How People Recall, Recognize, and Reuse Search Results

JAIME TEEVAN
Microsoft Research

When a person issues a query, that person has expectations about the search results that will be returned. These expectations can be based on the current information need, but are also influenced by how the searcher believes the search engine works, where relevant results are expected to be ranked, and any previous searches the individual has run on the topic. This paper looks in depth at how the expectations people develop about search result lists during an initial query affect their perceptions of and interactions with future repeat search result lists. Three studies are presented that give insight into how people recall, recognize, and reuse results. The first study (a study of *recall*) explores what people recall about previously viewed search result lists. The second study (a study of *recognition*) builds on the first to reveal that people often recognize a result list as one they have seen before even when it is quite different. As long as those aspects that the searcher remembers about the initial list remain the same, other aspects can change significantly. This is advantageous because, as the third study (a study of *reuse*) shows, when a result list appears to have changed, people have trouble re-using the previously viewed content in the list. They are less likely to find what they are looking for, less happy with the result quality, more likely to find the task hard, and more likely to take a long time searching. Although apparent consistency is important for reuse, people's inability to recognize change makes consistency without stagnation possible. New relevant results can be presented where old results have been forgotten, making both old and new content easy to find.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Graphical user interfaces*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*User issues*

General Terms: Design, Experimentation, Human Factors

Additional Key Words and Phrases: Refinding, recall, recognition, reuse, search, dynamic information, personal information management

ACM Reference Format:

Teevan, J. 2008. How people recall, recognize, and reuse search results. ACM Trans. Inform. Syst. 26, 4, Article 19 (September 2008), 27 pages. DOI = 10.1145/1402256.1402258 http://doi.acm.org/10.1145/1402256.1402258

Authors' address: One Microsoft Way, Redmond, WA 98005.

1. INTRODUCTION

Finding information is one of the most basic personal information management tasks [Jones 2007]. Most of the finding activities that take place on a person's desktop computer involve refinding, or the finding of previously viewed information. However, refinding is also a particularly common type of finding on the Web [Teevan et al. 2007]. Refinding is particularly challenging in environments like the Web where most information is outside of the searcher's direct control because previously viewed information can move, change, or disappear entirely without the searcher's knowledge. Potentially beneficial changes intended to support the finding of new information often interfere with refinding. For example, when a search engine improves the search results for a query it helps searchers encounter new, more relevant results. However, such changes can also interfere with people's ability to refind previously viewed results because results no longer appear where expected [Obendorf et al. 2007; Teevan et al. 2007]. Although it can be tempting to ignore this conflict and build systems that only support new-finding or refinding without consideration of the other behavior, people regularly do both simultaneously, by, for example, clicking on previously viewed results and new results during the same repeat search session [Teevan et al. 2007].

Consider Connie's searches for breast cancer treatments as an example of simultaneous new-finding and refinding. Connie was recently diagnosed with breast cancer and wants to learn about available treatments. The result list returned for her initial query for "breast cancer treatments" is shown on the left in Figure 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 appear too technical to interest Connie, and she is not generally interested in learning about alternative treatments, so she skips over the first couple of results in the list and decides to follow the fourth link to the About.com page. This initial interaction colors her future interactions with breast cancer search results.

As Connie explores treatment options, it is possible for the search engine to identify better results. Connie provides implicit feedback about what she considers relevant and irrelevant in the links she chooses to follow. She may also be willing to provide explicit feedback or query refinements because this topic is important to her. Further, her information need may evolve in predictable ways as she learns more about the topic, and new timely information about the latest treatments may become available as her search extends over time. Although new, more relevant results can benefit Connie, naïvely reranking the search results she has already seen to place the better results first is not necessarily the best way to help her satisfy her information need. Connie has developed expectations about what results the search result list for "breast cancer treatments" contains during her initial interaction with the list. If, for example, the About.com page she clicked on were no longer ranked about fourth in the list, she might have trouble returning to it because she is likely to look for it there.

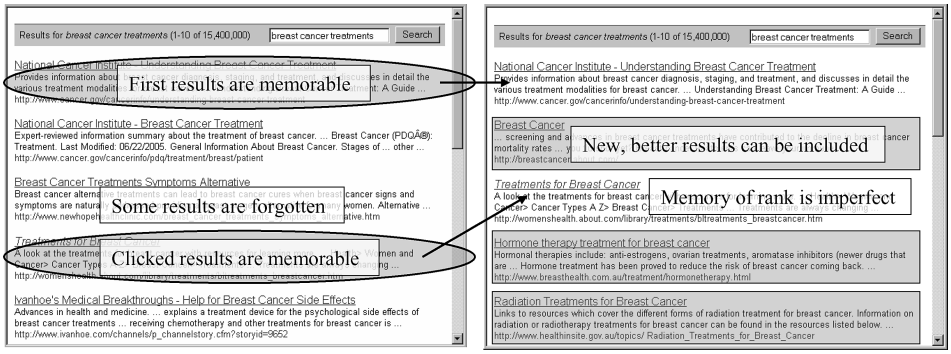

Fig. 1. On the left is the result list originally returned to Connie for the query "breast cancer treatments." On the right is the result list returned at a later date. It contains the results that Connie remembers having seen before where she expects them, while still including new results.

The three studies presented in this article and summarized in Table I build on each other to give insight into why changing Connie's search result list can cause her problems. They also address how Connie's past interactions can be accounted for to allow her to easily refind the information she found before while still interacting with newly relevant, timely results. The first study, a study of search result *recall* (Table I, column 1), explores what people find memorable about search result lists. Although the study reveals that people do not recall much, the searcher's limited memory can actually be used to the searcher's advantage. New, more relevant results can be placed where old results are forgotten to create a list that matches expectation but contains valuable new information. For example, the result list shown on the right in Figure 1 maintains the results that Connie is likely to remember from the original list (Figure 1, left), such as the first result and the result she clicked on, but also includes valuable new results, such as a page about hormone therapy treatment. The second study, a study of search result *recognition* (Table I, column 2), explores how what people recall about result lists influences how people recognize previously viewed search results. The study reveals that new search results lists look the same as previously viewed result lists as long as the memorable aspects of the lists are preserved. This is important because the third study, a study of search result *reuse* (Table I, column 3), demonstrates that when people believe they have seen a result list before, they are more likely to be able to reuse the previously found information contained in it, find what they are looking for quickly, be happy with the result quality, and find the task at hand easy. While matching the expectations a person develops during an initial search is clearly important for refinding, expectation matching does not necessarily interfere with the finding of new information. New results are easy to find when placed where old results have been forgotten.

2. RELATED RESEARCH

Refinding as a personal information management activity has recently attracted considerable interest [Aula et al. 2005; Bruce et al. 2004; Capra and

Table I. Summary Details the Three Studies Presented in this Paper
The first column represents a study of what people recall about search results, the second a study of what people recognize about search results, and the third of how people reuse search results.

