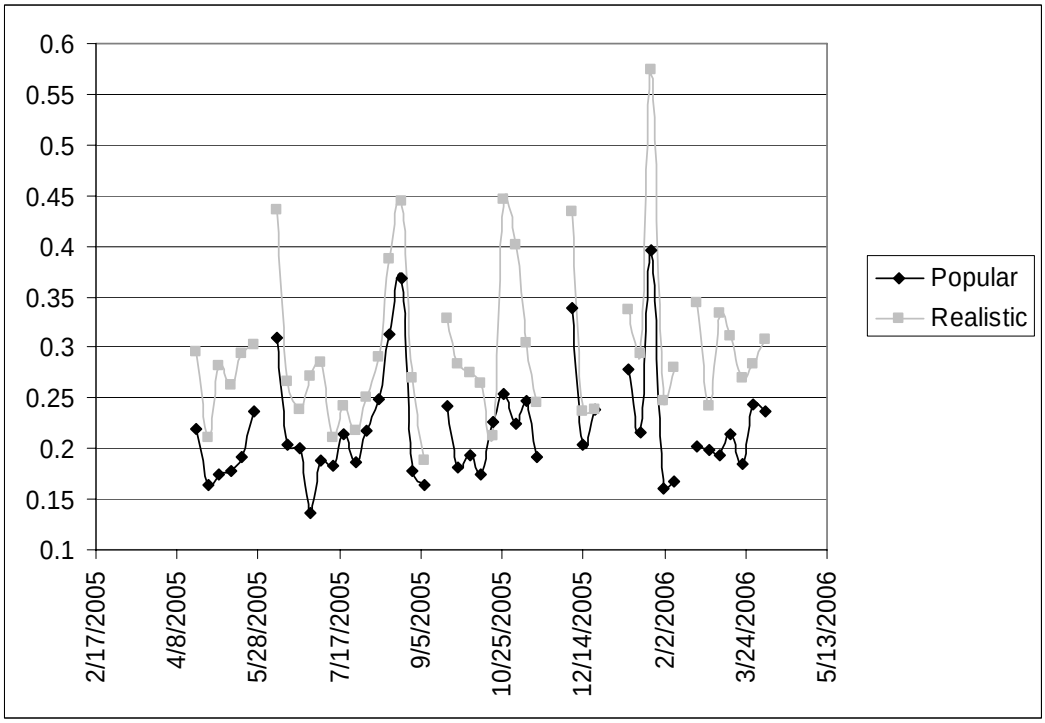


Figure 8-3. Weekly percentage difference between query result lists, for popular queries and for realistic queries.

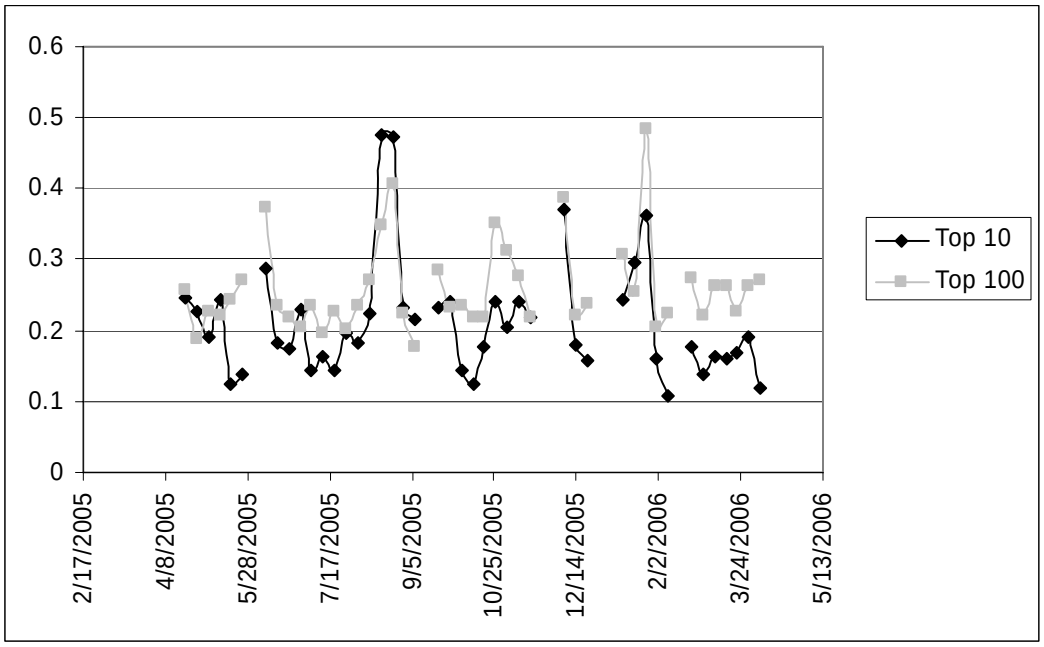


Figure 8-4. Weekly percentage difference between query result lists, for results in the top 10 and results in the top 100.

The early analysis of the effect of search result change on re-finding revealed that problems could occur not just because information becomes unavailable, but also because information becomes difficult to find because it is not where expected. For this reason, it is worth studying the rank change of results that remained in the result list over time. On average, results that did not disappear from the top 10 changed rank by 0.5 places weekly, and results that did not disappear from the top 100 changed rank by 4.7 places weekly. The pattern of change is similar to what was observed for the percentage difference. One interesting thing to note is that results were equally likely to move up in the list as they were to move down. Exactly 50% of all results that remained in the top 100 between weeks that also changed rank moved up, and exactly 50% moved down. This did not vary between popular and realistic queries, but results in the top ten were somewhat more likely to move down than up, with 52% moving down and only 48% moving up.

In all of this analysis, the result's URL is used as the measure of "sameness" between results. However, the same URL could point to different content over time, and different URLs could be used to point to the same content (e.g., if the People Magazine Web site is initially listed as http://www.people.com but later referred to as http://people.aol.com, the results will appear different although both point to same content). People, when browsing search results, likely do not often focus on the URL, but rather use the result's title and summary to help them re-locate information.

For this reason, the amount of change was studied between title and summary description for results that were the same between query result lists. Titles and summaries were compared by ignoring case, all HTML formatting, and any non-alphabetic characters. Thus, the change in title for the result http://en.wikipedia.org/wiki/Marijuana/ from

"Marijuana - Wikipedia, the free encyclopedia" to "Cannabis (drug) - Wikipedia, the free encyclopedia" was considered a change, but a subsequent change in title to "cannabis – drug: Wikipedia, The Free Encyclopedia" would not be considered a change.

Result titles were relatively consistent across time. Only 1.6% of the time that the same URL occurred in a list after a week did it appear with a different title. Summaries, which are longer, were more likely to vary, and 12.2% of the time that they recurred in a list they were different. The changes appeared to be both a function of how the search engine represented the result and a result of changes to the content of the result. Because people are able to easily recognize Web pages from truncated titles and URLs (Kaasten, Greenberg, & Edwards, 2002), it is likely many of the changes to how the result is displayed pass unnoticed.

This section has clearly demonstrated that search result lists are very dynamic. It is likely that the rate of change will increase as search engines are improved to take into account more information and use more complex ranking algorithms. With personalization, results to the same query change as the user's profile changes, with collaborative filtering, they change as the community's interests change, and with relevance feedback they change even as the user interacts with the results. As an example, the results returned by the personalized search system described in Chapter 5 change as the user received new emails, visits new Web sites, and authors new documents. Such change benefits the user by giving access to new and better information, but needs to be done carefully so as not to interfere with re-finding.

8.3 Re-Finding when the Web Changes

To better understand how people react to problems re-finding that are caused by informational changes, a study was conducted investigating the way people described the difficulties they encountered when returning to information on the Web. This was done by analyzing Web pages, collected via a Web search, where the phrase, "Where'd it go?" was used. A number of interesting observations arose from the analysis, including that the path originally used to locate the information target appeared very memorable, and that the temporal aspects of when the information had been seen before were rarely referenced in an absolute sense. People expressed a lot of frustration when problems arose, and often sought an explanation of why they could not find their target, even in the absence of a solution.

The study was conducted by analyzing instances, collected via a Web search, where the phrase, "Where'd it go?" was used to refer to a piece of information. The following quotation is an example from the data that illustrates a number of the observations discussed in greater detail (quotations are reported exactly as they occurred, without correction to spelling or grammar):

> I remember when I first joined these forums! There was little "Did you know" facts about Star Wars at the front page, but they were replaced with movie qoutes! Why did they disappear?

(a) (b) (c)

Figure 8-5. Three instances containing the phrase "Where'd it go?" The first (a) is a posting from a person looking for Web functionality. The second (b), titled "Where'd it go?", is a redirect page. The third (c) offers support finding information that has moved due to a site change.

The description emphasizes that the Star Wars facts were originally encountered on the forum's front page, and there was a trend in the data to emphasize the importance of the original path used to encounter the information target. On the other hand, time is not mentioned directly in the quotation, but rather alluded to by relating earlier access to a personal event. The study suggests that the temporal aspects of when the information was seen before were often referenced this way. Frustration, suggested in this example by the exclamation marks, was commonly observed, and it appeared that an explanation of why the change had occurred was often sufficient to allay frustration, even in the absence of a solution. In the example given above, instead of asking for a pointer to the missing information, the person asks for an explanation. The study presented here expands on prior studies of how people return to information by investigating how people coped when changes occurred to their information target and its environment.

8.3.1 Study Methodology

The instances of re-finding analyzed in this chapter were found by collecting Web pages that contained the phrase "Where'd it go?" via a Google Web search. Because Google only returns the top 1000 results, the search yielded 1000 pages of 5,340 reported. This set of pages could have been supplemented by performing the same search on other search engines, but there was considerable overlap among the result sets from different search engines, with 55% to 62% of the top 100 results already belonging to the Google set. Other phrases with similar meanings, such as "Where did it go?" and "I can't find it anymore," could also have been used to supplement the document set. "Where'd it go?" was selected because it was found to be the phrase most commonly used in the appropriate context. Additional instances found via other search engines or phrases appeared to merely enforce the phenomena observed in this chapter. This suggests that little would have been gained by supplementing the data collected.

The data were analyzed by making an initial pass through the data to develop a coding scheme and identify the 258 instances that contained expressions of difficulty re-finding information. A second pass was made to code this subset.

8.3.2 Overview of the Data Collected

Excluding duplicates and irrelevant results, the Web search yielded 258 instances where "Where'd it go?" was used to refer to a piece of information, several of which are shown in Figure 8-5. The topics of the Web pages collected ranged broadly, from technical software languages to teen sleeping disorders. The page format also varied. The data contained ten to twenty instances each of redirect pages (e.g., Figure 8-5(b)), Web logs (blogs), articles, and frequently asked question (FAQ)/help pages (e.g., Figure 8-5(c)). Most of the pages in the collection (165 pages, 64%) were message board and newsgroups postings (e.g., Figure 8-5(a)). The popularity of this format could be because "Where'd it go?" is informal and conversational, and thus appears commonly in informal and conversational settings like message boards.

The most common type of Web-based information being searched for was general Web content (e.g., Figure 8-5(c)). Web sites (e.g., Figure 8-5(b)) and message board postings were also frequent targets. Other less common targets included pictures, message board threads, information to download, and Web functionality (e.g., Figure 8-5(a)). Searches for non-Web information were similarly varied.

The phrase was not exclusively used when someone was unable to locate their target. For example, in 68 instances, or 26% of the total instances, it was used rhetorically. Rhetorical use was particularly common when the phrase occurred in a FAQ or on a redirect or help page. While these instances do not illustrate problems re-finding, they do provide insight into anticipated problems. However, this chapter focuses on how people describe the information they can't find. Thus the numbers reported in the analysis are based on the 165 instances where "Where'd it go?" was used by someone actively seeking a piece of information.

The most common reason the information target being sought was difficult to locate was that another person had changed the target or the information environment surrounding the target (e.g., Figure 8-5(c)). Problems also appeared to arise due to changes that occurred for other reasons, such as due to a Web site outage. There were no instances where "Where'd it go?" was used in reference to information that had changed because it was time dependent (e.g., last week's weather). This could be because people had strong expectations that time dependent information might change, and thus did not expect to be able to relocate it. Difficulties were not always due to the information target having moved, and in 15 instances (9%), it clearly had not. Instead, the seeker was simply unable to locate what was being sought. Consider the following posting, titled "Where'd it go???":

> I must be blind! I posted my intro and first time weigh in - I saw it posted - honest! and now its gone...unless I'm blind! lol Help?????

The posting had not moved, but instead had been posted on a different list than the seeker remembered. Still, the seeker believed the target had moved, and this belief of change, even when inaccurate, was present in virtually all cases.

8.3.3 Describing the Missing Information

How people described information they believed to have moved gives insight into how people cope when changes occur to their information target and its environment. The following section presents analysis of how people described their information target in the 165 instances collected where someone was actively searching for a piece of information.

Expressions of Frustration

People expressed frustration when they could not locate information. In 41 instances (25%), there was a clear statement of frustration, such as "Ahhh *pulls out masses of hair* Where'd it go?!?!" or "where'd it go … i'm panicing". Although there are many reasons why people might have felt such frustration, one explanation that appeared in the data was that losing information made people feel bad about themselves. In 18 of the cases, people who could not find information called themselves stupid or crazy (e.g., "I thought I was going crazy on this one") or assumed blame for their difficulties (e.g., "maybe i'm doing something wrong?"). As will be discussed later, an explanation of why the information target had moved was often a satisfactory answer. This could be because while explanations do not solve the problem, they remove the stress of having lost something and allay the fear of being stupid.

The large amount of frustration observed could also be due in part to the fact that people only went through the effort expressing their difficulties on the Web when a problem was particularly frustrating. Most people do not announce on the Web every time they have difficulty re-finding information. This is supported by the fact that in 13% of the instance (22 times), people who had not originally mentioned having trouble re-finding something agreed when someone else did, saying, "I noticed it too!" or, "I was wondering the exact thing. Where DID it go?"

Shared Context

The phrase "Where'd it go?" often appeared with very little explicit surrounding context. An illustrative example of this can be found in Figure 8-5 (a), where the information target is described only as a "thingy". Similarly, the person who posted the following could not name their target:

> I miss that little tab thingy on my profile that took me straight to my groups...that was convienient! Where'd it go?

Nonetheless, the intended audience in both cases understood what was being referred to, and both received responses. An instance of a particularly cryptic posting was posted under the title "ALRIGHT WHERE'D IT GO!":

> HEY! who thieved the guids to dotb solo'n, and neriad shall solo'n-i knowfaint poitns not the detailed particulars-so uh someone post the url, or email me or somthin

Even this confusing post was understood. Although several expressed puzzlement, one person posted an explanation:

> I do believe she/he is referring to the drums of the beast, and neriad shawl guides, mainly how to obtain each of them solo, most likey either a thread or a link on the old site would be my guess.

Relying on shared context relieved some of the burden from the seeker of expressing their information need. The types of context that were explicitly stated suggest what the seeker considered necessary to specify their target, and the following addressed the more commonly mentioned types.

The Importance of Path

The path via which the target was originally found appeared to be very important, and in 52 of the instances (31%) the path was explicitly mentioned. As an example, 17 times (10%) the query "Where'd it go?" clearly referred not to the asker's information target, but rather to a step along the path to the target. This is illustrated in the following quotation, where the target was a recipe, but the seeker asked for help getting to the containing Web site:

> Okay, where's the link? I wanted to try this quick and delicious recipe everyone raved about

Similarly, someone else asked for a pointer to a newspaper, despite their target being the obituaries it contained:

> Can anyone please provide info on the demise of the Jersey City Observer newspaper? In particular, whether or not it was bought a a competitor, and if so,and as importantly, where it's OBITs and other Personals may have be today?

In Chapter 3, this same behavior was observed for search in general, and several advantages to searching this way were suggested, such as that the source is often easier to locate than the target, and that the source can provide valuable information about the target, such as its trustworthiness.

Time is Relative

Time is often treated as a uniquely important feature in systems that support returning to information (e.g., Freeman & Fertig, 1995; Fertig, Freeman, & Gelernter, 1996; Ringel et al., 2003). However, the instances analyzed in this study did not contain many references to exactly when target was seen before. The temporal aspects of previous interactions with the information target were mentioned in 33 instances (20%), but less than half of those instances made specific references to time in terms of dates or time intervals. When they did, the event usually occurred that same day (e.g., "this morning", "earlire today", "half an hour ago").

Most references to time were vague (e.g., "recently", "earlier", "way back when", not in "quite a while", and not "for some time"). Consider as an illustrative example five different people's postings looking for an online intelligent agent that could be talked to via instant messaging. Only two of the postings made any reference to time at all:

1) OH MY GOD, where is SmarterChild, he's been ofline for a LONG time, and...WHERE DID HE GO?

2) Smarter Child has been offline for some time. What's going on?

It is impossible to tell how long the agent had been missing.

Five times time was referred to in a personal manner, related, for example, to a personal event, as in the quotation in the introduction ("when I first joined these forums"). Regularity of access was mentioned eight times. One person, looking for a Web site that had disappeared, said, "I check it almost every day". Another poster looked for an advertisement seen many times before:

> For awhile now, ive been seeing an advertisement … Now I cant find the Inside Sun advertisment … So, the question is, what happened to it?

Such mentions offer proof that missing information once existed, and that the seeker once knew how to find it.

8.3.4 Answering "Where'd it Go?"

In addition to looking at how people described missing information, the answers people received to "Where'd it go?" requests were analyzed in order to understand how the problems encountered were solved. Solutions ranged from explanations of what had happened, to work-arounds so the seeker could deal with not having the information, to actual resolutions. The three types of solutions (explanations, work around, and resolutions) were not mutually exclusive, and sometimes all three occurred in a single instance.

