Large Scale Query Log Analysis of Re-Finding

Sarah K. Tyler
University of California, Santa Cruz
Santa Cruz, CA, USA
skt@soe.ucsc.edu

Jaime Teevan
Microsoft Research
Redmond, WA, USA
teevan@microsoft.com

ABSTRACT

Although Web search engines are targeted towards helping people find new information, people regularly use them to re-find Web pages they have seen before. Researchers have noted the existence of this phenomenon, but relatively little is understood about how re-finding behavior differs from the finding of new information. This paper dives deeply into the differences via analysis of three large-scale data sources: 1) query logs (queries, clicks, result impressions), 2) Web browsing logs (URL visits), and 3) a daily Web crawl (page content). It appears that people learn valuable information about the pages they find that helps them re-find what they are looking for later; compared to the initial finding query, re-finding queries are typically shorter, and rank the re-found URL higher. While many instances of re-finding probably serve as a type of bookmark for a known URL, others seem to represent the resumption of a previous task; results clicked at the end of a session are more likely than those at the beginning to be re-found during a later session, while re-finding is more likely to happen at the beginning of a session than at the end. Additionally, we observe differences in cross-session and intra-session re-finding that may indicate different types of re-finding tasks. Our findings suggest there is a rich opportunity for search engines to take advantage of re-finding behavior as a means to improve the search experience.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – Query formulation, Search process.

General Terms: Human Factors, Measurement.

Keywords: Re-finding, query log analysis, Web search.

1. INTRODUCTION

Previous research has shown that 39% of all queries issued to a search engine are instances where a user returns to a Web page that that user has found before via a separate search [21]. Whether driven by a need to remember past information, to discover new dynamic content, or even mere chance, queries that lead to repeat clicks account for a large portion of search traffic. But despite its prevalence, relatively little is known about this type of re-finding behavior.

In this paper, we use large-scale log analysis to explore how re-finding differs from traditional new-finding for search. Via the analysis of the Live Search (now Bing) query logs, we investigate many features of re-finding queries, their associated clicks, and the sessions re-finding queries belong to. Through this analysis, we are able to understand the way people appear to tailor their queries to re-finding previously viewed content and pick up the threads of previous tasks. This analysis focuses on the *search engine* portion of re-finding behavior pictured in Figure 1.

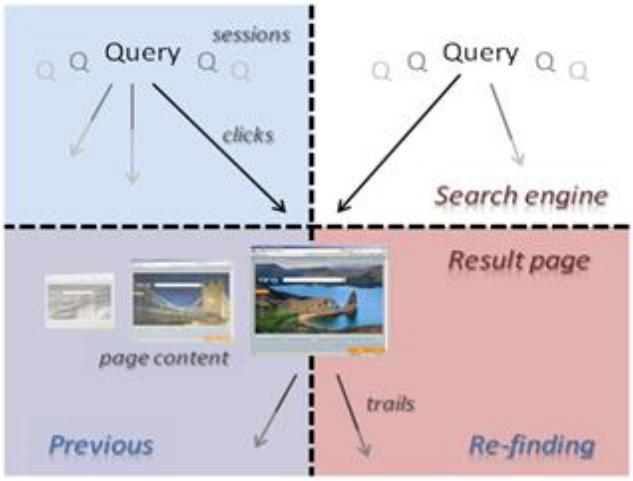

Figure 1. This paper looks at the different queries a person has used to find a particular page, as well as how those queries relate to the enclosing sessions, the content of the page over time, and the trails followed from the result page.

In addition to studying the aspects of re-finding that a search engine typically encounters, we also study aspects of the re-found result to better understand why the searcher might have been looking for that particular page and what they wanted to do once there (the *result page* portion of Figure 1). To study this, we supplement the query log analysis with analysis of the page's content, crawled daily, and with Web browser logs. This additional data enables us to study things like how the page content changes between visits and the consistency of the trails [26] people follow from the re-found pages. Table 1 lists the specific aspects discussed in this paper, broken down by whether the aspect pertains to the search engine (top portion of Figure 1) or result page (bottom portion of Figure 1), as well as whether the behavior is considered in isolation or as part of a sequence.