	Recall	Recognize	Re-use
Number of participants	119	165	42
Experimental design		Between subject	Within subject
Session 1	1 self-selected query	1 self-selected query	12 pre-selected queries
Session 2	Recall result list	Recognize result list	6 refinding tasks 6 new-finding tasks
Interval between sessions	1 hour	1 hour	1 day
Discussion	Section 3	Section 4	Section 5

Pérez-Quiñones 2005; Obendorf et al. 2007; Sanderson and Dumais 2007; Teevan et al. 2007]. Repeat searches like Connie's are a common way to revisit personal Web content [Obendorf et al. 2007]. Teevan et al. [2007] found that 33% of all search engine queries have been issued before by the same user. As discussed by Jones [2007], many researchers consider all information experienced by a person, including previously viewed Web information, to be personal information, even if the information remains outside of a person's direct control. The studies presented here focus on understanding how people recall, recognize, and reuse search result lists containing previously viewed Web information, but the lessons learned are likely to apply to search results containing other types of personal information.

Related research suggests that people's previous interactions with search result lists are important for their understanding of future result lists. Consistency in information presentation is important, and often changes to electronic information that should help the user (such as the inclusion of new treatment options in Figure 1) can get in the way. Dynamic menus, for instance, were developed to help people access menu items more quickly than traditional menus by bubbling commonly accessed 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 and Shneiderman 1989; Somberg 1987]. Problems resulting from change have been observed for search results as well [Obendorf et al. 2007]. In a study of search result stability, Selberg and Etzioni [2000] noted 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." Teevan et al. [2007] demonstrated the veracity of this statement via large scale log analysis. They found that searchers take significantly longer to click on a repeat search result during a repeat query when the result list had changed. Another example of the difficulties caused by result list change can be found in a study by White et al. [2002]. In this study, the authors tried to help people search by giving them lists of relevant sentences that were dynamically reranked based on implicit feedback gathered during the search. However, people did not enjoy the search experience as much or perform as well with the dynamic system as they did when the sentence list was static.

Some search tools, like the "Stuff I've Seen" system [Dumais et al. 2003], have been developed to specifically support refinding. Such tools typically help users find previously viewed documents by allowing them to take advantage of the large amount of metadata people know about refinding targets through sorting and filtering. However, such tools make the process of returning to a document different from the process by which it was originally encountered. The tools do not explicitly support the repeat use of search for refinding, and do not consider a searcher's expectation of where results will be ranked. For example, if Connie were to repeat her search for "breast cancer treatments" using a refinding search tool, she may be able to specify when she last saw the about.com page she liked and the page's domain, but her search results will most likely be ordered differently from when she first ran the query.

Information management systems that preserve consistency of interaction despite change permit their users to choose to interact with a cached version of their information space [Hayashi et al. 1998; Rekimoto 1999]. As an example, Rekimoto [1999] developed a system that allows people to use their desktops to "time travel" to specific information environments that existed in the past. Similarly, history-based Web tools [Komlodi 2004; Komlodi et al. 2006; MacKay et al. 2005] tend to do a good job of preserving the landmarks a searcher may remember about previously viewed content. However, operating within a static world denies users the opportunity to simultaneously discover new information. With such systems, Connie could not, for example, revisit previously found information on breast cancer treatments while still learning about newly available treatments. Support for simultaneous finding and refinding is important because the finding of new information while refinding is common. Teevan et al. [2007] found that 27% of repeat searches involve clicks on new results as well as previously clicked results.

The studies presented later in this article explore how to take advantage of what people recall about search result lists so that they can easily interact with old and new search results at the same time. A method is presented for creating result lists for previously issued queries where perceived consistency is maintained, so that result lists appear unchanged even though they include new and potentially better results. This method is modeled on the concept of *change blindness*. Change blindness is a visual phenomenon where obvious changes to a scene occur without the viewer's notice as a result of limitations on human memory capacity and attention [Simons and Rensink 2005]. As an example, the difference between the two photographs in Figure 2 is obvious when they are viewed side by side—one picture has a crosswalk and the other does not. But when the two pictures are flashed sequentially, separated by a small gap in time, most people cannot identify the difference—even when actively looking for a change. Several researchers in human-computer interaction have expressed interest in how change blindness might affect users' ability to interact with computer-based information [Durlach 2004; Nowell et al. 2001; Varakin et al. 2004]. Their research, however, has focused on the fact that people may miss important changes 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.

Fig. 2. An example of a large change to a photograph that may not be noticed due to change blindness. When viewed side-by-side, it is obvious that the lines of the crosswalk are present in only one picture. But when flashed sequentially, most people cannot identify the difference—even when looking for a change.

In the research presented here, changes like those to the search result list in Figure 1 are intended to pass unnoticed, much as the changes to the picture in Figure 2 do.

Although result lists that contain new information can be made to appear the same as previously viewed result lists, the study of search result reuse presented here demonstrates that the inclusion of new and better results nonetheless can help satisfy the user's information need. Usability improvements do not need to be noticed to benefit the user. A classic example of this 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 any change [Ahlström 2005].

3. HOW PEOPLE RECALL SEARCH RESULTS

Three studies were conducted to understand how people recall, recognize, and reuse search results. The first study looked at what people recall about the search result lists with which they interact. People's memories of result lists are important to understand because they in turn influence future recognition and reuse of the list. This section begins by describing the recall study methodology, and then presents the study's findings. A preliminary version of this study is presented in a prior poster [Teevan 2006].

3.1 Recall Study Methodology

The design of the recall study, like all three studies presented in this paper, consisted of two sessions. During Session 1 participants were exposed to one (or more, in the case of the reuse study) query result list, and during Session 2 participants were asked to recall (and, in the case of the other two studies,

recognize or reuse) the results seen during the first session. All three studies were conducted on the participant's own computer using interactive Web forms. No one person participated in more than one study. The general variation across the three studies is summarized in Table I. This section describes the methodological details that relate specifically to the recall study.

3.1.1 *Session 1.* During the first session of the recall study participants were asked to enter a self-generated query into a search box. In response to the query, a list of query results was fetched from a leading search engine and returned. Participants were asked to interact with the list as they would normally. While typical studies of list recall [Henson 1998; Murdock 1962] require all items to be attended to, participants were not required to view every result. By allowing natural interaction, the study revealed which aspects of the result lists were both attended to and remembered.

The queries people entered and the results they clicked on were logged. The observable behavior captured through the study was similar to behavior commonly observed for Web search. The average initial query length was 3.2 words, which is somewhat higher than, but comparable to, what has been found through Web query log analysis [Spink et al. 2001]. 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].

3.1.2 *Session 2.* A half hour later, participants were emailed a survey that asked them to recall the result list without referring back to it. The survey, shown in Figure 3, asked participants to remember the text of their query, the number of results returned, and basic information about each result, including its rank, title, snippet, URL, whether the URL was clicked, and if so, what the corresponding Web page was like. 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.