The question "Where'd it go?" was sometimes anticipated, used rhetorically by information providers trying to ease the re-finding of information they had changed. For example, "Where'd it go?" occurred twelve times in frequently asked questions (FAQs) (e.g., "Retrieving the Office Toolbar – Where'd it go?") and on help pages (e.g., Figure 8-5(c)). Other pages referenced a Macintosh manual's appendix titled "Where'd it go?" The appendix linked common tasks in other operating systems, such as Windows or older Macintosh versions, with the new operating system:

> "Where'd it Go?" is a cleverly conceived reference for OS 9 users. It isn't just some skimpy table that tells you which menu now contains a given command. It's a reasonably good translation dictionary for work habits that includes explanations of the new way to think about the task.

Clearly the problem of re-finding information that has changed is a significant enough problem for people to invest considerable effort helping others deal with it. As such, these instances provide insight into how information re-finding in dynamic environments is currently supported. For example, the fact that people remember the path that they originally encountered information was sometimes taken advantage of. The dataset contained twelve redirect pages (e.g., Figure 8-5(b)), and five "404: file not found" pages. These pages provided information about where and why the target had moved at the site it used to be located. Thus, while the previous analysis focused solely on those instances

where information was actually being looked for, the analysis in the rest of this chapter includes all of the 258 cases where "Where'd it go?" referred to information.

Explanations

The question "Where'd it go?" was often answered with an explanation of where "it" had gone. Even in the absence of an actual pointer to the sought after information, it appears explanations were important in allaying some of the frustration people felt at not being able to re-find information that had moved. Explanations were the most common solution observed, occurring in 33% of the instances (85 times). Explanations were particularly common when "Where'd it go?" was used rhetorically in reference to information that had became unavailable, occurring in 19 out of 23 such cases (83%). For example, all five of the "404: file not found" pages provided an explanation of what had happened to the information, as exemplified by the following:

> I haven't been able to maintain these pages the way I would like to. I've removed the pages I used to have here. If you need a link from one of my old pages, I may be able to retrieve the page from my archives. I'd be happy to send you, via e-mail, any information that was on those pages.

It appeared that explanations were so important that they were often made up. In 38 instances, "Where'd it go?" was asked with a hypothesis of where it had gone. In an illustrative example, someone noted a missing message board with a suggestion for why it might have disappeared:

> Nothing posted after December 6 went onto the board, then today it disappeared completely! Maybe Eric didn't pay his Web page hosting fee.

Replies also often guessed at what might have happened (22 times). While the following is not an explanation of why someone's post had moved, it is a hypothesis:

> Well cindi.....in my experience, if Spike doesn't like how a post is going, or if it is too off topic or controversial, he'll take it out. Which post was it? Sorry!

Explanations often seemed to be sufficient to allay the frustration of the searcher, and people who provided explanations were often thanked, but rarely followed up with. In fact, explanations were sometimes the sole target of the query. This was the case for the quotation in the introduction, and the following is a more extreme instance; here the person created a thread titled "Where'd it go?" despite having already found the target:

> Knox [a server] just seemed to disapear for a couple of minutes and then came back again

These cases where the target was already found highlight the importance of explanations when information moves.

Work-Arounds

Another solution, observed in 22 of the pages analyzed (9%), was to suggest a work around to deal with not having the desired information. For example, someone looking for functionality that had changed asked:

> Where'd it go to? I know I can use guides to manually center elements, but I kinda miss the Center command from FW4.

The respondent pointed the seeker to a worthy substitution, saying, "I found it, or something better, under Window|Align menu." Similarly, a "404: file not found" page suggested alternatives that might be of interest. The page, which once provided satirical content, recommended another Web site with comic information:

> For the time being, I (Pete) reccomend you go here and read some comics, as we all need our daily dose of funny, don't we.

Work-arounds were not always satisfactory, however. This is illustrated in the following instance where the seeker was provided with a successful work-around:

> whatever modules ARE working right now seem to be what i need… but--where'd it go off to? if i do need it in the future, how can i restore it?

In this case, the person still wanted an explanation, and perhaps even a resolution to the problem.

Resolutions

The information being looked for was successfully located in 82 of the cases (32%). An analysis of these instances where the problem was resolved suggests the importance of being involved with the change; when a definitive solution was provided, it was often provided by the person who had made the change. While this obviously occurred regularly when "Where'd it go?" was used rhetorically, it was also common when "Where'd it go?" was used by people actually trying to locate a piece of missing information. Of the 47 instances where people trying to locate information were told where it had gone, ten of the responses were clearly from the person who made the change. In the following instance, the person looking for a posting they had made was pointed to its new location by its mover:

> I moved it to the bug reports forum since it seems to be a bug that is effecting all premium stores.

The person who changed the information also was often the one to provide an explanation of why the information had moved. People trying to locate information received 52 explanations, and 22 of those were obviously from a person involved in the change. As an example, when people asked where a message board posting had gone, it was almost always the moderator who explained that it had been deleted. In another example, someone asked:

> I won the "Jr. Wolfer, 75 posts" contest, but, where did the "Contests and Stuff" section go? And I think the contests idea is pretty good, too. I'm wondering if you got rid of it?

The seeker received an explanation from the person who organized, and subsequently cancelled, the contest:

> Well, it's like that: Being a global moderator needs tons of posts, but the contest only required 75 posts, wich is a very litle number, so i cancelled, and maybe i'll put a new contest soon.

While it was often difficult for people not involved in the change process to locate missing information, people who were involved appeared to maintain a good understanding of the information and what had happened.

8.3.5 Multiple Users of the Same Information

People often had different intentions with the same information, as illustrated by the fact that the most common reason for information to move was another person. As a result, several interesting problems worthy of further investigation arose. For example, sometimes information was removed because people in general were not interested in it, despite the information being of interest to the seeker:

> I think they got removed because there were only about three of them, and they got old fast

Information was also sometimes removed because the information provider actively did not want the information to continue to be available. For example, the author of the following quotation references a previous posting he did not want others to be able to read:

> I was hoping nobody saw it, oops. I got taken in by that Metallica spoof going around the net. I found out it was a parody site so I deleted [the posting].

This same conflict was also evident in the seven instances when information was removed for copyright reasons:

> [T]he French site Elostirion.com was asked to take down the image of the Ringwraiths. You can still read the news on this story from this morning which ended with the confirmation of these characters in fact being uncloaked ringwraiths.

The conflict of interest between information users, who want the information they interact with to be persistent, and information providers, who want control over their information, is related to copyright issues that have arisen in making digital library documents persistent (Hearst, 1996).

Another interesting conflict that arose was highlighted by the large number of message board postings that went missing because they were deleted by moderators:

> The Web site you list is commercial and is the reason your post was removed. I have now edited out the site so you will understand. Please read the goals and rules of posting on sleepnet.com forums.

In these cases, the people looking for their past postings were not interested in finding the information for themselves, but rather in ensuring that others could see it. This was in direct conflict with the information providers, who had removed the posting because they explicitly did not want the content to be available.

Search engines also face problems because people have different intentions with the same information. Chapter 4 showed that individuals used the same query to refer to different information needs, and this motivated the personalized search system in Chapter 5. This chapter has demonstrated that different individuals' past experiences with information create different expectations when returning to information. The following chapter presents a system that personalizes results based on these expectations to allow users to find new information while not becoming confused, disoriented, or unable to find information they have seen before.

That which is static and repetitive is boring. That which is dynamic and random is confusing. In between lies art.

- John Locke (1632 - 1704)

Chapter 9

Supporting Finding and Re-Finding via Personalization

Based on the previous chapters, it appears that finding and re-finding are at odds with each other. The tools that best support finding present their users with new and relevant information, but doing this hinders the users' ability to re-find old information. The tools that best support re-finding present information in a consistent manner, but doing this hinders the users' ability to find new information. It is not enough to develop tools that support only one behavior or the other, as analysis of the Yahoo logs suggested people use search tools to locate old information and new information at the same time.

Fortunately, it is possible to reconcile these seemingly conflicting goals by personalizing search results not by ranking the most relevant results first, but rather by ranking them where the user expects them. In this chapter, the efficacy of this solution is demonstrated through the implementation of a system called the *Re:Search Engine*.

Recall the motivating example presented in Chapter 1. Connie was diagnosed with breast cancer and searched for a list of treatment options using the query "breast cancer treatments". As a result, she developed expectations about the result list returned for that query. Although naïve personalization can provide her with objectively better results, if the results do not match her expectations, the previous chapters suggest the new results will hinder her search. Levy (1994) observed, "[P]art of the social and technical work in the decades ahead will be to figure out how to provide the appropriate measure of fixity in the digital domain." One way to achieve the appropriate measure of fixity may be to take advantage of the fact that people do not remember much of the information that is presented to them.

Figure 9-1 illustrates the way Connie's lapses in memory can be used to her advantage. Because she only remembers a few results in the original list she saw, those few results can be held constant, while the new personalized results can be added. The merged list is likely to look the same as the old list to her, despite containing new, useful information, and thus can satisfy both her re-finding and new-finding needs.

The chapter begins a description of several studies used in the construction of the Re:Search Engine. Then the details of how the Re:Search Engine merges new results

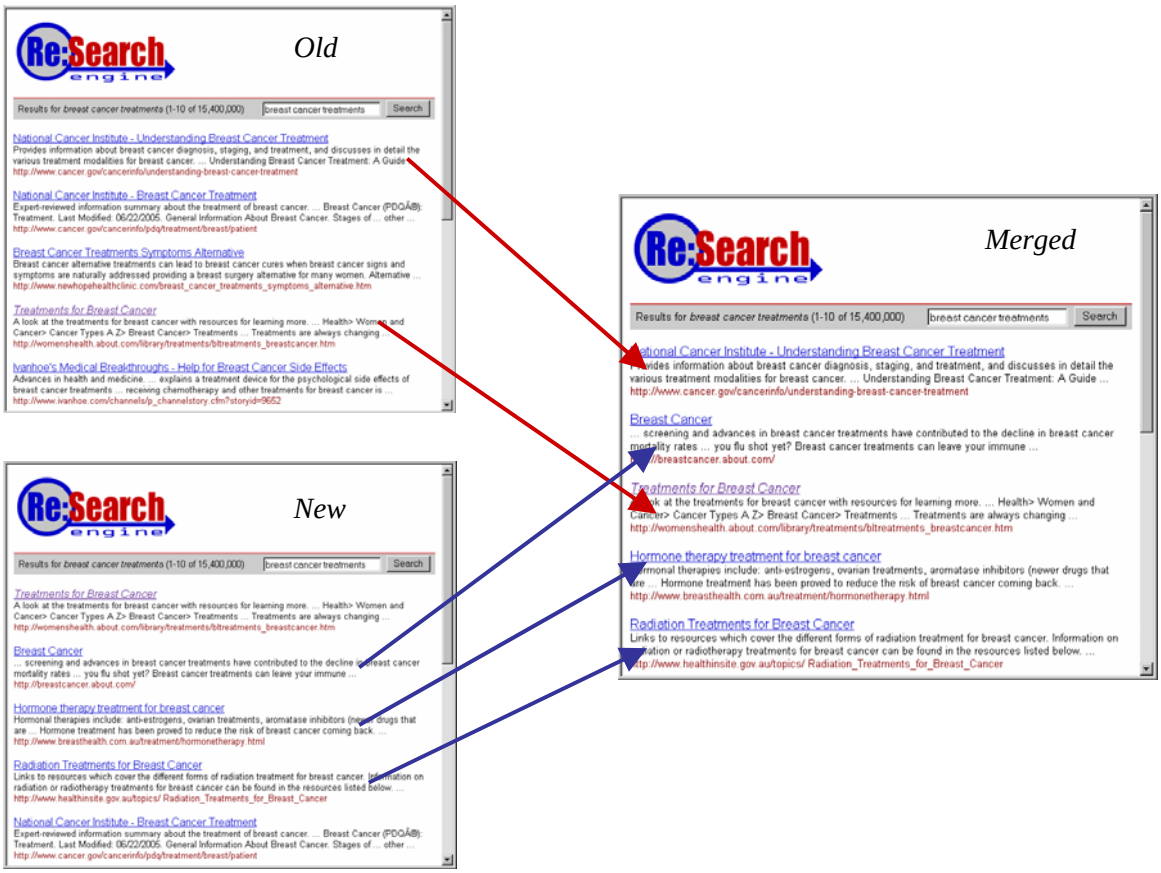


Figure 9-1. The Re:Search Engine in action. The result page labeled "*Old*" shows the results from when Connie first searched for "breast cancer treatments". The page labeled "*New*" shows the results when the query is next performed, personalized to rank the most relevant first. The "*Merged*" results shows how the Re:Search Engine combines what Connie is likely to remember having seen during her initial search with what is new in the personalized results.

into a previously viewed list are presented. Finally, the resulting system is demonstrated to successfully permit its users to interact with dynamic information without losing the context they develop through such interaction.

Studies Used to Build the Re:Search Engine

Two studies were conducted in the construction of the Re:Search Engine. This section begins with a discussion of a paper prototype study of a system that incorporates old and new information. The study was conducted to gain insight into the feasibility of the Re:Search Engine. The study highlights that it is important for the Re:Search Engine to understand and preserve the memorable aspects of previously viewed search result lists. For this reason, this section then presents the results of a study of what is memorable about search result lists. These findings inform the construction of the Re:Search Engine's architecture and underlying algorithms.

9.1.1 Paper Prototype Study

A paper prototype study was conducted to explore whether it is feasible to invisibly change information with which people have previously interacted. The study revealed that it is possible to change large amounts of information without the notice of participants, and, more importantly, with them still feeling in complete control of the information. Here the study methodology is briefly introduced, and the results discussed.

Study Methodology

Paper prototyping (Rettig, 1994) is a good way to test new interface ideas quickly, cheaply, and easily. A paper prototype study was conducted to observe how people interact with changing document clusters, and tested informally with fourteen participants. Half of the participants were male, and half female. Their computer experience ranged from limited to extensive, and their ages ranged from 21 to 55. This section describes why document clustering was studied and how clustering relates to the display of search result lists. It then discusses the implementation of the paper prototype.

The Relevance of Clustering

The paper prototype study was performed with a sample problem of clustering because document clustering algorithms create many of the same problems re-finding that are created by search engines. Both clustering algorithms and search engines involve lists of documents that can change as new information becomes available. Clustering algorithms are often able to roughly group documents into initial clusters quickly, but take more time that the user is willing to wait to create a good clusters. This paper prototype study looks at a situation where an initial quick-and-dirty clustering is presented to the user, and then is cleaned up as the user interacts with the information. Clusters of documents can be seen as similar to search result lists in that each cluster contains a result list of information that matches that cluster's topic. There are also many situations where it may be of interest to present the user with a quick-and-dirty first pass search result ranking, and then update that ranking as the participant interacts with it. For example, for a meta-search engine, not all search engines being queried may return results quickly, or for a search engine that uses implicit relevance feedback, the information gathered about the user during the user's interactions with the results may lead to improvements to the result ranking.

Clustering Implementation

Clustered documents were presented in lists of hyper-linked document titles and summaries, much as search results are displayed. Each cluster was represented by a tabbed pane. Clusters were described by a set of keywords and represented by a unique color. The documents in each cluster were ordered by their relevance to the cluster. Clusters, document ordering, and the words that represented the clusters were created as realistically as possible with the support of simple clustering algorithms (e.g., k-means) and simple information retrieval measures of the importance of words. A mock-up of the paper prototype can be seen in Figure 9-2.

The interface shown in Figure 9-2 was implemented in paper using the items shown in Figure 9-3. Figure 9-3(a) shows several examples of document title and summaries,

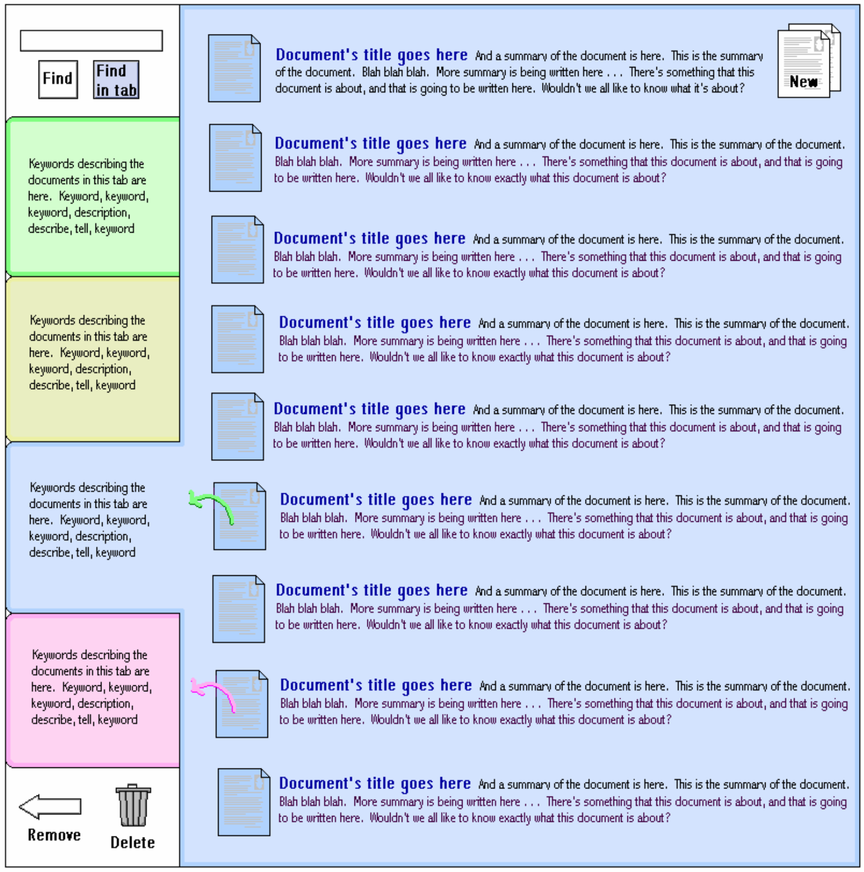
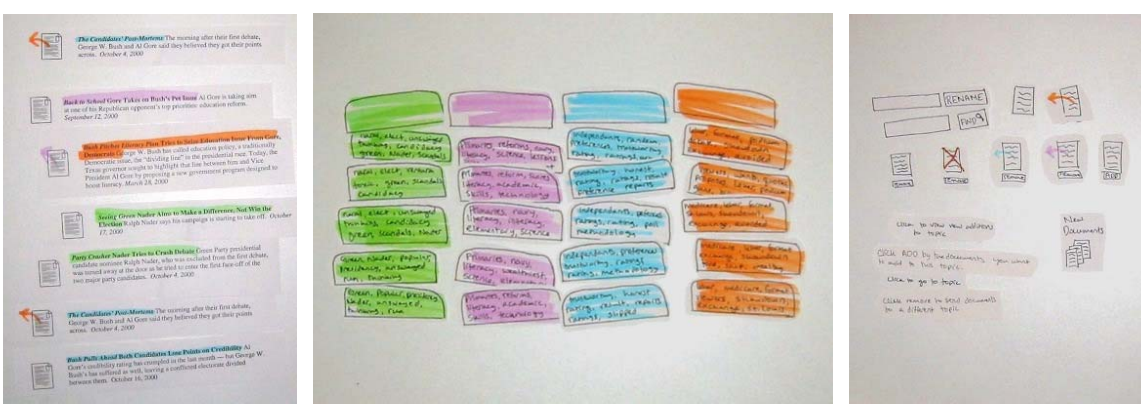


Figure 9-2. Mock-up of the paper prototype.