3.1.3 *Participants.* As with all three studies presented in this paper, participants were recruited via several mailing lists, including lists associated with the Massachusetts Institute of Technology, a book group, and parenting. People were not compensated for participation. Two hundred forty five people participated in Session 1 of the recall study, and 119 completed both sessions. The relatively high dropout rate (51%) may reflect the fact that recalling previously viewed search results is hard to do. Those people who did not complete the second session may have remembered very little about initial result list they saw during Session 1. However, the purpose of the study was to discover which aspects of a result list are more memorable than others, and that is most likely not a function of the absolute amount of information remembered.

The demographics of the 119 participants in this study and in the subsequent two studies are shown in Table II. A comparable number of men (52%) and women (45%) participated in the study. Most participants (64%) were between the ages of 25 and 39, but 18% were over 40, and 15% under 25. Ninety-seven

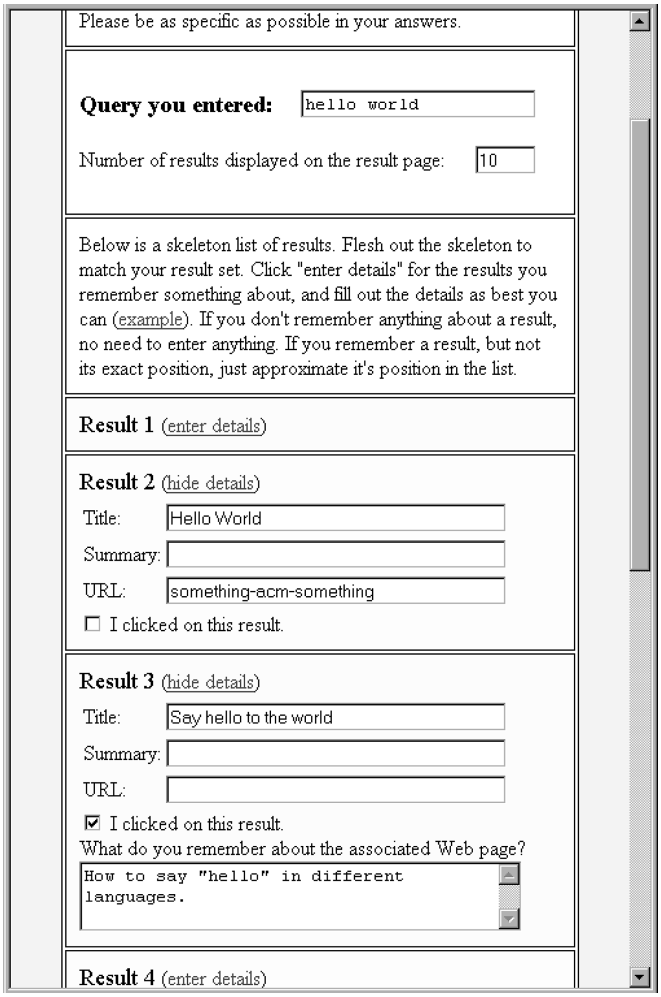

Fig. 3. The survey used in Session 2 of the recall study. The purpose of the survey was to prompt participants to recall previously viewed search results.

percent reported daily computer use. Twenty-seven percent of respondents were affiliated with MIT.

3.1.4 *Methodologies.* In general, during Session 2 participants recalled very little about the search results they interacted with during Session 1. Even though at most a few hours elapsed between the time when the search result list was originally seen and the follow-up survey, a mere 15% of all results presented during Session 1 were described in any way. The majority of participants remembered nothing about any of the results (mode = 0), and on average, only 1.47 of the results from the original list were described.

For analysis purposes, it was necessary to determine whether a result was accurately remembered or not. This was challenging because descriptions of the results could be quite vague (e.g., the title of a result was described as

Table II. Demographic Information for the Studies Presented in this Article
The first column represents a study of what people recall about search results, the second of what people recognize about search results, and the third of how people reuse search results.

	Recall	Recognize	Re-use
Number of Participants			
Session 1	245	208	92
Session 2	119	165	42
Gender (Session 2)			
Male	52%	29%	50%
Female	45%	68%	48%
Not reported	3%	3%	2%
Age (Session 2)			
18–24	15%	15%	12%
25–39	64%	68%	69%
40+	18%	17%	19%
Not reported	3%	0%	0%
Computer use (Session 2)			
Daily	97%	97%	100%
Affiliation (Session 2)			
MIT	27%	17%	31%

"something shakespeare" when the query was for "shakespeare sites"). Two independent coders matched the participants' descriptions of the recalled results with the actual result list they viewed with an 84% inter-rater reliability. One hundred and eighty nine results were described richly enough for both coders to make the same match, and these results were considered to have been "memorable." The memorable results were analyzed to provide insight into how to predict which results will be remembered and what ordering changes will be noticed.

3.2 Recall Study Results

3.2.1 *What Makes a Result Memorable.* Two main factors emerged from the data as affecting how likely a result was to be remembered: where in the result list it was ranked, and whether or not the result was clicked.

Figure 4 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 has been observed in cognitive psychology literature [Murdock 1962]. Those results that are presented first are more memorable than later results and the results presented last are somewhat more memorable than earlier results. Highly ranked results appear particularly memorable. This is probably because top results get more attention than lower-ranked results that require scrolling to view. People tend to click most on highly ranked results [Joachims et al. 2005], while results "below the fold" (typically result 7 and below) are often never seen at all [Granka et al. 2004]. It could also be due to the "primacy effect" [Murdock 1962], a cognitive phenomenon where the first items in a list are more memorable.

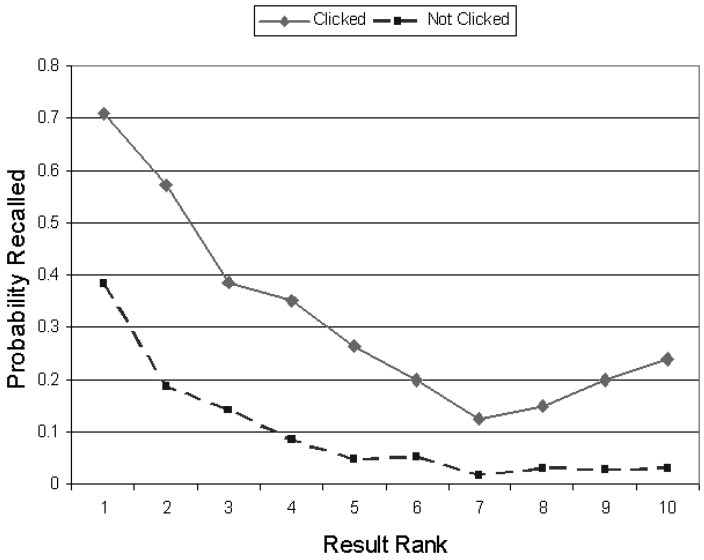

Fig. 4. The probability of recalling a result given rank. The probability generally decreases as a function of rank. Clicked results (solid line) were significantly more likely to be recalled ($p < 0.01$) than results that were not clicked (dotted line).

Whether a result was clicked affected how likely it was to be remembered. The importance of click through data has been studied for its value as an implicit measure to determine result quality [Joachims et al. 2005; Kelly and Teevan 2003]. In this analysis, click through is looked at as a way to determine how likely a result is to be remembered. Results that were clicked were significantly ($p < 0.01$) more likely to be recalled. Forty percent of the clicked results were remembered, compared with only 8% of the results that were not clicked.

Among the clicked results, the last results in the list appeared more memorable than previous results. The rise in the graph around result ten could be due to the "recency effect" [Murdock 1962], which indicates that the most recently attended to items in a list are particularly memorable. This would suggest why no similar increase appears among results that were not clicked. It is likely that the later-ranked nonclicked results were often not read. Thus the last result seen for nonclicked results varied as a function of the individual (e.g., the resolution of the participant's screens, how likely the participant was to review all of the results, etc.).

The last result clicked (which was not necessarily the last result in the list) appeared to be particularly memorable. A 12% increase in recall was observed if a result was the last result clicked, compared to other clicked results. The last result clicked may be particularly memorable because of the recency effect (it was also the last result seen), or because the result was what the participant was looking for. A result where the information contained in the result was actually used by the searcher is likely to be more memorable than a result that was viewed but merely examined.

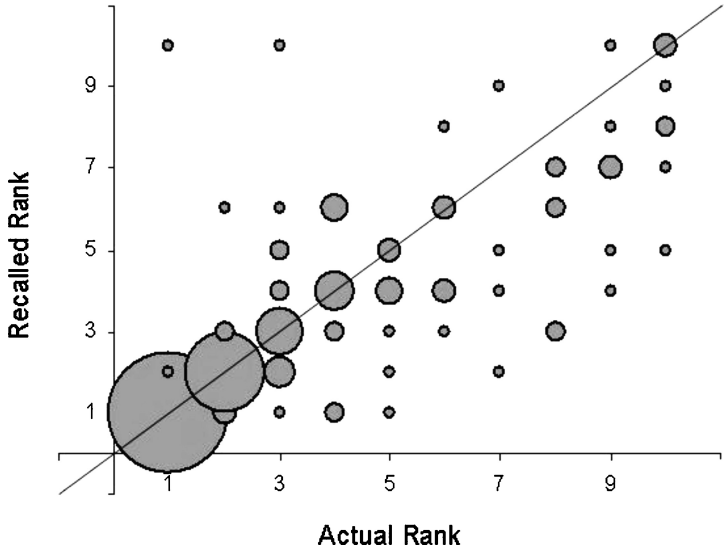

Fig. 5. The result's location in the result list as the participant remembered it, compared with the result's actual location. The size of each point represents the number of people remembering that combination.

3.2.2 *How Result Ordering Was Remembered.* Subjects' memories of result ordering were also analyzed to understand how changes to ordering might affect their ability to interact with a repeat search result list. Participants regularly made mistakes when recalling a result's rank. The recalled rank differed from actual rank 33% of the time. Mistakes were less common for early-ranked results. For example, the first result's rank was correctly recalled 90% of the time. Accuracy dropped for results the further down the list they were ranked. This can be seen graphically in Figure 5, which shows recalled rank as a function of actual rank. The importance of rank on memory implies that moving a result from the number one position in a result list is more likely to be noticed than moving a later-ranked result.

Figure 5 also illustrates another interesting trend in the data. The greater weight of the data occurs to the right of the diagonal line (along which actual and recalled rank are the same). This means that remembered results were much more likely to be recalled as having been ranked higher than they actually were. Those 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 could reflect 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.

It is interesting to consider the ramifications of the fact that people misremember result ranking. It suggests that it may be possible for a result list to look more like the result list a person remembers having seen than the actual list that person saw. In the following two studies there was a trend for results that were changed according to a model of how results are remembered to be

perceived as static more often even than result lists that were in fact static. While these findings are not significant, they could suggest that results should be placed where they are expected—even when that is not where the results originally occurred.

4. HOW PEOPLE RECOGNIZE SEARCH RESULTS

The ability of people to recognize a search result list as being the same as a previously viewed result list was evaluated in a study of search result recognition. The study showed that while most result lists that are different from an initial result list appear to be different, very different result lists can be recognized as the same if they maintain consistency in the recalled aspects. Here the recognition study methodology is presented and the study results are discussed.

4.1 Recognition Study Methodology

4.1.1 *Session 1.* The design of the first session of the recognition study was the same as the design of the first session of the recall study. Participants were asked to enter a single query of their choosing into a search box and interact with the returned results as they normally would. As with the recall study, the general search behavior observed during Session 1 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.

4.1.2 *Session 2.* A half hour after Session 1, participants were emailed a pointer to a survey that asked about the search they conducted during Session 1. In this survey, participants were asked to recognize whether a new result list was the same or different from the result list that they saw during Session 1. Participants who believed the Session 2 list was different were also asked whether the changed results were better, the same, or worse, and asked to describe any differences noticed. 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.

The result list that each participant was asked to recognize was constructed by merging together the results seen during Session 1 and a new set of results. The inclusion of new results in a previously viewed result list is only beneficial when the new results are more relevant to the searcher's needs. To reflect this desired usage scenario, the results returned during Session 1 were not actually the most relevant results available, but rather were the results 11 through 20 returned by the underlying search engine. This enabled higher quality results (results 1 through 10) to be merged with the Session 1 list during Session 2.

The merged list a participant saw during Session 2 was constructed in one of the following five ways:

1. *Random Merge*. Four of the results viewed during Session 1 were merged at random into an ordered result list containing the top six new results.

Table III.
Examples of the five different merge types explored in the recognition study. Assumes result 1 and 9 in the original list were clicked.

Merged Rank	Merge Type				
	Original	New	Random	Clicked	Intelligent
1	Old result 1	*New result 1*	*New result 1*	Old result 1	Old result 1
2	Old result 2	*New result 2*	Old result 2	Old result 9	Old result 2
3	Old result 3	*New result 3*	*New result 2*	*New result 1*	Old result 3
4	Old result 4	*New result 4*	*New result 3*	*New result 2*	*New result 1*
5	Old result 5	*New result 5*	Old result 3	*New result 3*	*New result 2*
6	Old result 6	*New result 6*	Old result 1	*New result 4*	*New result 3*
7	Old result 7	*New result 7*	*New result 4*	*New result 5*	Old result 9
8	Old result 8	*New result 8*	*New result 5*	*New result 6*	*New result 4*
9	Old result 9	*New result 9*	*New result 6*	*New result 7*	*New result 5*
10	Old result 10	*New result 10*	Old result 4	*New result 8*	*New result 6*

2. *Clicked Merge*. The results clicked during Session 1 were ranked first, followed by the new results. The exact number of results preserved varied as a function of how many results are clicked.
3. *Intelligent Merge*. Old and new results were merged with an attempt to preserve the memorable aspects of the list seen during Session 1 (described in greater detail below). On average, the merged list contained four results viewed during Session 1 and six new results.
4. *Original*. No merging was done. The result list was exactly the same as the originally viewed list. This is what a user of a system that cached previously viewed result lists would see.
5. *New*. The list was comprised of entirely new results.