(a) (b) (c)

Figure 9-3. Example items used during paper prototype study, including a list of documents and summaries (a), a number of tabs with different wordings (b), and various widgets, search boxes, and mouse-over help (c).

Figure 9-3(b) shows some example tabs, and Figure 9-3(c) shows some example widgets, such as search boxes and mouse-over help. These items were manipulated as the user interacted with them so that they behaved like a real computer. For example, when a participant "clicked" on a "hyper-linked" document title by tapping on it, the participant was given a piece of paper representing the document.

As participants interacted with the clusters of documents, changes to improve the clustering were introduced at regular intervals (e.g., the start and end of each task). Changes included changes to document ordering, changes to which documents belonged to which clusters, and changes to the words on the tab. Changes were grouped into one of two categories: 1) changes that were likely to be noticed by the user, and 2) changes that were not likely to be noticed by the user. Changes that were likely to be noticed by the user (Category 1) were brought to the user's attention and permission was needed before they happened. On the other hand, changes that were not likely to be noticed (Category 2) were permitted to happen as necessary.

If a scheduled change occurred to a piece of information that had not yet been viewed, it was considered to belong to Category 2, and such changes were allowed to happen freely. Since many of the instances where the underlying information is changing involve very large collections of data, it is quite likely that the user will never see most of it, making the issue of how to maintain the user's context trivial. For example, there is no reason for a search engine to preserve consistency in the results it displays to a user for queries that the user has never issued before. In the paper prototype study, unseen information included information in clusters that were not visited and information that was not ranked highly enough to be displayed.

However, it is possible that even changes to information that had been previously viewed could pass unnoticed. For example, a change in the result ordering between two unread documents whose summaries had been viewed may belong to Category 2 because the participant does not really remember anything about the documents even though they have been displayed. Paper prototyping was used to understand to which category various different changes belonged.

Tasks

The participants were asked to complete four tasks with the document clusters. The first three tasks were designed to encourage interaction with information in each of the clusters. The documents being clustered were about the 2000 United States presidential election, and tasks included, for example, "Compare the education platforms of Bush and

Table 9-1. Words on the paper prototype tab related to the Green Party.

Time Period	Cluster Wording
Initial	naral, elect, ventura, toxic, green, scandals, candidacy
Start of Task 1	naral, elect, unswayed, thinking, candidacy, green, scandals, nader
Start of Task 2	naral, elect, unswayed, thinking, candidacy, green, nader, scandals
Start of Task 3	green, popular, presidency, nader, unswayed, thinking, run
Start of Task 4	green, nader, popular, presidency, unswayed, run, thinking

Gore." The fourth task was intended to encourage re-finding, and asked the participant to find a memorable document from one of the three previous tasks.

Specific changes to each cluster were associated with specific points in each task. For example, the changes to the tab representing documents about the Green Party and Ralph Nader happened at the start of each task, and can be seen in Table 9-1.

Results

Participants expected to be able to re-find previously viewed information where they had seen it before. Users developed *conceptual anchors* into the previously viewed information, and when they wanted to re-find that information they used those conceptual anchors to retrieve it. If the conceptual anchors in the clustering implementation were preserved, the other information could change as needed. For example, a cluster is described by a set of keywords that are extracted based on common word occurrences within the documents contained in the cluster. Participants seemed to build a initial understanding of the content of each cluster from the keywords, and did not notice subsequent small word changes within the keyword list.

Participants did not seem to notice the order in which the documents were displayed, as long as the first document remained first, and all of the visible documents remained visible. Participants did care about which cluster a document was located in when first seen, but did not seem to mind if it later also showed up in a cluster where it was related by content.

It appears interfaces can be structured to support the construction of conceptual anchors. Each cluster was represented with a unique color. Participants quickly associated colors with clusters, and relied on the color to navigate rather than keywords. As evidence, one participant noticed a change to the words on a tab describing a particular cluster. When asked what she noticed, she said, "I noticed that the order of the words on the tab changed, not the color." Her statement suggests things like color and word order are relatively important, while things like the addition or subtraction of words (a change which she missed) are less important.

Often when changes were noticed, participants did not attribute the change to the system, but rather to themselves. For example, one participant, noticing a new and relevant word on a tab said, "Here's something I didn't notice before." Although large amounts of information changed during testing, test subjects generally expressed a feeling of control over the information, and often articulated surprise when informed that they had been working with information that was changing. "You say information was changing," one participant said, "but I did not feel like it was changing." Another participant said, "I'm in control of my own computer, I can decide to put things anywhere," apparently unaware of the fact that he had not been the sole agent deciding where things belonged.

9.1.2 Study of Conceptual Anchors in Search Result Lists

The paper prototype study suggests that search result lists can change as long as the conceptual anchors people develop into result lists are maintained. Thus it is important to

understand what common result list anchors are. To assist in the development of the Re:Search Engine, a study was conducted of which aspects of a result list are memorable, and thus are likely to be noticed when they change, and which are not memorable, and thus are unlikely to be missed if change. In this study, participants were first asked to interact with a result list and then later asked to recall the list. Based on the data collected, a model is developed of which aspects of a result list are memorable.

Study Methodology

To discover what people found memorable about search result lists, participants were asked to interact naturally with a list of 8 to 12 results for a self-generated query. While typical studies of list memory such as those discussed in Chapter 6 require all items to be attended to, participants were not required to view every result. In fact, studies suggest that participants were unlikely to view all of the results, and likely spent considerably more time looking at the first result in the list than at subsequent results (Granka, Joachims, & Gay, 2004).

By allowing natural interaction, the study revealed which aspects of the result lists were both attended to and recalled. Queries were issued to a leading Internet search engine via a Web form accessed from the participant's own computer, and clicked results were logged. After an interval of a half hour to an hour, participants were emailed a brief survey (shown in Figure 9-4) that asked them to recall their result list without referring back to it. Participants were asked to recall their query, the number of results returned, and basic information about each result, including its title, snippet, URL, whether it was clicked, and if so, what the Web page was like.

Approximately half of the 245 people who issued an initial query completed the follow-up survey. Removing survey responses that were clearly erroneous (e.g., the remembered query and results didn't match the initial query and results at all, or the titles, snippets and URLs were remembered exactly, suggesting the participant referred back to the result list while completing the survey) yielded 119 responses for analysis. A relatively even number of men (52%) and women (45%) responded to the follow-up survey. Most (64%) were between the ages of 25 and 39, but 18% were over 40, and 15% under 25. Ninety-seven percent reported daily computer use. Only 27% of respondents were affiliated with MIT. Typically, the follow-up survey was completed within a couple of hours of the initial search. Sixty-nine percent of all responses were received within three hours of the initial search, and all but five were received within a day.

The observable search behavior captured by the study was similar to behavior commonly observed for Web search. The average initial query length was 3.2 words, somewhat higher than, but comparable to, what has been found by others through Web query log analysis (Spink et al., 2001; Xue et al., 2004). When interacting with search results, participants on average followed 1.9 results, and this is comparable to the 1.5 clicks per result page observed by others on considerably larger datasets (Xue et al., 2004).

Because the recalled rank for a result did not necessarily correspond with the result's true rank, it was necessary to manually match the description of each recalled results with an actual result. Some results were easy to match, while others were impossible. For example, a result for the query "shakespeare sites" was described merely as "something shakespeare" – not very descriptive given all of the results in the original list had

something to do with Shakespeare. Two independent coders matched recalled descriptions to actual results, with an 84% inter-rater reliability. All of the data collected is included in the following analysis whenever possible, but if the real rank of a result is required only the 189 result descriptions on which the coders agreed are used.

Just as rank was often remembered inaccurately, so was the number of results returned in the initial result list. The correlation coefficient between the actual number of results returned and the number of results recalled was 0.09, which is not significantly different from 0. Not surprisingly, a large majority of the participants assumed ten results were returned regardless of how many actually were. Although participants received anywhere from 8 to 12 results, in order to clearly illustrate primacy and recency effects (discussed in Chapter 2) the figures below present data for exactly ten results – the first five results from the beginning of list, and the last five results from the end of the list.

The data collected through this study was analyzed to provide insight into how to recognize memorable results, and how to understand the relative likelihood that various different types of changes that could occur in a result list would be noticed.

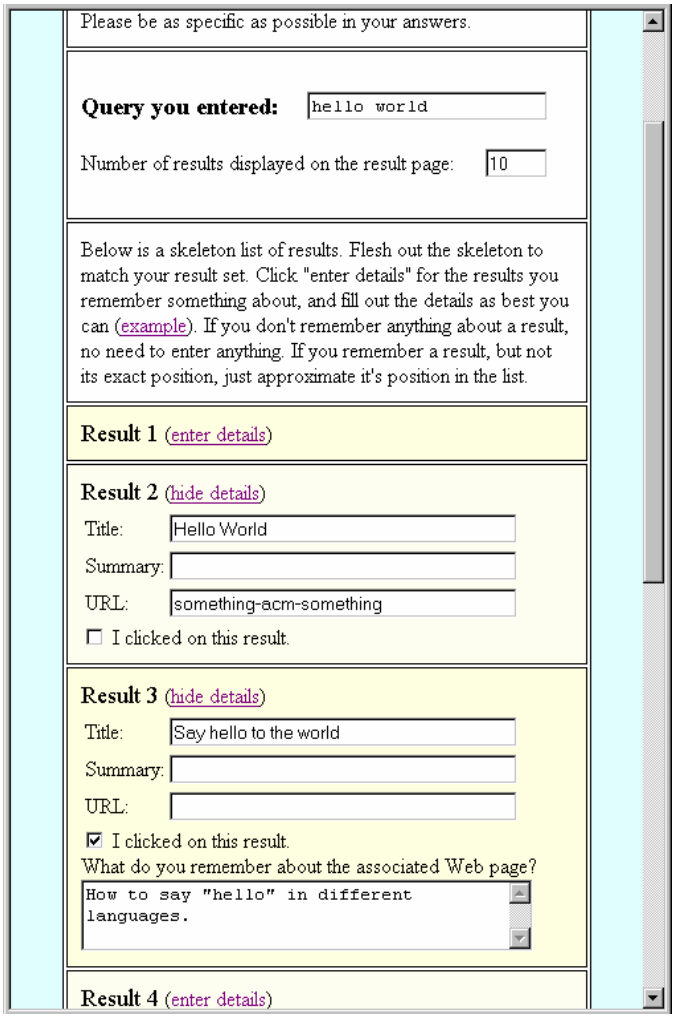


Figure 9-4. The follow-up survey asking participants to recall previously viewed results.

What Makes a Result Memorable

In general, participants recalled very little about the search results they interacted with. Even though only a few hours elapsed between the time when the search result list was originally seen and the follow-up survey, only 15% of all results viewed were described in any way. The majority of participants remembered nothing about any of the results (mode=0), and on average, participants described only 1.47 of the results from the original list.

A result was considered memorable if it was sufficiently described so that the two independent coders agreed on which result the description referred two. Two main factors emerged from the data as affecting how memorable a result was: where in the result list it was ranked and whether or not the result was clicked.

Rank Affects How Memorable a Result Is

As suggested by the previous studies of list memory discussed in Chapter 6, the likelihood of a result being remembered was affected by where in the result list the result occurred. Figure 9-5 shows the probability that a result was remembered given the result's rank for results that were clicked (solid line) and results that were not clicked (dashed line). The general shape of the curves is similar to what is seen in cognitive psychology literature. Those results presented first are memorable compared to later results (similar to the primacy effect) and those results presented last are somewhat more memorable than earlier results (similar to the recency effect).

Highly ranked results appeared particularly memorable. This is probably because top ranking search results get significantly more attention than lower ranked results that require scrolling to view. Results "below the fold" (typically result 7 and below) are often never seen at all. Further, people tend to believe highly ranked results are more

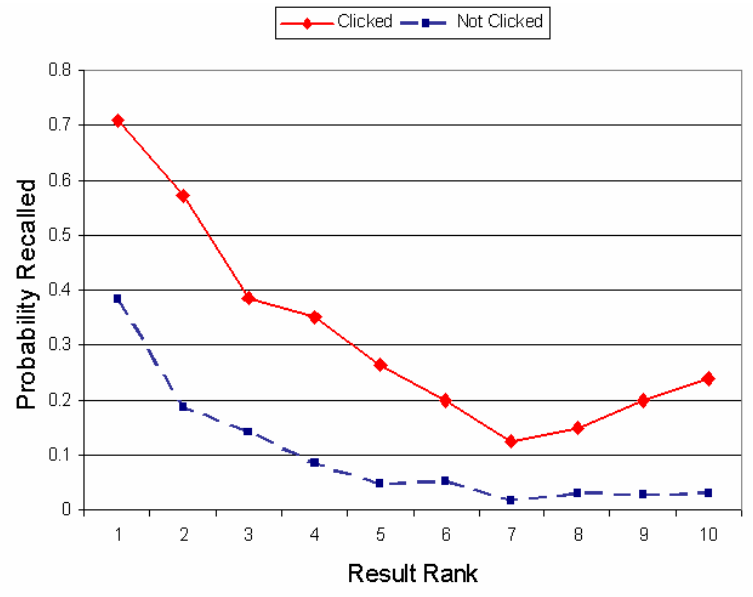


Figure 9-5. The probability of recalling a result given rank. The probability generally decreases as a function of rank. Clicked results were significantly more likely to be recalled ($p<0.01$).

relevant to their queries, regardless of whether the results actually are more relevant or not (Joachims et al., 2005). Results thought to be highly relevant probably stand out because they are of direct interest and distinctive (similar to the von Restorff effect).

While the last results in the list appeared more memorable for clicked results, such was not the case for results that were not clicked. This discrepancy may be because low ranked results that were not clicked were also often not read. Thus the last result seen for non-clicked results varied as a function of the individual (e.g., the resolution of the participant's screen, how likely the participant was to review all of the results, etc.).

Clicked Results are More Memorable

In addition to rank, whether a result was clicked or not affected how likely the result was to be remembered. As discussed in Chapter 2, the importance of click through data has been studied for its value as an implicit measure to determine search result quality. In this analysis, click through is looked at as a way to determine how likely a result is to be remembered.

A clear correspondence was observed between how likely a result was to be remembered and whether it was clicked. On average only 8% of results that were not clicked were recalled, where as significantly more results that were clicked were recalled (40%, $p<0.01$). The greater likelihood of a result that was clicked being recalled can be seen graphically in Figure 9-5.

The last result that was clicked was even more likely to be memorable. A 12% increase in recall was observed if the result was the last clicked, compared to other clicked results. The last result clicked may be more memorable because it was probably also the last result seen (similar to the recency effect), and because it was what the participant was looking for and thus distinctive (similar to the von Restorff effect).

Other Factors Affecting Recall

The number of times a result was visited also appeared to affect its likelihood of being recalled. A common comment that accompanied a result recalled with great detail was that it pointed to a Web page that the participant visited often. One participant remembered the exact URL for a flamenco Web site, and her ability to accurately remember the URL is perhaps explained by her description of the Website as a "good flamenco Website I have used before." Such comments were common despite the fact that participants were not explicitly asked to share how often they visited a particular result.

Some results were recalled that did not occur in the original result list. For example, a participant remembered a result with the URL "www.mlb.com" as being returned for the query "pittsburgh pirates", when in reality no such result was listed. Such phantom results probably represent pages that were found via other means but that conceptually belonged in the search result set.

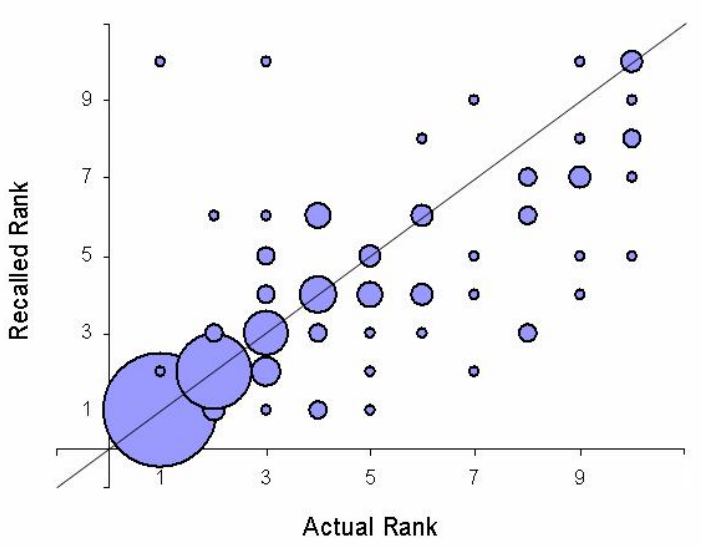


Figure 9-6. The result's location in the result list as the participant remembered it, compared with the result's actual location. The size of each point represents the number of people remembering that combination.