An example of each of these merge types can be seen in Table III.

The intelligent merge algorithm was built on the results of the recall study, and the details of its implementation can be found in prior work [Teevan 2007]. The intention of the intelligent merge is to preserve the memorable aspects of the original result list while including new results where previously viewed results have been forgotten. To do this, the value of each new result is balanced against the cognitive cost of changing the originally viewed result list. The value of including a new result in the merged list is calculated as a function of the result's rank as returned by the underlying search engine and how close to the top it will appear in the merged list. The value of including a previously viewed result in the merged list is calculated as a function of the result's likelihood of being remembered and how closely it will be ranked in the merged list to where it was ranked in the original result list. The recall study makes it possible to quantify how likely a previously viewed result is to be remembered (using whether it was clicked and its original rank, see Figure 4), and how likely that result is to be looked for at a particular location in the merged list (using its original rank, see Figure 5). All permutations of possible final lists that include at least a few old results and a few new results are considered, and the best result list is chosen. Several examples of merged lists are shown in Table IV.

Note that presenting old and new results in a single merged list is only one design alternative among many for simultaneously presenting both types of

Table IV.
The rank of new results and results from the original result list after an intelligent merge, as a function of what results were clicked.

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*

results. Another alternative, for example, could show old and new results in two separate side-by-side lists. The research presented here focuses on designs that maintain the familiar single ranked-list presentation. Neither old nor new content is not called out in the ranked list, by, for example, highlighting new results or adding to previously viewed results the date the result was visited. However, any alternative that presents both new and old information in a single list (whether it does so visibly or invisibly) faces the merging challenges that this work addresses.

4.1.3 *Participants.* A total of 208 people participated in Session 1, and 165 people completed Session 2. Each of the five types of merged lists was viewed by approximately 33 people. None of the people who participated in the recall study were included in this study. As in the recall study, people were not compensated for participation. Nonetheless, the response rate was much higher for the recognition study (79%) than for the recall study (49%). This may reflect the relative ease of recognizing information compared with recalling it.

Demographic information can be found summarized in Table II. 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.

4.2 Recognition Study Results

The results of the recognition study show that most methods for merging new results with previously viewed results create noticeably different result lists. However, as long as memorable aspects of the original result list are preserved, changes to the unmemorable aspects appear to go unnoticed. This finding is discussed in greater detail below, followed by evidence that people find result quality worse when they notice a change to the result list even when the quality is objectively better.

Table V.
Results from the recognition study. While participants noticed changes to the result list when changes were made naively (new, random, and clicked), they did not when memorable information was preserved (original and intelligent).

	User Judgment:	
Merge Type	Same	Different
New	19%	81%
Random	38%	62%
Clicked	41%	59%
Original	69%	31%
Intelligent	81%	19%

4.2.1 *Lists That Account for Previous Interactions Appear Unchanged.* As shown in Table V, 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 may have found memorable (i.e., the random merge, the clicked merge, the new result list). When the follow-up results list was comprised of entirely new results, participants reported it had changed 81% of the time. When four random results were held constant (random merge), 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. The differences between the three cases are not significant, except that the difference between the clicked merge and the new list is weakly significant ($p < 0.05$).

The remaining two cases (the original result list and the intelligent merge) 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 case where new results were merged in intelligently, where differences were noted only 19% of the time. This disparity is not significant, but as mentioned earlier could 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 intelligent merge and original list were significantly more likely to be considered the same as the original list than the random merge, clicked merge, or new result list ($p < 0.01$).

The significant difference between the intelligent merge and the two other merge algorithms may appear somewhat surprising, since, for example, the random and intelligent merges both contained the same number of new and old results. However, several important aspects varied between the two merge types, including which results were preserved and the preserved results'

ordering. The difference in perceived change suggests that these two aspects are important for recognition.

4.2.2 *Apparent Consistency Leads to Higher Perceived Quality.* While the recognition study revealed that it is possible for people to recognize a result list containing new information as being the same as a previously viewed result list, it is not obvious that people want new relevant results to appear the same as previously viewed 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 result lists that appeared to have changed during Session 2 was studied. 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 are significantly correlated [Joachims et al. 2005; Patil et al. 2005]). This was confirmed by an independent coder, who judged the new result list 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, $p < 0.01$), and in fact found the changed result list was worse 14% of the time. This suggests that the better result lists were judged to be of worse quality merely because they were different from what was expected based on the participants' previous interactions.

5. HOW PEOPLE REUSE SEARCH RESULTS

To further understand how apparent change affects people's ability to find and refind information, an third, more controlled study was conducted. In this study, people were asked during Session 2 to reuse search results found during Session 1, as well as to find new results. While a static result list works well to support refinding, it does not support the finding of new information. In contrast, a result list with only new results supports the finding of new information, but does not support refinding well. Returning results that appear static but contain new information appears to perform almost as well in both cases.

5.1 Reuse Study Methodology

5.1.1 *Session 1.* Session 1 of the reuse study differed somewhat from Session 1 of the recall and recognition studies. Instead of asking participants to issue one self-selected query, they were asked to conduct 12 preselected finding tasks, presented in a random order. Although these design decisions affect the realism of the study, using preselected tasks minimized task effects and using 12 tasks enabled a within-subject design. Controlling for as much variation as possible was important to support the observation of relatively small differences across the conditions being tested.

Although the 12 tasks were pregenerated, care was taken to ensure they were as realistic as possible. Two of the Session 1 task descriptions are shown

Table VI.
Two of the queries and associated tasks used in the reuse study. All participants conducted the same task for each query during Session 1, and randomly completed either a refinding or new-finding task for the query during Session 2.

	Session 1	Session 2	
Query	Task	Re-finding Task	New-Finding Task
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.
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).

in the second column of Table VI. Each of the 12 tasks was inspired by 12 queries identified from a major search engine's logs as having been issued in a manner resembling the target refinding behavior being studied. According to the logs, each query was issued twice by the same individual and both new and previously clicked results were logged as having been clicked the second time the query was issued. Because the interval between Session 1 and Session 2 in the reuse study was one day, the 12 queries in the log 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. The resulting queries were approximately 2.4 words long, which is a very typical query length [Spink et al. 2001]. Two tasks were generated from each of the 12 queries and their top 20 results. To ensure consistency across tasks, the top 20 results were filtered so that the answer to each task could be found in one and only one search result in the result list, and result snippets were edited to ensure the answer did not appear in a snippet.

During Session 1, one of the two tasks associated with each query was presented to each participant. As shown in Figure 6, for each task in Session 1 participants were given the task description, the query, and a list of ten results. To avoid positional bias, the search result that contained the answer to the task was placed at a random location in the list. Participants were asked to identify the result from the list of results that satisfied the task.

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.

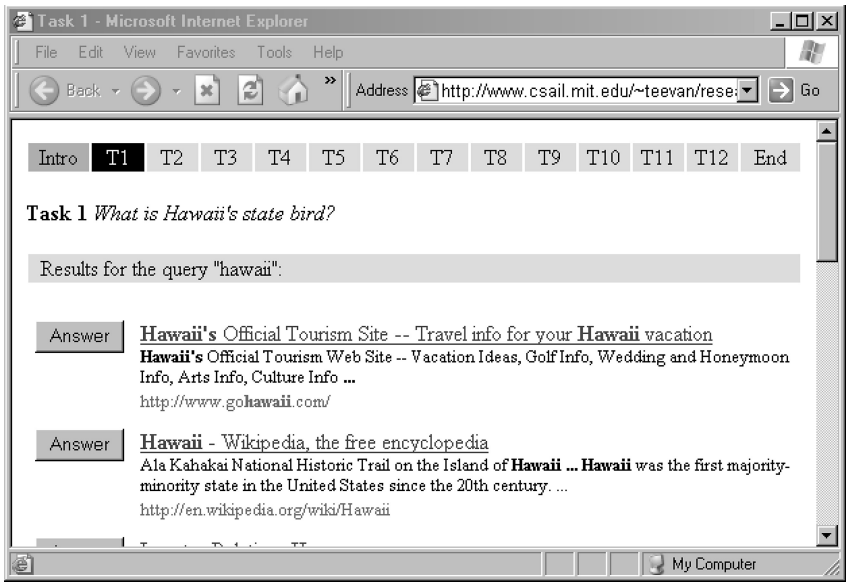

Fig. 6. An example task ("What is Hawaii's state bird") that participants completed during the reuse study.

5.1.2 *Session 2.* Session 2 of the reuse study was conducted the following day (mean = 1.0 days, median = 0.99 days). According to log analysis [Teevan et al. 2007], repeat searches are very common at this interval and involve repeat clicks 88% of the time and new clicks 27% of the time. During Session 2, 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 randomly designated refinding tasks (the same as the task conducted the previous day), and the other half were designated new-finding tasks (involved finding new information not previously available). Several refinding tasks and new-finding tasks for Session 2 can be seen in the right-hand columns of Table VI.

Four of the five merge types studied in the recognition study were used again in the reuse study: the original result list, the new result list, the random merge, and the intelligent merge. Because the clicked merge and the random merge appeared to be recognized similarly in the recognition study, one of the two merging was dropped in favor of collecting more data. The random merging was selected for the reuse study because it is easy to control the number of old and new results displayed.

To ensure consistency across task, each new-finding task was once again designed so that the answer could be found in one and only one of the new search results and not in the result snippet. The answer to the new-finding task could not be found in the initial search result list, but rather could only be found when new information was included in the result list for Session 2. As was done for the recognition study, the new search results were identified using results 1 through 10 returned from a Web search engine, while the search results returned during Session 1 were results 11 through 20.

Table VII.
Information about where the correct result occurred in the result list used in Session 2 of the reuse study, according to the merge type. The result that answered the refinding task always occurred in the original list. Although it was artificially maintained in the random and intelligent merges, it would have occurred naturally in 40% of the random merges and 78% of the intelligent merges. All lists with new results always contained the correct new answer. The correct new result was, on average, ranked earliest for the new list, next for the random merge, and lowest for the intelligently merge.

		Merge Type			
		Original	New	Random	Intelligent
Likelihood of containing the	Old answer	100%	0%	40%	78%
	New answer	0%	100%	100%	100%
Average rank of new answer		N/A	3.5	5.5	7

Each of the six refinding tasks a participant did was performed using the original result list, the random merge, or the intelligent merge. Each participant conducted two refinding tasks with each of these three merge types. To control for ordering effects, the ordering of which merge type was used for which task was selected randomly. The new list was not used for refinding tasks because refinding tasks could not be solved using only new information. Care was taken to ensure that the correct answer did appear in the randomly and intelligently merged lists. Had this not been done, the correct result would only have appeared in the random merging 40% of the time and in the intelligent merging 78% of the time, as shown in Table VII. The correct result was naturally preserved more often for the intelligent merge than the random merge because it takes into account the participant's previous interactions with the original list, and these interactions are influenced by which result is correct.

Each new-finding task was conducted with the new result list, the random merge, or the intelligent merge, and each participant conducted two of the new-finding tasks with each merge type. The original list was not included because the new-finding tasks could not be solved using the original list, whereas for each of these three merge types, the correct result for the new-finding task appeared. Because only the top six new results appeared in the random and intelligent merged lists, the correct result always appeared in the top six of the new result list as well. As can be seen in Table VII, this means the correct result occurred ranked more highly for the new result list (3.5th) than for the random (5.5th) or intelligent merged list (7th).

Note that the correct result for the refinding task and the correct result for the new-finding task only appeared simultaneously in the random merging and the intelligent merging. While the new result list may be good for the finding of new information, it cannot be used for refinding. And while the original list may be good for refinding previously viewed information, it cannot be used for new-finding.

Each task was timed, and the results that were clicked were logged. A task ended when the participant marked a result as relevant or gave up. As with Session 1, 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. 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 asked whether the list quality was better, worse, or the same.

5.1.3 *Participants.* Ninety-two people completed Session 1 of the reuse study, and 42 people completed both sessions. None of the participants took part in the recall or recognition studies. Because the reuse study was particularly involved, participants who completed both sessions were entered in a drawing for one of three $50 gift certificates.

Demographic information is summarized in Table II. 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. All reported daily computer use. Thirty-one percent were associated with MIT.

5.1.4 *Methodologies.* The data collected were analyzed to understand both of how well participants performed refinding and new-finding tasks under the different merge 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 included 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 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 interrupted tasks was discarded. In the analysis of refinding 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.

5.2 Reuse Study Results

The results of the reuse study suggest that knowledge was reused across the two sessions. Task performance during refinding in Session 2 was strongly correlated with whether the follow-up list looked the same as what the participant remembered from their initial search. Thus it is not surprising that the reuse study shows result reuse was easier for participants when using the intelligent merging—where the list tends to appear unchanged—than when using the random merging—where changes are often noticed.

5.2.1 *Knowledge is Reused across Sessions.* Table VIII 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

Table VIII.
Measures for new-finding and refinding tasks of the reuse study, broken down by session and task-type. The *p*-values for tasks performed during Session 1 and repeated during Session 2 are reported. Values significant at a 5% level are *italicized*.

	Session 1		Session 2				
	All Tasks		New-Finding		Refinding (v. Session 1)		
Measure	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 IX.
Measures for new-finding and refinding tasks of the reuse study, separated by whether participants thought the result list used for Session 2 was the same as the result list used during Session 1 or different. The *p*-values that are significant at a 5% level are *italicized*.

	New-Finding			Refinding		
Measure	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*

(137.2 seconds), but the refinding 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 VI. On the other hand, the Session 2 refinding 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 IX shows that for all measures except result quality, performance during refinding was significantly better than performance during the initial finding session. Clearly, the knowledge participants gained about the search results for the tasks during Session 1 helped them to refind information more quickly than they originally found it.