How Result Ordering Was Remembered

How result ordering was remembered was also analyzed to give insight into how changes to ordering might affect one's ability to interact with a search result list. Participants were often observed mistakenly recalling a result's rank. Recalled rank differed from actual rank 33% of the time. Mistakes tended to be less common for highly ranked results, and the first result's rank was remembered correctly 90% of the time. Accuracy dropped quickly as rank dropped, as can be seen graphically in Figure 9-6. This implies that moving a result from the number one position in a result list is more likely to be noticed than moving a lower ranked result.

Figure 9-6 also illustrates another interesting trend in the data. The greater weight of the data occurs to the right of the identity line. This means that remembered results were much more likely to be recalled as having been ranked higher than they actually were. Recalled results moved up in the result list 24% of the time, significantly more often than they moved down (10% of the time, $p<0.01$). The trend to remember results as highly ranked probably reflects the fact that remembered results were more likely to be relevant to the participant's information need and thus in the participant's mind "should have been" ranked more highly than they actually were.

Although psychology literature suggests that relative ordering is important for recall (Henson, 1998), swaps in result ordering occurred in 10% of the responses where a swap was possible (i.e., a participant's remembered at least two responses).

It is interesting to consider the ramifications of the fact that people misremember result ranking and that people sometimes recall phantom results as having occurred in the list even when they didn't. These findings suggest that it is actually possible for a result list to look *more* like the result list that a person remembers having interacted with than the

actual list they did interact with. In the studies presented later in this chapter, there is a trend for the Re:Search Engine results appear unchanged more often than static result lists. While these findings are not significant, it could be a result of the Re:Search Engine doing a good job of placing results where they are expected – even when that is not where they originally occurred.

9.2 Re:Search Engine Architecture

The Re:Search Engine was designed using the results of these two studies. The engine consists of a Web browser toolbar plug-in that interfaces with a preexisting search engine (e.g., Google, Yahoo, or a personalized search system). When a person issues a query that they have issued before, the Re:Search Engine fetches the current results for that query from the underlying search engine and merges the newly available information with what the user is likely to remember about the previously returned search results.

The architecture design of the Re:Search Engine is shown in Figure 9-7. The user's query is passed through an index of past queries that the user has issued. The index returns queries that are similar to the one just issued and a score for each past query that represents how similar it is to the current query. These queries are then used to retrieve from a cache the results the user viewed when searching for each past query. This set of results, along with the live results for the current query from the underlying search engine, are merged together, using the query scores to weight how important each different result set is. The resulting list of search results is presented to the user. Finally, the query the user issued is added to the index of past queries and the merged result list is added to the result cache. Each of these components is described in greater detail below.

All of the data the Re:Search Engine collects resides locally on the user's machine. This has the disadvantage of tying the use of the Re:Search Engine to a particular machine, but

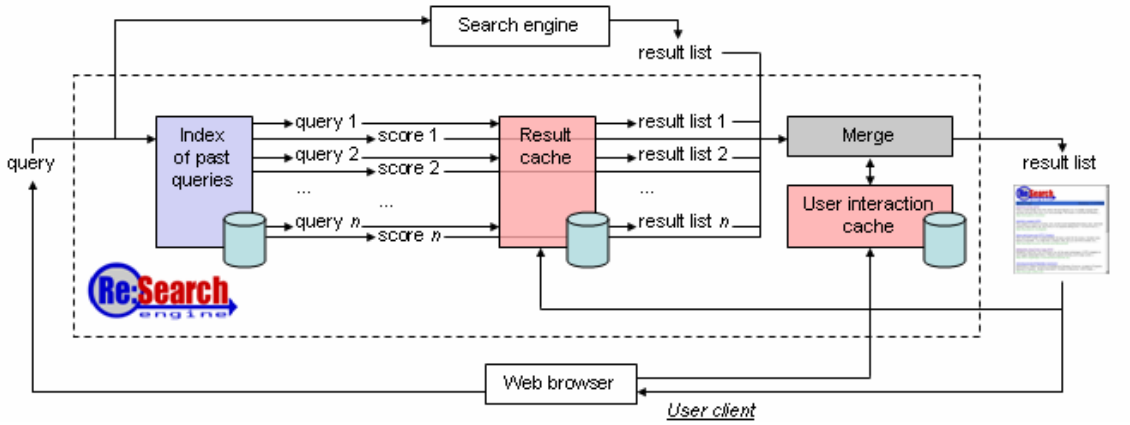


Figure 9-7. The architecture of the Re:Search Engine. The user's current query is matched to past queries, and the results for the past queries are retrieved from a cache. These results are then merged with the live search engine results based on how memorable the results are, and the resulting result list is presented to the user.

such a design decision ensures that the relatively large amount of personal information that the Re:Search Engine stores will remain private.

9.2.1 Index of Past Queries

This section discusses the index of past queries. When a user first issues a query to the Re:Search Engine, the query is passed through an index of past queries the user has issued in order to determine if the user intends for the current search to retrieve previously viewed information, and, if so, what queries were previously used to retrieve such information. Once queries that are similar to the current query have been gathered, the current query is added to the past query index.

The index of past queries functions in a similar manner to a traditional document index used in information retrieval, except that the "documents" that are indexed are past query strings. Query strings are stemmed, and stop words and capitalization are removed. Matching queries in this manner covers many of the common changes to queries observed Chapter 7, such as changes to word order, white space, capitalization, and word form.

Each past query (*pq*) is given a score based on how closely it matches the current query (*cq*). The score is computed using a standard information retrieval scoring function known as *tf.idf* (Salton, 1998):

$$\text{Score}_{pq} = \sum_{t \text{ in } cq} pq_t \log(N/n_t) \qquad \textbf{(6)}$$

where N is the number of past queries the user issued, and n_t is the number of past queries in which term t occurs. This scoring function reflects the fact that past queries that match due to terms the user searches for rarely are more likely to mean the same thing than terms that match because of terms the user searches for often.

Chapter 7 showed that not all queries with similar text are repeat queries. In fact, if a user is in the middle of a search session, it is likely that if several variants of the same query are issued, the user actively wants to see new results with each variant. For example, if Connie thought the results she received for her query for "breast cancer treatments" returned too many alternative or experimental treatments, she might try searching for "established breast cancer treatments". This search should not merge in the results for the query issued immediately prior. For this reason, past queries that are similar but that occurred recently are ignored.

Although the index of past queries permits flexible query matching, the Re:Search Engine's interface is designed to encourage users to communicate re-finding intent by encouraging them to exactly duplicate previously issued queries. The index of past queries is used to support sophisticated query completion in the search box. Thus if Connie begins typing, "cance..," into the search box, her previous query for "breast cancer treatments" will be suggested.

9.2.2 Result Cache

If the query the user issued is determined to be related to one or more previous searches run by the user, the results corresponding to the previous searches are fetched from a result cache using the pervious queries returned by the past query index. The result cache is a straightforward cache that maps an exact query string to the search results list presented to the user for that query. Only the most recently viewed set of results for a particular query is stored in the cache. For example, when Connie issues the query "breast cancer treatments" a second time, the merged results shown in Figure 9-1 replace the old results in her result cache.

9.2.3 User Interaction Cache

Once past results that might be relevant to the user's current query are fetched, they are merged with the live search results to produce a result list consisting of old and new information to return to the user. Because the merge algorithm is designed to help users take advantage of the context built on the query topic during past searches, it requires an understanding of how memorable the past results the user interacted with are. The user's browser is instrumented to gather implicit information about memorability, and the interactions are stored in the user interaction cache.

Currently the user interaction cache only stores past results that the user clicked on, but there are many other possible implicit cues that could use to understand which results are memorable to the user. Possible cues worth investigating include dwell time on the result's Web page, the number of times a particular result is accessed, and more sophisticated measures such as mouse tracking or eye tracking. Additionally, active information could be gathered. For example, the system could be extended to allow the user to mark results that they believe are worth remembering.

9.2.4 Merge Algorithm

Preserving the feeling of consistency for an individual interacting with changing information requires that the memorable information, or conceptual anchors, be maintained. Merging new results into an old result list requires that the value of the new information be balanced with the cognitive cost of changing the already viewed information. This section describes how this is done.

Valuing a New Result at a Particular Rank

Each new result that could be returned is given a *benefit of new information* score based on the expected benefit the as-yet-unseen result will provide to the user. If scoring information is available from the underlying search engine, the result's score can be used to represent the expected benefit, otherwise the result's rank can provide a proxy. The value of the new information is a function both of how likely the new information is to be relevant to the query (the benefit of new information score) and how likely the information is to be seen. Because results are more likely to be seen if they are ranked towards the top of the final result list, the total benefit of the new information included in a result list is based both on each new result's benefit and the result's location in the list.

Valuing an Old Result at a Particular Rank

Each result in the original list is assigned a *memorability* score that represents how memorable the result is. This score is based on whether the result was clicked and its rank in the result list, using a smoothed version of the probability of a result being remembered given the rank and whether it was clicked (see Figure 9-5). The value of the old information included in the final result list is a function both of how memorable that information is and how likely the information is to appear where expected. For this reason, old results are assigned a *cost of change* that represents how likely the result is to be remembered in a particular location. This is calculated using a smoothed version of the probability of a result of a particular rank being recalled at a different rank (see Figure 9-6). The value of an old result at a particular rank is a function both of its memorability score and its cost of change.

Choosing the Best Possible List

During the merge process, all permutations of possible final lists that include at least three old results and three new results are considered, and the result list with the highest total benefit of both old and new information is selected. The value of a list is a function of the benefit of new information it contains, and the memorability of the old information. There is obviously a trade-off between preserving a lot of the information that was originally seen and presenting as much new information as possible. Requiring that both old and new results be included in the final list ensures that some context is maintained while not allowing the list to stagnate.

Although considering all permutations of possible result lists naively is expensive, the merge algorithm can be implemented efficiently by representing the problem as a min-cost network flow problem (Goldberg, 1997), as shown in Figure 9-8. Ten unites of flow are sent through the graph, each unit representing one result in the final result list. Seven units are sent passed to nodes representing the new results, and seven are passed to nodes

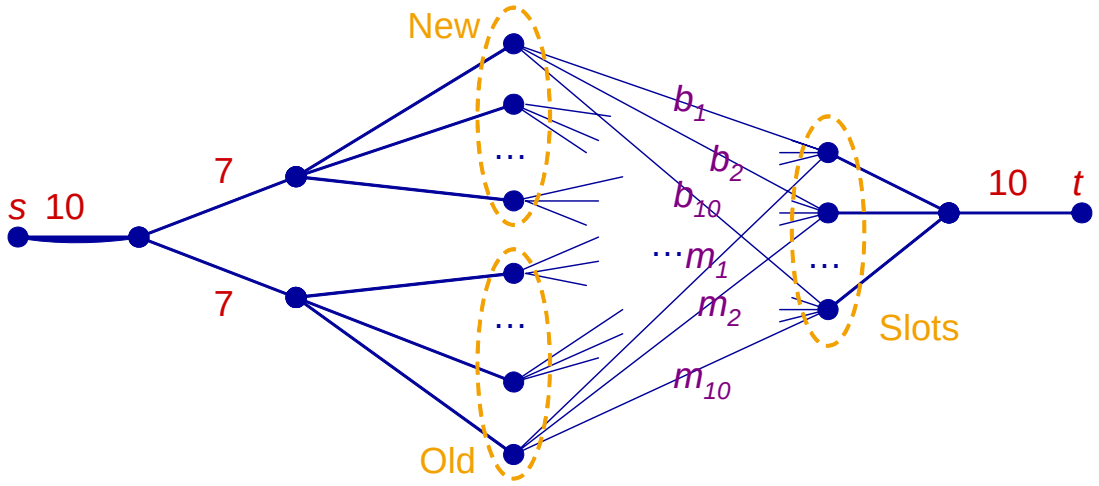


Figure 9-8. Graph representation of the merge algorithm. All edges have unit flow, with the exception of the edges labeled in red. All edges have zero cost, with the exception of the edges connecting the nodes representing the new and old results to the slots.

Table 9-2. Rank of old and new results after merging.

***Merged* Rank**	**Results clicked in original result list:**		
	None	9	1, 2, 6, 8
1	Old result 1	Old result 1	Old result 1
2	Old result 2	Old result 2	Old result 2
3	Old result 3	Old result 3	Old result 3
4	Old result 4	*New result 1*	*New result 1*
5	*New result 1*	*New result 2*	*New result 2*
6	*New result 2*	*New result 3*	Old result 6
7	*New result 3*	Old result 9	Old result 8
8	*New result 4*	*New result 4*	*New result 3*
9	*New result 5*	*New result 5*	*New result 4*
10	*New result 6*	*New result 6*	*New result 5*

representing the old results. This ensures that at least three units must pass through the old results and at least three units must pass through the new results.

The nodes representing the new results are connected to the ten slots representing the final result list with unit capacity edges that have a cost equal to the inverse of the benefit of the new information. The nodes representing the old results are similarly connected to the ten final result lists slots with unit capacity edges that have a cost equal to the inverse of how memorable the results are. All other edges have zero cost.

The cost of change and the benefit of new information can be weighted to express the relative value of new and old information. This weighting should be a function of the individual using the Re:Search Engine, the elapsed time since the original list was seen, and the certainty that the person wants new information versus old information. In the implementation tested, when no results were clicked the merging produced a list that began with four old results and ended with six new results. When low ranked results from the original result list were clicked, the clicked results were preserved in the new merged result list while higher ranked previously viewed results were dropped. Several examples of merged lists are shown in Table 9-2.

For simplicity, users are assumed to remember perfectly which result page a result occurred on (e.g., whether the result occurred in the top ten, or in results 11-20). Because the results for a query are never expected on a different result page than where they were seen, each old result page can be treated independently of other result pages during the merge. The highest ranking new information available is always is merged in, regardless of what particular page is requested.

9.3 Performance of the Re:Search Engine

How well the Re:Search Engine performs can be understood on several levels. At the most basic level, the engine can be said to work if the merged list of results looks unchanged to its users. This is studied by comparing the merged results with the remembered results. The Re:Search Engine can also be tested by looking at whether the

merging makes re-finding easier while not interfering with finding. This is studied through a controlled study of finding and re-finding. To truly understand whether the Re:Search Engine improves the search experience overall and how it affects search behavior requires a longitudinal study, and this remains as future work.

9.3.1 Comparing Results with Remembered Results

To test how well lists merged by the Re:Search Engine preserve the memorable aspects of the original search result lists, a study was run to investigate how often a changed list looked unchanged to someone who had interacted with the original list. The study showed that intelligent merging enabled new results to be included virtually unnoticed while changes included with naïve mergings were often noticed. In this section, the study methodology is presented and the study results are discussed.

Study Methodology

The effectiveness of the merged result list in masking changes was tested with a similar experimental setup to the first study. The primary difference in this second study is that participants were asked to recognize their original result list rather than recall it. Participants again ran an initial search using a self-generated query and were asked to interact with the result list as they normally would. A half hour to an hour later, they were emailed a pointer to a follow-up survey, shown in Figure 9-9. In the follow-up survey participants were presented with a new search result list and asked whether it was the same as the list they saw initially or not. If the participant noticed a change, he or she was asked whether the changed results were better, the same, or worse, and asked to describe any differences noticed.

The follow-up result list was drawn from one of five cases:

> ***Original*** Result list is exactly the same as the original result list (results 11 through 20 for the query from the underlying search engine).
>
> ***New*** Result list comprised of entirely new results (the top ten results from the underlying search engine).
>
> ***Dumb Merging – Random*** Four results were randomly preserved from the initial result list, and the rest were changed. Four results were chosen because that was the average number of results preserved in the *Intelligent Merging* case.
>
> ***Dumb Merging -- Clicked*** Results that were clicked during the participant's initial interaction with the original list are ranked first, followed by new results. This case represents an attempt to put the most relevant results from the original list first while providing new information.
>
> ***Intelligent Merging*** Old and new results merged according to the Re:Search Engine's intelligent merging algorithm so that memorable aspects of the old results were preserved. This is the case being tested.