5.2.2 *Apparent Consistency Supports Refinding.* This section looks at how perceived change affected participants' performance. The next section provides details about actual and perceived change as a function of the type of list used during Session 2. Table IX shows people's performance along several different metrics for new-finding and refinding tasks, separated by whether the participant thought the result list they were given during the Session 2 task was the same as the result list they interacted with during Session 1 or different. For new-finding tasks, there was no significant difference in performance for any of the measures between instances when a person noticed a change to the list and when they did not. On the other hand, performance on refinding 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 vs. 2.21), took less time

Table X.
The percentage of time participants thought results in Session 2 of the reuse study were the same as what they saw during Session 1, as a function of task and merge type. The p-values that are significant at a 5% level are *italicized*.

Task Type	Merge Type	Results Perceived to be the Same	p-value (significance)		
			Random	Intelligent	New/Original
New-Finding	Random	50%		0.062	*0.006*
	Intelligent	61%	0.062		*0.001*
	New	38%	*0.006*	*0.001*	
Refinding	Random	60%		*0.008*	*0.006*
	Intelligent	76%	*0.008*		0.920
	Original	76%	*0.006*	0.920	

to complete the task (39.5 seconds vs. 94.8 seconds), and answered more tasks correctly (97% vs. 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 vs. 3.30) and the task difficulty to be lower (1.50 vs. 2.21).

These results suggest that perceived change generally correlates with one's ability to refind previously viewed information, but does not greatly correlate with one's ability to find new information. While people could notice change more often when they have difficulties refinding, the following section suggests that it is the perception of change that causes difficulties.

5.2.3 *Appearance of Change a Function of Follow-Up List.* As expected given the recognition study, whether the result list appeared to change between sessions was a function of merge type and whether the follow-up task was a refinding task or a new-finding task. Table X shows the percentage of time participants thought results were the same as a function of task and merge type. For new-finding tasks, the result list appeared the same as the associated list from Session 1 38% of the time when an entirely new list was returned. This is significantly ($p < 0.01$) less frequent than when the list contained new information merged in a random manner (appeared the same 50% of the time) or in an intelligent manner (appeared the same 61% of the time). There was no significant difference between the intelligent merge and the random merge.

For refinding tasks, the intelligently merged list actually appeared to be the same as often as the original list appeared the same, even though the merged list contained six new results. Seventy-six percent of the time both lists were marked as unchanged. This is significantly ($p < 0.01$) more likely than for the random merge. That participants noticed changes to the intelligent merge during new-finding as often as they did with the random merge, but less often during refinding, suggests that when they needed new information they were able to locate it, but that when new information was not central to their task, that information passed unnoticed.

5.2.4 *Intelligent Merging Good for Finding and Refinding.* Table XI shows how long it took participants to perform new-finding and refinding tasks, broken down by merge type. The amount of time taken to complete a refinding task

Table XI.
The time it took participants to complete the Session 2 task of the reuse study, as a function of task and merge type. The *p*-values that are significant at a 5% level are shown in *italics*.

Task Type	Merge Type	Task Time (seconds)		*p*-value (significance)		
		Mean	Median	Random	Intelligent	New/Original
New-Finding	Random	153.8	115.5		*0.037*	0.267
	Intelligent	120.5	85.5	*0.037*		0.280
	New	139.3	92	0.267	0.280	
Refinding	Random	70.9	37.5		*0.037*	*0.008*
	Intelligent	45.6	23	*0.037*		0.554
	Original	38.7	26	*0.008*	0.554	

was the lowest when a static result list was used, next best when the results were merged intelligently, and the worst when they were merged randomly.

The other measures indicate a similar trend, although not always at the same levels of significance. For example, participants were more likely to correctly answer the refinding task using an unchanged list or intelligently merged list, when compared to the random merging, but the difference is significant only for the unchanged list. However, it is worth noting that the difference between the two mergings is likely greater than observed. If the result being refound is not preserved, refinding is impossible. The study was designed to enforce the preservation of the refinding result, but as seen in Table VII this happened naturally more often for intelligent merge than the random merge (78% vs. 40% of the time). 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 merge type for the finding of new information was not significant. However, finding new information with the intelligently merged lists happens significantly faster than with a random merging. This may be because there is some knowledge reuse even when finding new information in the context of previously viewed results. In those cases, the participant may have learned which results do not contain the answer, and knows to avoid them, while with the random merging they may find it necessary to repeat their review of information they have seen before.

It is worth noting that the rank of the correct result for new-finding tasks was significantly lower ($p < 0.001$) when the results were intelligently merged than for either the new list or the random merging—appearing, as shown in Table VII, on average 7th in the list as opposed to 5.5th (random merging) or 3.5th (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 new results with the old list according to the random merging algorithm, on average two of the four results were merged in ahead of the correct result. In contrast, the intelligent merging was 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.

The subjective performance with each merge type was also analyzed. Refinding with the intelligent merging was considered weakly significantly ($p < 0.05$)

easier compared to refinding with the random merging. Result quality was considered weakly significantly ($p < 0.05$) better for the new list for new-finding tasks than for the random merging. This may be because the correct result was ranked higher for the new list for new-finding tasks (appearing 3.6th rather than 5.6th), but is unlikely since the intelligently merged list quality was higher than the random merged list, despite the correct result being ranked on average 1.4 results lower.

In general, for refinding tasks, task performance with the original result list appears to be best, followed by performance with the intelligently merged list, and then the random 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 random 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 seems to be the best compromise to support both finding and refinding. A static, unchanging result list works well to support refinding, 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 refinding well. The intelligent merging performs close to the best in both cases, while the random merging does comparatively worse.

6. CONCLUSION AND FUTURE WORK

This article has presented three studies of search result recall, recognition, and reuse. The studies demonstrated the importance of consistency during refinding. Result lists that appeared to change across query sessions created problems for participants. Although the ability to find new information can appear to be at odds with the ability to refind, a solution was presented where new results were ranked where changes to the result list would not be noticed. This allowed people to find new information as easily as if they were given all new information and to refind information as easily as if nothing had changed. To truly understand whether maintaining consistency improves the search experience or affects search behavior requires a longitudinal study, and this remains as future work.

The studies presented here assume some period of time has passed between repeat visits to search result lists. It will be interesting to explore how new results can be included without notice in lists that are actively being used. This could allow search engines to improve results using real time implicit relevance feedback without disrupting the user's search. Research into this area is currently under way.

Effectively meeting people's expectations based on previous interactions in dynamic information environments like the Web is essential to successfully supporting people's complex finding and refinding behavior. 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 and Microsoft, for example, have introduced desktop folders that base their content on queries and change as new information becomes available. As 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." The studies explored here are a good first step towards that end.

ACKNOWLEDGMENTS

This research has benefited greatly from valuable input from David R. Karger, Susan T. Dumais, Mark S. Ackerman, and Robert C. Miller.