Note that the original search results participants were presented with were not the best possible results. Instead of initially returning the top ten results from the Web search engine, results ranked 11 through 20 were returned. The new search results that were

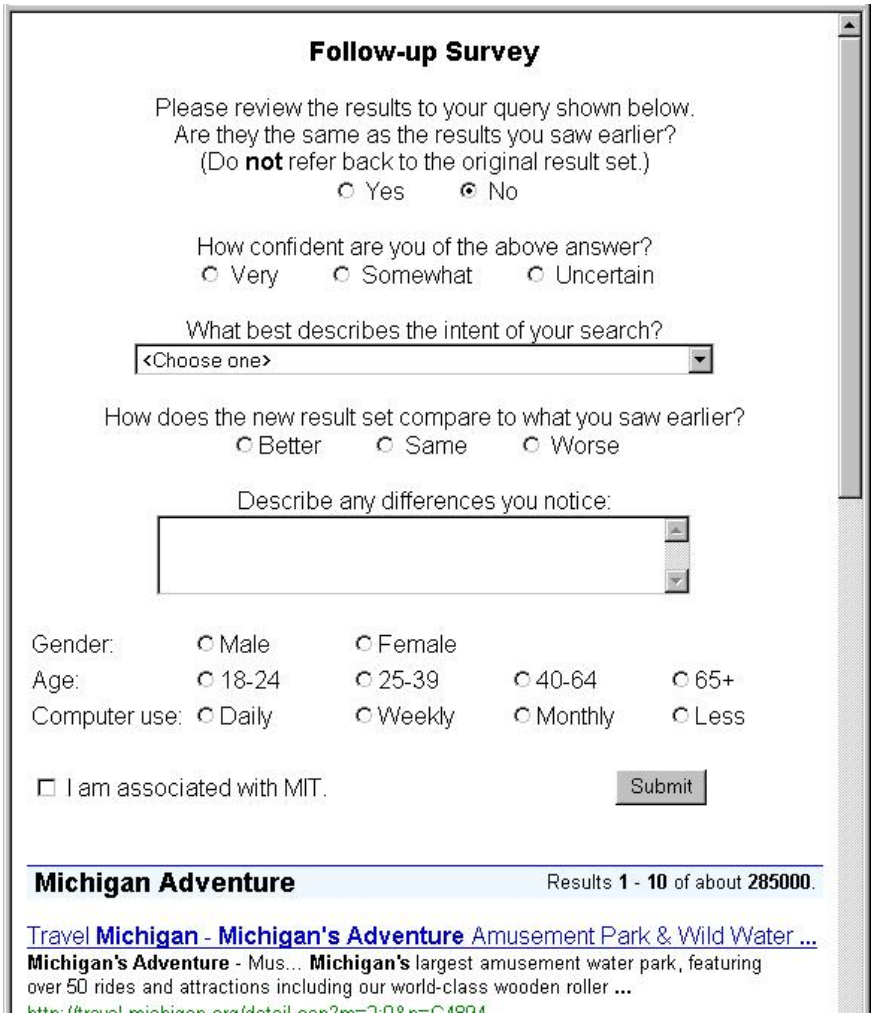


Figure 9-9. The follow-up survey asking participants whether the new search results look the same as the previously seen results for the same query

incorporated into the follow-up list were results one through ten. This design decision reflected the intended usage scenario of such a merging algorithm, where the new results to be included should ideally be better than the original results.

A total of 208 people completed the initial survey and 165 people completed the follow-up survey. The response rate was much higher for this study than for the recall study, which probably reflects the relative ease of recognizing information compared with recalling it. Because participants should not actively try to remember the initial search result list, the study was conducted with a between-subject design. Each of the five cases was completed by approximately 33 people. None of the people who participated in the initial recall study, used to develop the merge algorithm, were included in the test of the merge algorithm.

Fewer men (29%) than women (68%) participated in the study. Most participants (68%) were between the ages of 25 and 39, but 17% were over 40, and 15% under 25. Ninety-seven percent reported using a computer daily. Only 17% of respondents were affiliated

with MIT. Typically, the follow-up survey was completed within a couple of hours of the initial search. Sixty-three percent of all responses were received within three hours of the initial search, and all but ten were received within a day.

As with the previous study, the general search behavior observed of the participants was comparable to what has been reported in other larger scale studies. The average query length was 2.8 words, and the number of results clicked averaged 1.1 per query.

Results

The results of this study show that while most methods for incorporating new information into a result list create noticeably different lists, the merging of new information so as to preserve the memorable aspects of the result list goes unnoticed. This finding suggests it is indeed possible to sneak new information into a result list. While new information should not necessarily always be included in an imperceptible manner, the results imply that there is benefit to doing so; when changes to the result list were noticed, participants found the new result list quality to be lower, even though the changes were beneficiary.

Merged Results Look Unchanged

The percentage of follow-up result lists that were perceived to be the same for each of the five cases is shown in Table 9-3. Differences were noticed most often for the three cases where new information was included in the follow-up list without consideration of what the searcher found memorable. When the follow-up results list was comprised of entirely new results, participants reported the list had changed 81% of the time. When four random results were held constant (*Dumb Merging – Random*), the change to the remaining six results was noticed 62% of the time, and when the clicked results were listed first and all other results were new, the change was noticed 59% of the time. In all three of these cases, respondents were generally very confident a change had been made when they observed one. The differences between the three cases are generally not significant, although there is a weakly significance relationship between *Dumb Merging – Clicked* and *New* ($p<0.05$).

The remaining two cases (*Original* and *Intelligent Merging*), represent instances where information from the original result list that might be memorable to the participant was not permitted to change – in the former case to the point of not including any additional new information. Even when the result list did not change at all, participants sometimes believed a change had occurred (31% of the time). In fact, participants were more likely to believe the result list had changed when all results were the same than for the *Intelligent Merging* case, where differences were observed only 19% of the time. This disparity is not significant, but could possibly reflect the fact that the intelligently merged list may actually look more like the list the participant remembers than the actual original result list. While there was no significant difference between the two, the result lists from both the *Intelligent Merging* and *Original* cases were significantly more likely to be considered the same as the original list than any of the other three cases were ($p<0.01$).

The results of this study are consistent with the preliminary paper-prototype study discussed in Section 0 that found it was possible to maintain a feeling of consistency for individuals interacting with dynamic clusters of documents not by keeping the clusters static, but rather by keeping static only that information that was memorable.

Table 9-3. Results from the list recognition study. While participants noticed changes to the result list when changes were made naively, they did not when memorable information was preserved.

New Result list comprised of entirely new results		Confidence in judgment		
		Very	Some	Not
Same	19%	20%	60%	20%
Different	81%	76%	24%	0%
Dumb Merging – Random Four random results held constant		Confidence in judgment		
		Very	Some	Not
Same	38%	40%	40%	20%
Different	62%	69%	31%	0%
Dumb Merging – Clicked Clicked ranked first, then new		Confidence in judgment		
		Very	Some	Not
Same	41%	36%	50%	14%
Different	59%	75%	15%	10%
Original Results shown are the same as original result list		Confidence in judgment		
		Very	Some	Not
Same	69%	36%	41%	23%
Different	31%	30%	60%	10%
Intelligent Merging Old and new results merged intelligently		Confidence in judgment		
		Very	Some	Not
Same	81%	36%	61%	3%
Different	19%	50%	50%	0%

Changed but not Better

While it seems apparent from the study that new information can be unobtrusively merged into previously viewed search result lists, it is not obvious that people want new relevant results to look the same as old results. When a person searches, he or she is looking for relevant material, so it could be that it is best just to return relevant results regardless of past context.

To explore whether noticeable change is problematic, the perceived quality of the old and new results presented to the participants was compared. Recall that the new results incorporated into the original list were ranked higher by the underlying search engine, and thus were likely of higher quality (rank and relevance for the top twenty results are significantly correlated – this is seen both in Chapter 4 and in research by Joachims et al. (2005) and Patil et al. (2005)). The new result list was judged by an independent coder to be better than the original result list 81% of the time. Nonetheless, when the participants noticed a change, they were significantly less likely to find the changed result list to be

better than the original result list (46% of the time, $p<0.01$). Further, they thought the changed result list was worse 14% of the time.

People's judgments of relevance have been shown to be influenced by rank (Joachims et al., 2005; Patil et al., 2005), and this study confirms the findings of Chapter 7 they are likely also influenced by expectation. Thus the expectation that a person develops that certain results will appear influences what that person considers relevant. This is another indication that consistency of the type explored in this chapter is likely to be important for providing results that appear relevant.

9.3.2 Merged Results Make Finding and Re-Finding Easy

The previous study demonstrated that the Re:Search Engine can insert new information into a result list without attracting the user's notice. However, although the result lists look static and contain new information, it is not necessarily the case that users can use them to effectively find and re-find. This section presents a study that shows that the Re:Search Engine does indeed make re-finding easy while not interfering with the finding of new information. The study methodology is described, and the results presented.

Study Methodology

The study involved two parts: 1) an initial session where participants conducted initial finding tasks, and 2) a follow-up session where participants conducted finding and re-finding tasks. This section presents demographic information about the participants, and then discusses the initial session (Session 1) and the follow-up session (Session 2) in greater detail. It concludes with some details about how the data were analyzed.

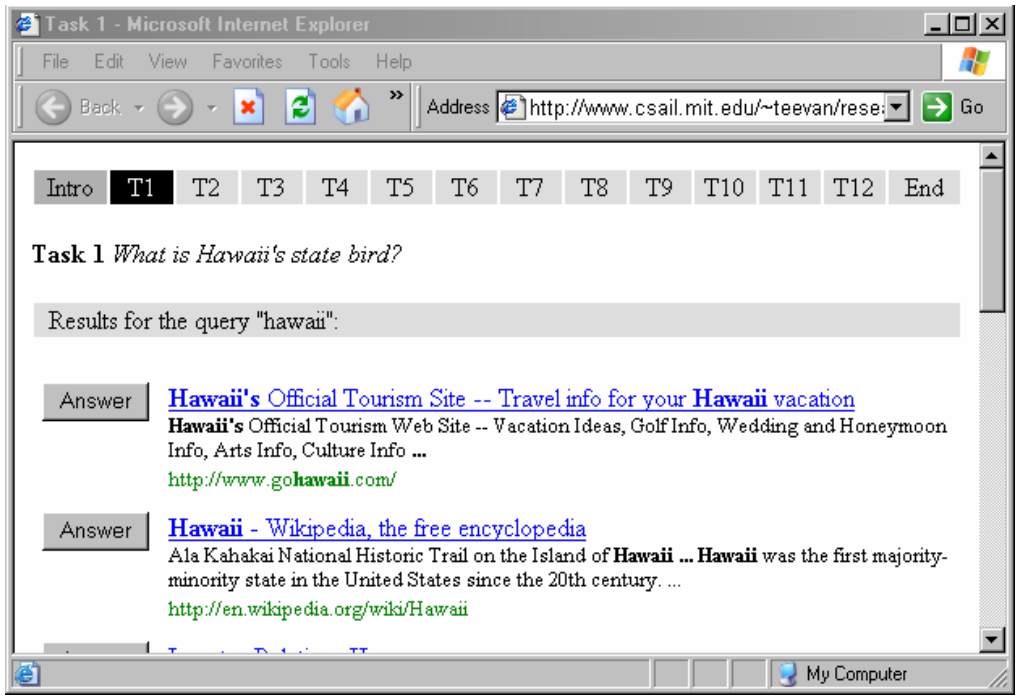


Figure 9-10. An example task used to evaluate the Re:Search Engine.

Table 9-4. Queries and their associated tasks used in the Re:Search Engine evaluation. During Session 1, all participants conducted the same task. During Session 2, they randomly either completed a re-finding task or a new-finding task.

Query	*Session 1*	*Session 2*	
	Task	**Re-Finding Task**	**New-Finding Task**
caramel apples	How much does a kit to make caramel apples cost?	Find the kit to make caramel apples that you found the cost of yesterday.	Find a caramel apple recipe that uses condensed milk.
fashion television	What is the name of the person who hosts the *Fashion Television* TV series?	What is the name of the person who hosts the *Fashion Television* TV series?	What is the theme song for the *Fashion Television* TV series?
metabolism	Find a calculator for basal metabolism, and find which activity level most closely matches yours.	Find the calculator for basal metabolism that you found yesterday where you found which activity level most closely matches yours.	Is the origin of the word "metabolism" Greek or Latin?
ram cichlid	Find a picture of a pair of Ram Chichlids guarding their eggs.	Find the picture you found yesterday of a pair of Ram Chichlids guarding their eggs.	Find a picture of a Ram Chichlid with a plain white background (no water).
stomach flu	Find a site that suggests some symptoms your child with the stomach flu might have that would warrant calling the doctor, including a swollen, hard belly and a fever higher than 102.5 degrees.	Find the site you found yesterday that suggests some symptoms your child with the stomach flu might have that would warrant calling the doctor, including a swollen, hard belly and a fever higher than 102.5 degrees.	Find a site that tells you what to expect if your child has the stomach flu and you're heading for the hospital.
video game cheats	Find a site that has links to different video game cheat sites, including specifically the Cheat Factory, AskCheats.com and Cheat Monkey.	Find the site you found yesterday that has links to different video game cheat sites, including specifically the Cheat Factory, AskCheats.com and Cheat Monkey.	Find the video game cheat site that calls itself "Multiplayer Gaming's Home page."
comcast	In what year was the Comcast Corporation founded?	In what year was the Comcast Corporation founded?	Does Comcast currently offer any full time jobs in accounting or finance?
big truck screensavers	Find a screensaver that features a semi-truck or big rig.	Find the *same* screensaver that features a semi-truck or big rig that you found yesterday.	Find a *different* screensaver featuring a semi-truck or big rig than you found yesterday.
free job posting boards	If you are interested in a position working at a hotel, can you submit your resume to a hotel-specific job Web site for free?	Find the hotel-specific job Web site you found before where you can submit a resume for free.	What company claims to be, "The world's largest FREE job and resume database?
ethyl mercaptan	Who discovered ethyl mercaptan?	Who discovered ethyl mercaptan?	How do you say "ethyl mercaptan" in French?
why become an avon rep	Does Avon provide any specialized training (e.g., selling tips or information on how you can build your sales)?	Find the site you found yesterday that tells you if Avon provides any specialized training (e.g., selling tips or information on how you can build your sales)?	How many representatives are in the Avon network worldwide?
prentice hall	Find the Prentice Hall companion Web site designed to be used with Janson's *History of Art* book, which includes features like online quizzes and writing activities.	Find again the Prentice Hall companion Web site designed to be used with Janson's *History of Art* book, which includes features like online quizzes and writing activities.	What division of Prentice Hall has the motto, "Tomorrow's solutions for today's professionals"?

Participants

A total of 42 people participated in the study. Fifty percent of the participants were male, and 48% were female (one did not report gender). A majority (69%) of the participants were between the ages of 25 and 39, but 12% were between 18 and 24, and 19% were between 40 and 64. Thirty-one percent were associated with MIT.

Session 1

During the initial session, participants conducted 12 finding tasks in random order. Tasks were inspired by queries identified in the Yahoo logs (introduced in Chapter 7) that were issued twice by the same individual and for which both new and previously viewed results were clicked during the second search session. Because the interval between the initial session and the follow-up session is one day, the 12 queries with an interval closest to a day were selected. Queries that might offend study participants, such as pornographic queries ("aunt peg") or gun-related queries ("taurus revolvers"), were ignored. Queries were approximately 2.4 words long, which is comparable to the average query length in the log data and what has been seen in other studies.

Participants were given a task and a list of results for each query. An example can be seen in Figure 9-10. The list of results was based on results 11 through 20 returned by Google. Results for the second session were based on results 1 through 10, reflecting the idea that result quality should generally improve during changes. To ensure consistency across tasks, each task was designed so that the answer could be found in one and only one search result and not in the result snippet. If more than one search result contained the answer, the superfluous results were replaced by results ranked further down. The search result that contained the answer was placed at a random location in the list. The queries and their corresponding tasks can be seen in Table 9-4.

Each task was timed. Results that were clicked were logged. A task ended when the participant marked a result as relevant or gave up. Participants were asked not to spend too much time on any one task, and encouraged to give up if they felt more than five minutes had passed. Following the task, participants were asked to report how easy the task was, how interesting it was, how relevant they found the results, and how long they thought the task took. The survey can be seen in Figure 9-11.

Session 2

A follow-up session was conducted the following day (mean=1.0 days, median=0.99 days). An interval of a day was selected because previous studies of memorability presented in this dissertation (e.g., the log analysis presented in Chapters 7 and 8, and the recall study presented in Section 9.1.2) suggest information about results lists is forgotten quickly. Task has been shown to influence list memory (Tulving & Thomson, 1973). The study tasks were not self-motivated, and re-finding was imposed artificially, so participants seemed likely to forget a considerable amount of information within a day. According to the log analysis presented in Chapter 7, repeat searches are very common at this interval, and involve repeat clicks 88% of the time and new clicks 27% of the time.

Participants were again given 12 finding tasks in random order, each associated with the same 12 queries used for the initial 12 tasks. Half of the tasks were designated re-finding tasks (the same as the task conducted the previous day), and the other half were

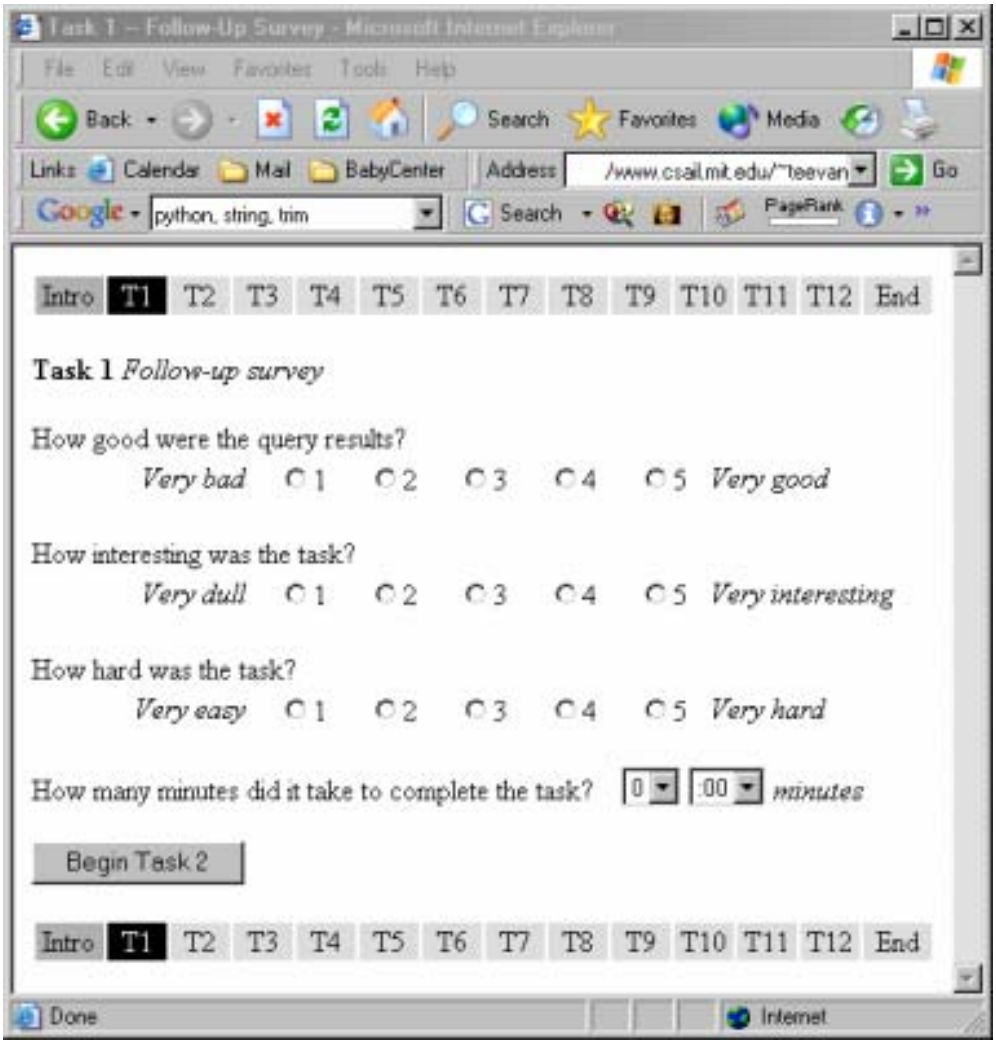


Figure 9-11. Follow-up survey for Session 1.

designated new-finding tasks (involved finding new information not previously available). The re-finding tasks and new-finding tasks for Session 2 can be seen in Table 9-4. New-finding tasks were constructed so that the answer could not be found in the initial search result list, but rather could only be found if new information is made available. The result list associated with the new task is based on results 1-10 returned by Google.

To ensure consistency across task, each task was once again designed so that the answer could be found in one and only one search result and not in the result snippet. If more than one result contained the answer, the surplus results were replaced by results ranked further down. The search result that contained the answer was placed at random in one of the top six locations in the new result list. Requiring the result to be ranked among the top six ensured that it would appear in any merged lists as well as in the new list, as only the top six new results are used for merging.

There were four types of lists the participants were given to perform each follow-up task:

Original The list is exactly the same as the original search result list the participant interacted with during the initial session. This is what the user of a system that cached previous search results would interact with. Note that this list contains the answer to any re-finding tasks, but not to tasks that involve finding new information.

New The list contains entirely new results. Because the new results are based on results 1-10 returned by Google, they are of higher quality than the original result list. Note that this list contains the answer to tasks that involve finding new information, but not to tasks that involve re-finding.

Dumb Merging The list is a random merging of the original result list with the new result list. The result from the original result list and three other random results from the original results are selected and merged into the top six results from the new list at random. This list contains the answer both to re-finding tasks and new-finding tasks. Note that if the merging had been done entirely randomly without forcing the correct result to remain in the new list, the correct result would only be preserved in 40% of the cases.

Intelligent Merging This is the list the Re:Search Engine would return for a re-finding task, with a few exceptions. So that an amount of old information consistent with the random list is preserved, exactly four results from the original result list are preserved (the exact number preserved is flexible in the Re:Search Engine, and varies between three and seven). Because this list, like the random list, should contain the answer both to the re-finding tasks and the new-finding tasks, the result with the answer to the re-finding task is preserved even if it would be dropped. This is done by bumping up its memorability score as necessary. In 77% of the cases the re-finding result was preserved naturally.

Each task was randomly assigned to be a re-finding task or a new-finding task, so that there were six re-finding tasks and six new-finding tasks. Each re-finding task was conducted with either the original list, the random list, or the Re:Search Engine list. The new list was not included because the re-finding tasks could not be solved using the new list. Each new-finding task was conducted with either the new list, the random list, or the Re:Search Engine list. The original list was not included because the new-finding tasks could not be solved using the original list.

Again, each task was timed. Results that were clicked were logged. A task ended when the participant marked a result as relevant or gave up. Following the task, participants were asked to report how easy the task was, how interesting it was, how relevant they found the results, and how long they thought the task took, as shown in Figure 9-11. Additionally, they were asked if they thought the result list was the same or different from the result list for the same query from the previous day. If they noticed a difference, they were also asked to report whether the list quality was better, worse, or the same.

Data Analysis

The data collected were analyzed to get an idea both of how well participants performed re-finding and new-finding tasks under the different list conditions, and how positively they perceived the experience. Performance was measured through analysis of the

number of clicks and the amount of time it took the participant to find the answer to the task, and the percentage of tasks that were answered correctly. Subjective measures include perceived result quality (1 to 5, from low quality to high quality) and perceived difficulty of the task (1 to 5, from very easy to very hard). Significance was calculated using standard least-squares regression analysis with fixed effects for each user.

Because participants were encouraged to give up after they felt five minutes had elapsed, task time was capped at five minutes. If a participant gave up or found an incorrect answer, their task time was recorded as five minutes. Timing information for any task that was interrupted was discarded. In the analysis of re-finding in Session 2, only those tasks for which the participant correctly found the target during Session 1 were considered – otherwise the merging of old and new information, which forced the preservation of the correct answer, did not necessarily preserve the result the participant originally found.

Results

Table 9-5 shows the average performance and subjective measures for the tasks, broken down by session and task type. On average during Session 1 participants took 120.2 seconds to complete a task. The new-finding tasks took slightly longer to complete (137.2 seconds), but the re-finding tasks were preformed in only 51.3 seconds on average. The small time discrepancy between the new-finding tasks of Session 1 and the new-finding tasks of Session 2 is likely a result of the tasks being different, as can be seen in Table 9-4. On the other hand, the Session 2 re-finding tasks correspond directly to the tasks used in Session 1 and performance for the two tasks can be directly compared.

The *p*-value reported in the right hand column of Table 9-5 shows that for all measures except result quality, performance during re-finding was significantly better than performance during the initial finding session. Clearly, the knowledge participants gained about the search tasks they performed in Session 1 helped them to re-find information more quickly than they originally found it. This section looks at what factors contributed to the ability to re-use knowledge in greater depth. It begins by showing that

Table 9-5. Basic results for study, broken down by session and task-type. The *p*-value for the difference between the tasks performed during Session 1 and later repeated during Session 2 is reported. The *p*-values that are significant at a 5% level are shown in *italics*.

Measure	Session 1		Session 2				
	All Tasks		New-Finding		Re-Finding (v. Session 1)		
	Mean	Median	Mean	Median	Mean	Median	*p*-value
Number of results clicked	2.35	1	3.50	2	1.54	1	*0.001*
Task time (seconds)	120.2	77	137.2	96	51.3	29.5	*0.001*
Percent correct	84%	100%	76%	100%	94%	100%	*0.001*
Result quality (1-5)	3.37	3	3.18	3	3.58	4	0.200
Task difficulty (1-5)	2.18	2	2.60	2	1.63	1	*0.001*

Table 9-6. Measures for new-finding and re-finding tasks, separated by whether the participant thought the result list given to them during Session 2 was the same as the result list they interacted with during Session 1 or different. The *p*-values that are significant at a 5% level are shown in *italics*.

Measure	*New-Finding*			*Re-Finding*		
	Same	Different	*p*-value	Same	Different	*p*-value
Number of results clicked	2.55	2.92	0.916	1.24	2.21	*0.001*
Task time (seconds)	148.6	120.4	0.488	39.5	94.8	*0.001*
Percent correct	74%	81%	0.382	97%	82%	*0.009*
Result quality (1-5)	3.38	3.12	0.394	3.73	3.30	*0.006*
Task difficulty (1-5)	2.55	2.46	0.617	1.50	2.21	*0.001*

task performance during re-finding in Session 2 was strongly correlated with whether the participant noticed a change to the result list, but did not affect task performance for new-finding tasks. As reported earlier in this chapter (Section 9.3.1), people are significantly less likely to notice change when the changes are incorporated through the intelligent merge process than when they are given a new list or a dumb merging, and this finding is confirmed in this study. Given people do not notice change when it happens intelligently, and given change interferes with re-finding, it is not surprising that the study shows knowledge re-use was easier for participants when using the intelligent merging than when using the dumb merging.

Noticeable Change Interferes with Re-Finding

Noticeable change appears to interfere with re-finding, but not with new-finding. Table 9-6 shows the performance for new-finding and re-finding tasks, separated by whether the participant thought the result list they used for Session 2 was the same as the result list they interacted with during Session 1 or different.

There was no significant difference between instances when a person noticed a change to the list and when they did not for any of the measures for new-finding tasks. On the other hand, performance on re-finding tasks was significantly better when the result list was believed to be the same as the original result list. People clicked fewer results (1.24 v. 2.21), took less time to complete the task (39.5 seconds v. 94.8 seconds), and answered more tasks correctly (97% v. 82%). The subjective user experience was also better when participants thought the list was the same. They generally found the result quality to be higher (3.73 v. 3.30) and the task difficulty to be lower (1.50 v. 2.21). These results suggest that change generally interferes strongly with one's ability to re-find previously viewed information, but does not greatly affect one's ability to find new information.

Merge Algorithm Affects Likelihood Change will be Noticed

While these results are interesting, it is possible that people were more likely to notice change for tasks they found difficult. However there was a strong correlation between the list merge type and task difficulty. This study, like the study described in Section 9.3.1, revealed that people were unlikely to notice changes to the intelligently merged

Table 9-7. The percentage of time participants thought results were the same as a function of task and list type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	List Type Used During Session 2	Results Perceived to be the Same	*p*-value (significance)		
			Dumb	Intelligent	New/Original
New-Finding	Dumb merging	50%		0.062	*0.006*
	Intelligent merging	61%	0.062		*0.001*
	New result list	38%	*0.006*	*0.001*	
Re-finding	Dumb merging	60%		*0.008*	*0.006*
	Intelligent merging	76%	*0.008*		0.920
	Original result list	76%	*0.006*	0.920	

search result list. The probability that a change was noticed for each of the experimental conditions is shown in Table 9-7. For both the finding tasks and the re-finding tasks, participants were less likely to notice changes when the results were intelligently merged, compared with the dumb merging and with an entirely different result set. Although the intelligent merging contained six new results, people thought it was the same as the original list just as often as they thought the original list was the same.

Intelligent Merging Works Well for Finding and Re-Finding

Given that participants did not notice change when it happened intelligently, and given change interfered with re-finding, it is not surprising that the study shows knowledge re-use was easier for participants when using the intelligent merging than when using the dumb merging. In many cases, in fact, performance with the intelligently merging was comparable to the ideal scenario of either an entirely static result list for re-finding, or a list with new information for finding.

Tables 9-8, 9-9, and 9-10 show the results for how participants performed on new-finding and re-finding tasks, broken down by list type. While many measures exhibit a trend, most are not significant. However, the amount of time taken to complete a re-finding task was significantly the lowest when a static result list was used, next best when the results were merged intelligently, and significantly the worst when they were merged dumbly.

Participants were more likely to correctly answer the re-finding task using an unchanged list, when compared to the dumb merging. Participants were also more likely to correctly answer the task for re-finding tasks, but the difference is not significant. However, it is worth noting that if the correct result were not preserved during either the intelligent merging or the dumb merging, it would be impossible for the participant to re-find the information regardless. Since the correct result was preserved significantly more often for the intelligent merging than the dumb merging (78% of the time v. 40% of the time), it is likely that had the target not been required to remain in the list a more striking difference would have been seen.

In general, the difference between performance measures for each list type for new-finding was not significant. However, the intelligently merged lists do appear to happen significantly faster than with a dumb merging (Table 9-8). This may be because there is

Table 9-8. The time it took participants the Session 2 task as a function of task and list type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	List Type Used During Session 2	Task Time (seconds)		*p*-value (significance)		
		Mean	Median	Dumb	Intelligent	New/Original
New-Finding	Dumb merging	153.8	115.5		*0.037*	0.267
	Intelligent merging	120.5	85.5	*0.037*		0.280
	New result list	139.3	92	0.267	0.280	
Re-finding	Dumb merging	70.9	37.5		*0.037*	*0.008*
	Intelligent merging	45.6	23	*0.037*		0.554
	Original result list	38.7	26	*0.008*	0.554	

Table 9-9. The number of results that participants clicked during Session 2 task as a function of task and list type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	List Type Used During Session 2	Num. of Results Clicked		*p*-value (significance)		
		Mean	Median	Dumb	Intelligent	New/Original
New-Finding	Dumb merging	4.13	2.5		0.146	0.085
	Intelligent merging	3.34	2	0.146		0.808
	New result list	3.15	2	0.085	0.808	
Re-finding	Dumb merging	1.66	1		0.804	0.107
	Intelligent merging	1.72	1	0.804		0.154
	Original result list	1.26	1	0.107	0.154	

Table 9-10. The percentage of tasks participants answered correctly during Session 2 as a function of task and list type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	List Type Used During Session 2	% Correct		*p*-value (significance)		
		Mean	Median	Dumb	Intelligent	New/Original
New-Finding	Dumb merging	74%	100%		0.185	0.960
	Intelligent merging	83%	100%	0.185		0.144
	New result list	73%	100%	0.960	0.144	
Re-finding	Dumb merging	88%	100%		0.061	*0.015*
	Intelligent merging	96%	100%	0.061		0.543
	Original result list	99%	100%	*0.015*	0.543	

some knowledge re-use even when re-finding. In those cases, the participant has learned which results do not contain the answer, and knows to avoid them, while with the dumb merging they may find it necessary to repeat their review of information they've seen before. It is worth noting that the rank of the correct result for new-finding tasks was significantly ($p<<0.001$) lower when the results were intelligently merged than for either

Table 9-11. The quality of the results, judged by participants for tasks conducted during Session 2, as a function of task and list type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	List Type Used During Session 2	Result Quality (1-5)		*p*-value (significance)		
		Mean	Median	Dumb	Intelligent	New/Original
New-Finding	Dumb merging	2.94	3		0.330	*0.034*
	Intelligent merging	3.19	3	0.330		0.247
	New result list	3.38	4	*0.034*	0.247	
Re-finding	Dumb merging	3.42	3		0.054	0.364
	Intelligent merging	3.70	4	0.054		0.288
	Original result list	3.61	4	0.364	0.288	

Table 9-12. The difficulty of the task, judged by participants for Session 2 task as a function of task and list type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	List Type Used During Session 2	Task Difficulty (1-5)		*p*-value (significance)		
		Mean	Median	Dumb	Intelligent	New/Original
New-Finding	Dumb merging	2.72	2		0.519	0.275
	Intelligent merging	2.61	2	0.519		0.650
	New result list	2.51	2	0.275	0.650	
Re-finding	Dumb merging	1.79	2		*0.023*	0.061
	Intelligent merging	1.53	1	*0.023*		0.669
	Original result list	1.57	1	0.061	0.669	

the new list or the dumb merging – appearing on average seventh in the list as opposed to 5.6th (dumb merging) or 3.6th (new list). The reason for this is that, as mentioned earlier, the correct result was always placed in the top six results in the new list. When merging the new results with the old list according to the dumb merging algorithm, on average two of the four results would be merged in ahead of the correct result. On the other hand, the intelligent merging is likely to preserve the first couple of results since they are the most memorable, and thus merge more results ahead of the correct result. Nonetheless, despite the lower rank of the correct result, participants were still able to find the results faster.

Table 9-11 and Table 9-12 show the subjective performance with each list type. Re-finding with the intelligent merging was considered significantly ($p<0.05$) easier compared to re-finding with the dumb merging. Result quality was considered significantly better for the new list for new-finding tasks than for the dumb merging. This may be because the correct result was ranked higher for the new list for new-finding tasks, but is unlikely since the correct result was ranked much lower for the intelligently merged list.

In general, for re-finding tasks, task performance with the original result list appears to be best, followed by performance with the intelligently merged list, and then the dumb

merging. Undoubtedly, had the case using a new result list been tested, task performance would have been the worst, given the solution to the task could not be found in the result list. For new-finding tasks, performance was generally best with the new result list, followed by the intelligent merging, followed by the dumb merging. Again, had the original result list been tested for the finding task, performance would have almost certainly been the worst, since the solution was not present.

Given these findings, the intelligent merging used by the Re:Search Engine seems to be the best compromise to support both finding and re-finding. A static, unchanging result list works well to support re-finding, but does not support the finding of new information. In contrast, a result list with new information works well to support the finding of new information, but does not support re-finding well. The intelligent merging performs closely to the best of both in both cases, while the dumb merging does comparatively worse.

History is the witness that testifies to the passing of time; it illumines reality, vitalizes memory, provides guidance in daily life and brings us tidings of antiquity.

- Cicero (106 BCE - 43 BCE)

Chapter 10

Conclusion and Future Work

This dissertation has explored Web search personalization as a way to help people find what they are looking for faster and more easily. Successfully supporting finding and re-finding is challenging, but offers many rewards. Part I focused on finding. Chapter 3 studied how people search, and highlighted the fact that people do not like to exert the effort to detail what they are looking for during their searches. Chapter 4 focused on how individuals search, and showed that personalization is important if search systems hope to meet the individual search goals of different people issuing the same query. Chapter 5 demonstrated that it is possible to significantly improve result quality via implicit personalization by using a rich user profile.

Part II dove into re-finding. Chapter 7 revealed re-finding is prevalent, and discussed the behavior's characteristics. The importance of consistency in re-finding was explored in Chapter 8, and changes to search result lists were shown exacerbate problems re-finding. Support for finding through personalization, as presented in Part I, appeared at odds with support for re-finding, because personalization increases the rate of change to result lists. Chapter 9 presented a solution to this apparent conflict. It introduced the Re:Search Engine, a system that personalizes search results not by ranking the most relevant results first, but rather by ranking them where the user expects them.

In this chapter the contributions of this thesis are revisited, and the future directions inspired by them are highlighted. Ultimately, this research should help make it possible for people to spend less time searching for information, and more time making productive use of it to get things done.

10.1 Contributions

As discussed in Chapter 1, the thesis makes five important contributions to the areas of information retrieval and human computer interaction. These are reviewed here:

- First, this thesis showed that for directed search people preferred to orienteer over teleport. It also gave insight into the differences in search behavior between individuals, suggesting that people use different step sizes while orienteering and

that people have very different notions of what is relevant to even fairly unambiguous queries.