REFERENCES

AHLSTRÖM, D. 2005. Modeling and improving selection in cascading pull-down menus using Fitts' law, the steering law and force fields. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Kellog, W., Zhai, S., Van der Veer, G. C., and Gale C., Eds. ACM Press, New York, NY, 61–70.

AULA, A., JHAVERI, N., AND KÄKI, M. 2005. Information search and re-access strategies of experienced Web users. In *Proceedings of the 14th International Conference on World Wide Web*, Ellis, A. and Hagino, T., Eds. ACM Press, New York, NY, 583–592.

BRUCE, H., JONES, W., AND DUMAIS, S. 2004. Keeping and refinding information on the Web: What do people do and what do they need? In *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology (ASIS&T)*. Information Today, Medford, NJ.

CAPRA, R. AND PÉREZ-QUIÑONES, M. A. 2005. Using Web search engines to find and refind information. *IEEE Comput. 38*, 10, 36–42.

DUMAIS, S. T., CUTRELL, E., CADIZ, J. J., JANCKE, G., SARIN, R., AND ROBBINS, D. C. 2003. Stuff I've Seen: A system for personal information retrieval and reuse. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Callan, J., Cormack, G., Clarke, C., Hawking, D., and Smeaton, A., Eds. ACM Press, New York, NY, 72–79.

DURLACH, P. J. 2004. Change blindness and its implications for complex monitoring and control systems design and operator training. *Hum.-Comput. Interact. 19*, 4, 423–451.

GRANKA, L. A., JOACHIMS, T., AND GAY, G. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. Sanderson, M., Järvelin, K. Allan, J., and Bruza, P., Eds. ACM Press, New York, NY, 478–479.

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 *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia: Links, Objects, Time and Space—Structure in Hypermedia*, Akscyn, R., Ed. ACM Press, New York, NY, 87–96.

HENSON, R. 1998. Short-term memory for serial order: The start-end model. *Cognitive Psych. 36*, 73–137.

JOACHIMS, T., GRANKA, L., PAN, B., HEMBROOKE, H., AND GAY, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'05)*, Baeza-Yates, R. A., Ziviani, N., Marchionini, G., Moffat, A., and Tait, J., Eds. ACM Press, New York, NY, 154–161.

JONES, W. 2007. Personal information management. *Ann. Rev. Inform. Science Tech. 41*, 453–504.

KELLY, D. AND TEEVAN, J. 2003. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum 37*, 2, 18–28.

KOMLODI, A. 2004. Task management support in information seeking: A case for search histories. *Comput. Hum. Behav. 20*, 163–184.

KOMLODI, A., SOERGEL, D., AND MARCHIONINI, G. 2006. Search histories for user support in user interfaces. *J. Amer. Soc. Inform. Sci. Tech. 57*, 6, 803–807.

LEVY, D. 1994. Fixed or fluid? Document stability and new media. In *Proceedings of the ACM European Conference on Hypermedia Technology*. Ritchie, I. and Guimarães, N., Eds. ACM Press, New York, NY, 24–31.

MACKAY, B., KELLAR, M., AND WATTERS, C. 2005. An evaluation of landmarks for refinding information on the Web. In *Proceedings of the Conference on Extended Abstracts on Human Factors in Computing Systems (CHI'05)*. Van Der Veer, G. C. and Gale, C., Eds. ACM Press, New York, NY, 1609–1612.

MITCHELL, J. AND SHNEIDERMAN, B. 1989. Dynamic versus static menus: An exploratory comparison. *ACM SIGCHI Bull. 20*, 4, 33–37.

MURDOCK, B. B. 1962. The serial position effect of free recall. *J. Experi. Psych. 64*, 482–488.

NOWELL, L., HETZLER, E., AND TANASSE, T. 2001. Change blindness in information visualization: A case study. In *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'01)*, Andrews, K., Roth, S., and Wong, P. C., Eds. IEEE Press, Los Alamitos, CA, 15–22.

OBENDORF, H., WEINREICH, H., HERDER, E., AND MAYER, M. 2007. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Rosson, M. B. and Gilmore, D. J., Eds. ACM Press, New York, NY, 597–606.

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 the International Conference on Human-Computer Interaction (INTERACT'05)*, Costabile, M. F. and Paterno, F., Eds. Springer, Berlin, Germany, 117–129.

REKIMOTO, J. 1999. Time-machine computing: A time-centric approach for the information environment. In *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology*, Zanden, B. V. and Marks, J., Eds. ACM Press, New York, NY, 45–54.

SANDERSON, M. AND DUMAIS S. 2007. Examining repetition in user search behavior. In *Proceedings of the 29th European Conference on IR Research, Advances in Information Retrieval*. Amati, G., Carpineto, C., and Romano, G., Eds. Springer, Berlin, Germany, 597–604.

SELBERG, E. AND ETZIONI, O. 2000. On the instability of Web search engines. In *Proceedings of the 6th Conference on Content-Based Multimedia Information Access, Recherche d'Informations Assistee par Ordinateur (RIAO'00)*, Mariani, J. and Harman, D., Eds. CID, Paris, France, 223–235.

SIMONS, D. J. AND RENSINK, R. A. 2005. Change blindness: Past, present, and future. *Trends Cognitive Sciences 9*, 1, 16–20.

SOMBERG, B. L. 1987. A comparison of rule-based and positionally constant arrangements of computer menu items. In *Proceedings of the Conference on Human Factors in Computing Systems and Graphics Interface (CHI+GI' 87)*. Carroll, J. M. and Tanner, P. P., Eds. ACM Press, New York, NY, 255–260.

SPINK, A., WOLFRAM, D., JANSEN, B. J., AND SARACEVIC, T. 2001. Searching the Web: The public and their queries. *J. Amer. Soc. Inform. Sci. Techn. 52*, 3, 226–234.

TEEVAN, J. 2006. How people recall search result lists. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI'06)*, Olson, G. and Jeffries, R., Eds. ACM Press, New York, NY, 1415–1420.

TEEVAN, J. 2007. The Re:Search Engine: Simultaneous support for finding and re-finding. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, Wellner, P. and Hinckley, K., Eds. ACM Press, New York, NY, 23–32.

TEEVAN, J., ADAR, E., JONES, R., AND POTTS, M. 2007. Information re-retrieval: Repeat queries in Yahoo's query logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, Kraaij, W., de Vries, A. P., Clarke, C. L. A., Fuhr, N., and Kando, N., Eds. ACM Press, New York, NY, 151–158.

TOGNAZZINI, B. 1999. A quiz designed to give you Fitts. http://asktog.com/columns/022Designed ToGiveFitts.html.

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. *Hum.-Comput. Interact. 19*, 4, 389–422.

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 the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'02)*, Järvelin, K., Beaulieu, M., Baeza-Yates, R., and Myaeng, S. H., Eds. ACM Press, New York, NY, 57–64.

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 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Neill, D. B. and Moore, A. W., Eds. ACM Press, New York, NY, 118–126.

Received June 2007; revised February 2008; accepted April 2008