- Second, this thesis presented a novel application of relevance feedback to the realm of implicit search result personalization. It demonstrated that implicit relevance feedback based on large amounts of information about the searcher can significantly improve search result quality.
- Third, this thesis offered evidence that repeat Web search queries are extremely prevalent. It highlighted common features of repeat searches and showed that people often search for new information using the same query results they use to find previously viewed information.
- Fourth, this thesis presented a model of what people remember about search results based on a large scale study of human memory. The study showed that what people remember about a result is a function of their interaction with the item and the item's location in the list.
- Fifth, this thesis presented an algorithm that uses the model of what people remember about search result lists to invisibly merge new information into a previously viewed search result list. Doing this allows people to effectively find both new and old information using the same search result list.

10.2 Future Work

There are a number of interesting directions for future work suggested by the research presented here. This section relates what has been shown about people's search behavior to the design of future tools. It highlights several interesting ways that appear particularly promising to support finding and re-finding.

10.2.1 Better Support for Finding

This section explores two approaches to supporting finding that are worth pursuing: 1) better support for the common search behavior of orienteering, as discussed in Chapter 3, and 2) better support for personalization, as motivated by Chapter 4 and explored in Chapter 5.

Better Support for Orienteering

Chapter 3 presented several advantages to orienteering, including that it appeared to lessen the cognitive burden of finding information, help people better understand their answer, and give people a sense of location during their search. These advantages provide insights for the construction of future search tools that go beyond merely providing keyword search.

To lessen the cognitive burden of search, people used a considerable amount of meta-information during their search that was not available for use by keyword search engines. While search engines are expanding to include meta-data, specifying an information need up front was sometimes more difficult than orienteering to information, and even, in

some cases, impossible. A better way of incorporating meta-data is to use meta-data for browsing, as it is in the systems being developed by Yee et al. (2003), Dumais et al. (2003), and Cutrell et al. (2006).

People were observed to often look for a particular information source in order to help their information target. Searching for the source instead of directly for the target lessened the cognitive burden of search because people often remembered more about the source than about the information target itself. Thus, it is particularly important to support the use of meta-data for sources of information, such as Web homepages or email messages. While the personalized search system presented in Chapter 5 takes advantage of this to boost results from previously viewed domains, next generation search tools could take this further and learn users' habitually used or trusted sources and make them easily accessible, similar to Maglio and Barrett (1997). Tools could also help people identify the correct source for a given target by previewing the content contained in the source – for example, by flagging email messages that contain email addresses, times, dates or locations.

Orienteering helped participants understand and trust the answers they found. Search tools could enable a similar understanding by showing the context of any results provided (e.g., the source as discussed above, or, in the case of question answering, the context of the answer, as done by Lin et al., 2003). Further, search tools could direct search or navigation to sources trusted by the user. To help users understand and accept negative results, search tools could also allow the user to take part in the search process, for example by helping people exhaustively search small areas such as Web pages or individual files.

Orienteering also helped people maintain a sense of location during their search. One technique people used to maintain this sense of location was URL manipulation, which could be better supported by future search tools. In addition, people sometimes knew their target but not the path to that target. To keep users from having to make a blind jump to the target, a next generation search tool could return several paths to potential targets to help the user navigate incrementally. To maintain a sense of location, people often used keyword in small steps: e.g., first Google, then site search and then page search. A search tool could integrate all three of these variable sized searches into one tool, to keep people from having to think about different tools and different interfaces for each step in their search.

Another way a next-generation search tool could support stepping behavior would be to automatically refine people's information as they interact with it by, for example, clustering it or suggesting query refinements. Given such a system, the comparison between filers and pilers provides insight into the type of personalization that should be supported. Because certain individuals tended to use search engines to take larger steps toward their information needs while others took smaller steps, the size of the refinements could vary according to how comfortable the user is taking large steps to locate information. Large, disjoint refinements would be appropriate for a user that prefers using keyword search to take large steps toward their information, while smaller, similar refinements would be more appropriate for finer-grained navigation.

Better Support for Personalization

Chapter 5 looked at one way to support personalization based on the findings of Chapters 3 and 4. The parameters of the personalization procedure explored represent only a small subset of the space of parameterizations. As an example of an extension, the user profile could incorporate a more complex notion of time and current interest by being more influenced by documents seen recently than documents seen a long time ago. Within the relevance feedback framework, it would be interesting to explore tuning the algorithm's parameters, using more complex terms (e.g., phrases), and incorporating length normalization. It might also be worthwhile to examine alternative frameworks for incorporating differential term weighting to personalize search.

In the personalization experiments presented, no one parameter setting consistently returned better results than the original Web ranking, but there was always some parameter setting that led to improvements. This result highlights the opportunity for using machine learning to select the best parameters based on the query at hand. This selection could be done based on the individual (e.g., the user's interests change often, so recently seen documents might be weighted more heavily in the construction of the user profile), the query (e.g., the query term is very common in the user's personal index, suggesting that a great deal of personalization is needed), and properties of the result set (e.g., the documents in the result set have widely varying scores, suggesting that personalization would be useful).

Differences across queries are particularly interesting to explore. There was some indication that personalization was more effective for shorter queries and more ambiguous queries (measured by the number of Web documents matching the query). For example, the queries "discount hotel London", "activities Seattle" and "trebuchet" improved with personalization but the queries "habanero chiles", "Snoqualmie ridge QFC" and "Pandora ranking 60 level-1 level-2" did not. However the effects were quite variable and were not statistically reliable given the relatively small numbers of queries used in the experiment.

In another direction, it would almost certainly be useful to introduce additional classes of text- and non-text based content and activities in the construction of interest profiles. These profiles could incorporate information about activities such as current or recent software application usage, the pattern of topics of content at the current or recent focus of attention, and the history of locations visited by the user as sensed via GPS and other location-tracking methodologies.

10.2.2 Better Support for Re-Finding

This section explores several approaches that are likely to lead to better re-finding support. It begins with a discussion of how better organizational support can help people re-find, and then discusses how search history tools can be improved. This thesis focuses on how change affects re-finding, and the section concludes with a discussion of how change can be better supported so as to not interfere with re-finding.

Better Support for Organization

Re-finding is a complementary action to keeping. When people encounter valuable information, they decide how to keep it based in part on their expected future information needs. There is often a tradeoff between investing more time during the initial encounter to keep the information or more time later to re-find it. For example, when Connie, in Chapter 1, located the About.com Web site she liked, she could have invested the time to bookmark it. Instead, she chose to invest more time later to re-find the site when and if it became necessary. Successful re-finding support could affect people's organizational behavior.

One value that information organization has is that it can remind the user where information is stored. Going forward it will be interesting to consider how search tools can provide reminding functionality without organizational overhead. As such tools are deployed and used, researchers will need to explore how they affect information seeking behavior. Re-finding tools may do well to integrate the organization of information (or creation of context) with search. For example, systems like Google's Gmail and Phalt (Cutrell et al., 2006) do not include folders per se, but rather folders are simply queries that filter a user's information based on metadata such as the email's sender field.

Better Support for History Management

There are a number of existing search history tools intended to support re-finding that do not intend to also support the finding of new information (e.g., Raghavan & Sever, 1995; Bharat, 2000; Komlodi, 2004; Komlodi, Soergel, & Marchionini, 2006). Most current search history designs retain a time ordered query history for their users, recording queries and clicked URLs. The research in this thesis suggests several implications for such tools in addition to the approach explored in Chapter 9 of maintaining context.

Analysis of the Yahoo logs presented in Chapter 7 indicated that different users make use of repeat queries differently. By understanding the behavior of an individual user, a search history list could be customized to contain those queries that are important to a given user. For example, because people repeated queries at different rates, the size of the list of past queries could vary by user. Because there was some periodicity in repeat clicks, search histories could also be customized based on the time of day. Users with a large number of navigational queries may also benefit from a direct link or preview of the query's associated Web page (possibly labeled with the frequent query term).

Better Support for Change

Looking forward, effective management of changing information will be essential to successfully supporting re-finding. The growing ease of electronic communication and collaboration, the rising availability of time dependent information, and the introduction of automated agents, suggest information is becoming ever more dynamic. Even traditionally static information like a directory listing on a personal computer has begun to become dynamic; Apple, for example, has introduced "smart folders" that base their content on queries and change as new information becomes available. To make it possible for people to have access to new and better information without becoming distracted or disoriented, the following three areas are worth pursuing:

Improve and Generalize the Model of what is Memorable

This thesis has shown that simple measures like click-through and rank can be effective at identifying memorable search results. However, because the memorable aspects of a result list are likely to generalize to other lists of information (e.g., email inboxes and directory listings), it is worthwhile to improve upon the existing model. There are a number of complex features worth exploring, including dwell time, scrolling behavior, and eye tracking data. While each additional feature may provide little new information alone, it may be possible to learn a good combination of features.

To truly generalize the model beyond lists of information, it will be important to look at an even broader range of features. Francisco-Revilla et al. (2001b) conducted a study investigating the types of changes to Web pages people perceived to be important, that suggests features like color and structure are worth exploring. The research by Rosenholtz et al. (2005) may provide a good general model. The authors use a model of visual clutter to place new visual information in clear areas so that the addition is obvious, but the same model could be used to sneak new information into cluttered areas.

Effectively Use the Model

Just as a model of clutter can be used to hide or highlight new information, a high quality model of what is memorable can be used not just to hide change, but also to highlight it. If it is possible to identify whether a person is interested in new information or in re-finding old information, change can be presented in a task-appropriate way. Identifying repeat Web searches can be challenging, because people often do not re-issue past queries exactly, but outside of the realm of search results, re-finding behavior can be even more difficult to recognize. Perhaps as a person starts to follow a known path, what they are looking for could be predicted, and the path to it preserved and highlighted.

It will also be interesting to explore whether the approach described in this thesis can generalize to real time information interactions. For example, White, Ruthven, and Jose (2002) observed people had trouble interacting with a result list that changed as they used it due to implicit relevance feedback, but proper consideration of the individual's expectations may alleviate the problem. Change blindness literature suggests that it might be possible to provide new information to the user as they interact with it by introducing a distraction.

Present Change at the Right Time

This thesis has focused specifically on *what* new information to display and *how* to display it. However, understanding *when* to display new information (either so it is noticeable to the user or not) is crucial. Combining this research models of attention (e.g., Horvitz et al., 2003; McCrickard & Chewar, 2003; Ho & Intille, 2005) will allow for a more nuanced understanding of how to perform interruptions, and how to avoid them when not absolutely necessary.

Hopefully the research presented in this thesis will lead to real improvements to the real world systems that people use to find information, old and new.

If we knew what it was we were doing, it would not be called research, would it?
- Albert Einstein (1879 - 1955)

Bibliography

Ackerman, M. S. and Halverson, C. (1998). Considering an organization's memory. In *Proceedings of CSCW '98*, 39-48.

Adler, L. M. (1957). A modification of Kendall's tau for the case of arbitrary ties in both rankings. *Journal of the American Statistical Society*, 52: 33-35.

Ahlström, D. (2005). Modeling and improving selection in cascading pull-down menus using Fitts' law, the steering law and force fields. In *Proceedings of CHI '05*, 61-70.

Allan, J. (2006). HARD Track overview in TREC 2005 high accuracy retrieval from documents. In *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*.

Alvarado, C., Teevan, J., Ackerman, M. S., and Karger, D. R. (2003). Surviving the information explosion: How people find their electronic information. Technical Report AIM-2003-006, MIT AI Lab.

Anick, P. (2004). Using terminological feedback for Web search refinement: A log-based study. In *Proceedings of WWW '04*, 88-95.

Asch, S.E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41: 1230-1240.

Aula, A., Jhaveri, N., and Käki, M. (2005). Information search and re-access strategies of experienced Web users. In *Proceedings of WWW '05*, 583-592.

Bates, M. (1979). Information search tactics. *Journal of the American Society for Information Science*, 30: 205-214.

Bates, M. (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., and Frieder, O. (2004). Hourly analysis of a very large topically categorized Web query log. In *Proceedings of SIGIR '04*, 321-328.

Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In *Proceedings of Information Retrieval '93*, 55-66.

Belkin, N. J., Marchetti, P. G., and Cool, C. (1993). Braque: Design of an interface to support user interaction in information retrieval. *Information Processing and Management*, 29(3): 325-344.

Bernard, H. R. (1994). *Research Methods in Anthropology: Qualitative and Quantitative Approaches.* Landham, MD: Altamira Press.

Bharat, K. (2000). SearchPad: Explicit capture of search context to support Web search. In *Proceedings of WWW '00*, 493-501.

Bhavnani, S. K., and Bates, M. J. (2002). Separating the knowledge layers: Cognitive analysis of search knowledge through hierarchical goal decompositions. In *Proceedings of the American Society for Information Science and Technology*, 204-213.

Boardman, R. and Sasse, M. A. (2004). "Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 583-590.

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

Bruce, H., Jones, W., and Dumais, S. (2004). Keeping and re-finding information on the Web: What do people do and what do they need? In *Proceedings of ASIST '04*.

Budzik, J. and Hammond, K. (1999). Watson: Anticipating and contextualizing information needs. In Proceedings of ASISIT '99, 727-740.

Byrne, M. D., John, B. E., Wehrle, N. S., and Crow, D. C. (1999). The tangled Web we wove: A taskonomy of WWW use. In *Proceedings of CHI '99*, 1999, 544-551.

Capra, R. G. and Pérez-Quiñones, M. A. (2003). Re-finding found things: An exploratory study of how users re-find information. Technical Report, Virginia Tech.

Capra, R. G. and Pérez-Quiñones, M. A. (2005). Mobile refinding of Web information using a voice interface: An exploratory study. In *Proceedings of the 2nd Latin American Conference on Human-Computer Interaction (CLIHC 2005)*.

Capra, R. G. and Pérez-Quiñones, M. A. (2005). Using Web search engines to find and refind information. *IEEE Computer,* 38 (10), 2005, 36-42.

Cater, K., Chalmers, A., and Dalton, C. (2003). Varying rendering fidelity by exploiting human change blindness. In *Proceedings of the 1st international conference on Computer graphics and interactive techniques in Australasia and South East Asia*.

Claypool, M., Le, P., Wased, M., and Brown, D. (2001). Implicit interest indicators. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*, 33-40.

Cockburn, A., Greenberg, S., Jones, S., Mckenzie, B., and Moyle, M. (2003). Improving Web page revisitation: Analysis, design and evaluation. *IT and Society*, 1(3): 159-183.

Cutrell, E., Robbins, D., Dumais, S., and Sarin, R. (2006). Fast, flexible filtering with Phlat. In *Proceedings of CHI '06*, 261-270.

Czerwinski, M., Horvitz, E., and Cutrell, E. (2001). Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI 2001 Conference*, 167-170.

Czerwinski, M., Horvitz, E., and Wilhite, S. (2004). A diary study of task switching and interruptions. In *Proceedings of CHI '04*, 175-82.

Dalal, Z., Dash, S., Dave, P., Francisco-Revilla, L., Furuta, R., Karadkar, U., and Shipman, F. (2004). Managing distributed collections: Evaluating Web page changes, movement, and replacement. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital libraries*, 160-168.

Davis, P., Maslov, A., and Phillips, S. (2005). Analyzing history in hypermedia collections. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, 171-173.

Ducheneaut, N. and Bellotti, V. (2001). E-mail as habitat: An exploration of embedded personal information management. *interactions*, 8(5): 30-38.

Dumais, S. T., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., and Robbins, D. (2003). Stuff I've Seen: A system for personal information retrieval and re-use. In *Proceedings of SIGIR '03*, 72-79.

Dumais, S., Cutrell, E., and Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of CHI '01*, 277-284.

Dumais, S., Cutrell, E., Sarin, R., and Horvitz, E. (2004). Implicit Queries (IQ) for contextualized search. In *Proceedings of SIGIR '04*, 594.

Durlach, P. J. (2004). Change blindness and its implications for complex monitoring and control systems design and operator training. *Human-Computer Interaction,* 19(4): 423-451.

Eastman, C. M. and Jansen, B. J. (2003). Coverage, relevance and ranking: The impact of query operators on Web search engine results. *TOIS*, 21(4): 383-411.

Feldman, S. (2004). The high cost of not finding information. *KM World*. Retrieved January, 2006 from http://www.kmworld.com/Articles/ReadArticle.aspx?ArticleID=9534.

Fertig, S., Freeman, E., and Gelernter, D. (1996). Lifestreams: An alternative to the desktop metaphor. In *Conference Companion on Human Factors in Computing Systems: Common Ground*, 410-411.

Fetterly, D., Manasse, M., Najork, M., and Wiener, J. (2003). A large-scale study of the evolution of Web pages. In *Proceedings of the 12th International World Wide Web Conference*, 669-678.

Ford, N., Wilson, T. D., Foster, A., Ellis, D., and Spink, A. (2002). Information seeking and mediated searching. Part 4. Cognitive styles in information seeking. *Journal of the American Society for Information Science and Technology*, 53(9): 728-735.

Francisco-Revilla, L., Shipman, F. M., Furuta, R., Karadkar, U., and Arora, A. (2001). Managing change on the Web. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 67-76.

Francisco-Revilla, L., Shipman, F. M., Furuta, R., Karadkar, U., and Arora, A. (2001). Perception of content, structure, and presentation changes in Web-based

hypertext. In *Proceedings of the twelfth ACM conference on Hypertext and Hypermedia*, 205-214.

Freeman, E. and Fertig, S. (1995). Lifestreams: Organizing your electronic life. In AAAI Fall Symposium: AI Applications in Knowledge Navigation and Retrieval.

Friedman, B., Kahn, P. H., and Hagman, J. (2003). Hardware companions? - What online AIBO discussion forums reveal about the human-robotic relationship. In *Proceedings of CHI '03*, 273-280.

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

Gauch, S., Chafee, J., and Pretschner, A. (2004). Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems*, 1(3-4): 219-234.

Goldberg, A.V. An efficient implementation of a scaling minimum-cost flow algorithm. *Journal of Algorithms*, 22(1): 1-29.

Good, N. S. and Krekelberg, A. (2003). Usability and privacy: A study of KaZaA P2P file-sharing. In *Proceedings of CHI '03*, 137-144.

Graphic, Visualization, and Usability Center (1998). GVU's Tenth WWW User Survey. Retrieved September, 2004 from http://www.gvu.gatech.edu/user_surveys/survey-1998-10.

Harada, S., Naaman, M., Song, Y. J., Wang, Q., and Paepcke, A. (2004). Lost in memories: interacting with photo collections on PDAs. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, 325-33.

Hawking, D. (2000). Overview of the TREC-9 Web track. In *Proceedings of TREC '00*, 87-102

Hawking, D. and Craswell, N. (2001). Overview of the TREC-2001 Web Track. In *Proceedings of TREC '01*, 61-68.

Hayashi, K., Nomura, T., Hazama, T., Takeoka, M., Hashimoto, S., and Gudmundson, S. (1998). Temporally-threaded workspace: A model for providing activity-based perspectives on document spaces. In *Proceeding of HyperText '98*.

Hearst, M. (1996). Research in support of digital libraries at Xerox PARC, Part I: The changing social roles of documents. *D-Lib Magazine*, May 1996.

Henson, R. (1998). Short-term memory for serial order: The Start-End Model. *Cognitive Psychology*, 36: 73-137.

Herder, H. (2006). Forward, back, and home again: Analyzing user behavior on the Web. Ph.D. Thesis, University of Twente.

Hicks, D. L., Leggett, J. J., Nürnberg, P. J., and Schnase, J. L. (1998). A hypermedia version control framework. *ACM Transactions on Information Systems (TOIS)*, 16(2): 127-160.

Ho, J. and Intille, S. (2005). Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of CHI '05*, 909-918.

Hölscher, C. and Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Networks*, 33: 337-346.

Horvitz, E., Kadie, C. M., Paek, T., and Hovel D. (2003). Models of attention in computing and communications: From principles to applications. Communications of the ACM, 46(3): 52-59.

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin and Review*, 2: 105-112.

Ingham, D., Caughey, S., and Little, M. (1996). Fixing the "broken-link" problem: The W3Objects approach. *Computer Networks and ISDN Systems*, 28(7-11): 1255-1268.

Intille, S. (2002). Change blind information display for ubiquitous computing environments. In *Proceedings of the Fourth International Conference on Ubiquitous Computing,* 91-106.

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

Järvelin, K. and Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of SIGIR '00*, 41-48.

Jeh, G. and Widom, J. (2003). Scaling personalized Web search. In *Proceedings of WWW '03*, 271-279.

Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf and C. Burges and A. Smola (Eds.), *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA: MIT Press.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR '05*, 154-161.

Jones, W. (in press). *Keeping Found Things Found: The Study and Practice of Personal Information Management*. San Francisco, CA: Morgan Kaufmann.

Jones, W. and Teevan, J. (in press), Eds. *Readings in Personal Information Management*. Seattle, WA: University of Washington Press.

Jones, R. and Fain, D. C. (2003). Query word deletion prediction. In *Proceedings of SIGIR '03*, 435-436.

Jones, W., Bruce, H., and Dumais, S. T. (2001). Keeping founds things found on the web. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 119-126.

Jones, W., Bruce, H., and Dumais, S. T. (2003). How do people get back to information on the Web? How can they do it better? In *Proceedings of Interact '03*, 793-796.

Jones, W., Dumais, S. T., and Bruce, H. (2002). Once found, what then?: A study of "keeping" behaviors in the personal use of Web information. In *Proceedings of ASIST '02*, 391-402.

Jul, S. and Furnas, G. W. (1997). Navigation in electronic worlds: Workshop report. *SIGCHI Bulletin*, 29(2): 44-49.

Kaasten, S., Greenberg, S., and Edwards, C. (2002). How people recognize previously seen WWW pages from titles, URLs and thumbnails. In *Proceedings of HCI '02*, 247-265.

Kahle, B. (1997). Preserving the Internet. *Scientific American,* 276(3):82-83.

Kamvar, M. and Baluja, S. (2006). A large scale study of wireless search behavior: Google mobile search. In *Proceedings of CHI '06*, 701-709.

Keenoy, K. and Levene, M. (2005). Personalisation of Web search. *ITWP*, 201-228.

Kelly, D. and Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In *Proceedings of SIGIR '04*, 377-84.

Kelly, D. and Cool, C. (2002). The effects of topic familiarity on information search behavior. In *Proceedings of the Second ACM/IEEE Joint Conference on Digital Libraries (JCDL '02),* 74-75.

Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2): 18-28.

Kelly, D. and Teevan, J. (in press). Understanding what works: Evaluating PIM tools. In W. Jones and J. Teevan (Eds.), *Readings in Personal Information Management*. Seattle, WA: University of Washington Press.

Kim, K. S. and Allen, B. (2002). Cognitive and task influences on Web searching. *Journal of the American Society for Information Science and Technology*, 52(2): 109-119.

Klimt, B. and Yang, Y. (2004). The Enron Corpus: A new dataset for email classification research. In *Proceedings of the European Conference on Machine Learning*, 217-226.

Koehler, W. (2002). Web page change and persistence: A four year longitudinal study. *Journal of the American Society for Information System Technology*, 52(2): 162-171.

Koenmann, J. and Belkin, N. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of CHI '96*, 205-212.

Komlodi, A. (2004). Task management support in information seeking: A case for search histories. *Computers in Human Behavior*, (2): 163-184.

Komlodi, A., Soergel, D., and Marhionini, G. (2006). Search histories for user support in user interfaces. *Journal of the American Society for Information Science and Technology*, 57(6): 803-807.

Kritikopoulos, A. and Sideri, M. (2003). The Compass Filter: Search engine result personalization using Web communities. In *Proceedings of ITWP '03*.

Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5): 361-371.

Lansdale, M. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1): 458-465.

Lansdale, M., and Edmonds, E. (1992). Using memory for events in the design of personal filing systems. *International Journal of Man-Machine Studies*, 36: 97-126.

Lau, T. and Horvitz, E. (1999). Patterns of search: Analyzing and modeling Web query refinement. In *Proceedings of the Seventh International Conference on User Modeling*, 119-128.

Lawrence, S. and Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280: 98-100.

Levy, D. (1994). Fixed or fluid? Document stability and new media. In *Proceedings of European Conference on Hypertext*.

Lieberman, H., Nardi, B., and Wright, D. J. (2001). Training agents to recognize text by example. *Autonomous Agents and Multi-Agent Systems*, 4(1-2): 79-92.

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. R. (2003). The role of context in question answering systems. In *Proceedings of CHI '03*, 1006-1007.

Liu, F., Yu, C., and Meng, W. (2002). Personalized Web search by mapping user queries to categories. In *Proceedings of CIKM '02*, 558-565.

Lyman, P. (2003). Archiving the World Wide Web. *LOOP: AIGA Journal of Interaction Design Education*, 7.

Maglio, P. P. and Barrett, R. (1997). How to build modeling agents to support Web searchers. In *Proceedings of UM '97*, 5-16.

Malone, T. E. (1983). How do people organize their desks? *ACM Transactions on Office Information Systems*, 1(1): 99-112.

Marchionini, G. (1995). Information Seeking in Electronic Environments. New York: Cambridge University Press.

Marshall, C. C. and Bly, S. (2005). Saving and using encountered information: Implications for electronic periodicals. In *Proceedings of CHI '05*, 111-120.

McCrickard, D.S. and Chewar, C.M. (2003). Attuning notification design to user goals and attention costs. Communications of the ACM, 45(3): 67-72.

McDonald, S. and Stevenson, R. J. (1998). Effects of text structure and prior knowledge of the learner on navigation in hypertext. *Human Factors*, 40(1): 18-27.

McKeown, K. R., Elhadad, N., and Hatzivassiloglou, V. (2003). Leveraging a common representation for personalized search and summarization in a medical digital library. In *Proceedings of ICDL '03*, 159-170.

Mitchell, J. and Shneiderman, B. (1989). Dynamic versus static menus: An exploratory comparison. *ACM SIGCHI Bulletin*, 20(4): 33-37.

Mizzaro, S. (1997). Relevance: The whole history. *JASIST*, 48(9): 810-832.

Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of SIGIR '94*, 272-281.

Morrison, J., Pirolli, P., and Card, S. (2001). A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. In *Proceedings of CHI '01*, 163-164.

Muramatsu, J. and Pratt, W. (2001). Transparent queries: Investigating users' mental models of search engines. In *Proceedings of SIGIR '01*, 217-224.

Murdock, B. B. (1962). The Serial Position Effect of Free Recall. *Journal of Experimental Psychology*, 64: 482-488.

Nardi, B., and Barreau, D. (1995). Finding and reminding: File organization from the desktop. *ACM SIGCHI Bulletin*, 27(3): 39-43.

Nielsen, J. (1998). Personalization is overrated. Retrieved March, 2006 from http://www.useit.com/alertbox/981004.html.

Nielsen, J. (2006). Usability 101: Introduction to usability. Retrieved March, 2006 from http://www.useit.com/alertbox/20030825.html.

Nowell, L., Hetzler, E., and Tanasse, T. (2001). Change blindness in information visualization: A case study. In *Proceedings of INFOVIS '01,* 15-22.

Ntoulas, A., Cho, J., and Olston, C. (2004). What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of WWW'04*, 1-12.

O'Day, V. and Jeffries, R. (1993). Orienteering in an information landscape: How information seekers get from here to there. In *Proceedings of CHI '93*, 438-445.

Østerbye, K. (1992). Structural and cognitive problems in providing version control for hypertext. In *Proceedings of European Conference on Hyptertext*.

Palen, L. and Salzman M. (2002). Voice-mail diary studies for naturalistic data capture under mobile conditions. In *Proceedings of CSCW '02*, 87-95.

Park, S-T., Pennock, D. M., Giles, C. L., and Krovetz, R. (2002). Analysis of lexical signatures for finding lost or related documents. In *Proceedings of SIGIR '02*, 11-18.

Patil, S., Alpert, S.R., Karat, J., and Wolf, C. (2005). "THAT's What I was Looking For": Comparing User-Rated Relevance with Search Engine Rankings. In *Proceedings of Interact '05*, 117-129.

Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., and Breuel, T. (2002). Personalized search. *Communications of the ACM*, 45(9): 50-55.

Preece, J. (1998). Reaching out across the Web. *Interactions*, 5(2): 32-43.

Raghavan, V. and Server, H. (1995). On the reuse of past optimal queries. In *Proceedings of SIGIR '95*, 344-350.

Rainie, L. and Shermak, J. (2005). Pew Internet and American Life Project: Data memo on search engine use. Retrieved January, 2006 from http://www.pewinternet.org/pdfs/PIP_SearchData_1105.pdf.

Ramakrishnan, N. (2005). The traits of the personable. In B. Mobasher and S. S. Anand (Eds.), *LNCS/LNAI State-of-the-Art Survey on Intelligent Techniques in Web Personalization*. Berlin: Springer-Verlag.

Ravasio, P., Schär, S. G., and Krueger, H. (2004). In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Transactions on Computer-Human Interaction,* 11(2): 156-180.

Reich, V. and Rosenthal, D. S. H. (2002). LOCKSS: A permanent Web publishing and access system. *D-Lib Magazine*, June.

Rekimoto, J. (1999). Time-machine computing: A time-centric approach for the information environment. In *Proceedings of UIST '99*, 45-54.

Rensink, R. A. (2002). Internal vs. external information in visual perception. In *Proceedings of the Symposium on Smart Graphics,* 63-70.

Rettig, M. (1994). Prototyping for tiny fingers. *Communications of the ACM*, 37(4): 21-27.

Ringel, M., Cutrell, E., Dumais, S. T., and Horvitz, E. (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. In *Proceedings of Interact '03*.

Rosenholtz, R., Li, Y., Mansfield, J., and Jin, Z. (2005). Feature congestion: A measure of display clutter. In *Proceedings of CHI '05*, 761-770.

Ross, N. C. M. and Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science*, 51(10): 949-958.

Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2): 95-145.

Salton, G. (1998). Automatic Text Indexing Using Complex Identifiers. In *Proceedings of the ACM conference on Document processing systems*, 135-144.

Schamber, L. (1994). Relevance and information behavior. *ARIST*, 29: 3-48.

Selberg, E. and Etzioni, O. (2000). On the instability of Web search engines. In *Proceedings of RIAO '00.*

Sellen, A. J., Murphy, R., and Shaw, K. (2002). How knowledge workers use the Web. In *Proceedings of CHI '02*, 227-234.

Shen, X. and Zhai, C. X. (2003). Exploiting query history for document ranking in interactive information retrieval. In *Proceedings of SIGIR '03*, 377-378.

Shiri, A. A. and Revie, C. (2003). The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment. *Journal of Information Science*, 29(6): 517-26.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum*, 33(1): 6-12.

Somberg, B. L. (1986). A Comparison of Rule-Based and Positionally Constant Arrangements of Computer Menu Items. In *Proceedings of CHI/GI '86*, 255-260.

Sparck Jones, K., Walker, S., and Robertson, S. A. (1998). Probabilistic model of information retrieval: Development and status. Technical Report TR-446, Cambridge University Computer Laboratory.

Speretta, M. and Gauch, S. (2004). Personalizing search based on user search history. Retrieved September, 2006 from http://www.ittc.ku.edu/keyconcept/

Spink, A., Wolfram, D., Jansen, B. J., and Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science*, 52(3): 226-234.

Spink, A., Jansen, B. J., Wolfram, D., and Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3): 107-109.

Strauss, A. and Corbin, J. (1990). *Basics of Qualitative Research Grounded Theory Procedures and Techniques.* Newbury Park, CA: Sage Publications.

Suchman, L. A. (1987). *Plans and Situated Actions*. New York: Cambridge University Press.

Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004). Adaptive Web search based on user profile constructed without any effort from user. In *Proceedings of WWW '04*, 675-684.

Tauscher, L. and Greenberg, S. (1997). How people revisit Web pages: Empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47(1): 97-137.

Tauscher, L. and Greenberg, S. (1997). Revisitation patterns in world wide Web navigation. In *Proceedings of CHI '97*, 399-406.

Teevan, J. (2001). Displaying dynamic information. In *Proceedings of CHI'01*, 417-418.

Teevan, J. (2004). How people re-find information when the Web changes. MIT AI Memo AIM-2004-012.

Teevan, J. (2005). The Re:Search Engine: Helping people return to information on the Web. In *Proceedings UIST '05*.

Teevan, J. (2005). The Re:Search Engine: Helping people return to information on the Web. In *Proceedings of SIGIR '05*.

Teevan, J., Adar, E., Jones, R., and Potts, M. (2005). History repeats itself: Repeat queries in Yahoo's query logs. In *Proceedings of SIGIR '06,* 703-704.

Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of CHI '04*, 415-422.

Teevan, J., Capra, R. G., and Pérez-Quiñones, M. A. (in press). How people find their personal information. In W. Jones and J. Teevan (Eds.), *Readings in Personal Information Management*. Seattle, WA: University of Washington Press.

Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Beyond the commons: Investigating the value of personalizing Web search. In *Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA)*.

Teevan, J., Dumais, S.T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of SIGIR '05*, 449-456.

Terry, W. S. (2005). Serial position effects in recall of television commercials. *The Journal of General Psychology*, 132(2): 151-163.

Tichy, W. F. (1985). RCS: A system for version control. *Software - Practice and Experience*, 15(7): 637-654.

Tognazzini (1999). A quiz designed to give you Fitts. Retrieved September, 2006 from http://www.asktog.com/columns/022DesignedToGiveFitts.html.

Tolle, J. E. (1984). Monitoring and evaluation of information systems via transaction log analysis. In *Proceedings of SIGIR '84*, 247-258.

Tulving, E. and Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychology Review*, 80: 352-373.

Varakin, D. A., Levin, D. T., and Fidler, R. (2004). Unseen and unaware: Implications of recent research on failures of visual awareness for human-computer interface design. *Human-Computer Interaction*, 19(4): 389-422.

Wang, P., Berry, M. W., and Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology,* 54(8): 743-758.

Weiss, R. (2003). On the Web, research work proves ephemeral. *Washington Post*, November 24: A08.

Wen, J. (2003). Post-valued recall Web pages: User disorientation hits the big time. *ITandSociety,* 1(3): 184-194.

Wen, J.-R., Nie, J.-Y., and Zhang, H.-J. (2002). Clustering user queries of a search engine. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining.*

White, R., Ruthven, I., and Jose, J.M. (2002). Finding relevant documents using top ranking sentences: An evaluation of two alternative schemes. In *Proceedings of SIGIR '02*, 57-64.

Whittaker, S. and Hirschberg, J. (2001). The character, value, and management of personal paper archives. *Transactions of Computer-Human Interaction*, 8(2): 150-170.

Whittaker, S. and Sidner, C. (1996). Email overload: Exploring personal information management of email. In *Proceedings of CHI '96*, 276-283.

Wildemuth, B. M. (2003). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3): 246-258.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation, 55*(3): 249-270.

Wilson, T. D. (2000). Human information behavior. *Informing Science, 3*(2): 49-55.

Xue, G.-R., Zeng, H.-J., Chen, Z., Yu, Y., Ma, W.-Y, Zi, W., and Fan, W. (2004). Optimizing Web search using Web click-through data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 118-126.

Yee, H., Pattanaik, S., and Greenberg, D. P. (2001). Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments. *ACM Transactions on Graphics (TOG)*, 20(1):.39-65.

Yee, K-P., Swearingen, K., Li, K., and Hearst, M. (2003). Faceted metadata for image search and browsing. In *Proceedings of CHI '03*, 401-408.

Zhang, Y. and Callan, J. (2005). Combining multiple forms of evidence while filtering. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.