Following a discussion of related work and the definition of some useful terminology, we discuss the data sets used to study the different aspects of re-finding shown in Table 1 in greater detail. We then present our findings, including:

- The query used to re-find a result is typically better than the query used to initially find it. Re-finding queries are shorter than the first observed query associated with a given URL click, and rank the re-found URL higher. When re-finding occurs across multiple sessions, the re-finding query is also more common than the previous query.

WSDM'10, February 4-6, 2010, New York City, New York, USA.

- Re-finding queries tend to converge. When a person repeatedly uses a search engine to find the same result, the query used may differ some initially, but will become consistent.
- The need associated with a URL appears to be consistent when it is found by the same individual. A user who clicks a previously clicked result is more likely to follow the same path than other users clicking the same URL.
- Session-level and cross-session re-finding are very different. Cross-session queries change more substantially and in different ways than intra-session queries do. Cross-session re-finding may involve picking up a previous task. The queries at the beginning of a session are particularly likely to involve re-finding results found at the end of a previous session.

These findings reveal a rich opportunity for search engines to take advantage of re-finding behavior to improve the search experience. We end with a discussion of how this might be done.

Table 1. The features of re-finding studied in this paper, broken down by whether the behavior relates to the search engine or Web page being re-found, and whether it is considered in isolation or as part of a sequence.

Behavior		By itself	In a sequence
Search Engine	**Text**	Query length Substantial v. minimal Popularity	Sessions Re-finding chains
	Clicks	Result rank	Click order Query specificity
Page	**Text**	Textual content	Change in content
	Clicks	Initial link followed	Trail followed

2. RELATED WORK

There is good evidence that people tend to stay within a known space on the Web. Well over half of all of the Web pages a person visits are pages that person has seen before [8, 15, 19], and a third of the queries a person issues to Web search engines involve re-finding a previously found result [21].

Jones et al. [5] studied how people keep information encountered on the Web, and found that people store Web-based information for future use in many ways, including by doing nothing and relying instead on tools to help return. Numerous tools have been built and studied in support of revisitation, such as the Web browser back button [19], bookmarks [1], and browser histories [11]. Although a search engine is one of the most common tools used to return to previously viewed Web pages [6] and people develop expectations about the repeat search results based on the results they have seen before [20], most search engines currently do little to explicitly support re-finding [3].

Search tools that use an individual's past search behavior to improve the search experience do so via personalization [9; 13; 23]. For example, Teevan et al. [23] used previous clicks to indicate preferred sites for an individual to get information from. But personalization research has almost exclusively been conducted in support of finding new information, as opposed to re-finding. Raghavan and Sever [16] recognized that good queries are hard to formulate, and looked at storing complex queries for future re-use. Several recent search tools, such as the Re:Search Engine [22] and the SearchBar [14], are focused around re-finding as a core scenario. We believe that a better understanding of how people use search engines to re-find information can serve to inform the development of future search tools.

Web browser log-based studies of Web site re-visitation provide clues about how search engines are used for re-finding. For example, several studies [2, 15] have found that search engines are disproportionately useful when users want to return to infrequently revisited sites. Query log analyses focusing on long term querying trends [17] have tended to deal with aggregate user populations. For example, some researchers have investigated queries in aggregate over time to understand changes in popularity [24] and the uniqueness of topics at different times of day [4].

A few researchers have looked at long term querying trends dealing with individual users over time. Wedig and Madani [25] found that topics for a user are consistent over time and different from one another, and that some users repeat clicks over long time periods. Teevan et al. [21] showed that re-finding and repeat queries were very prevalent, and explored how queries used to re-find changed and how well future clicks could be predicted following repeat queries. Sanderson and Dumais [18] examined the temporal properties of an individual's repeated searches and clicks. They focused on the aspects of repeat queries related to time, finding, for example, that navigational queries are repeated over longer periods of time than non-navigational queries.

In this paper we build on this existing research to look more deeply at re-finding queries. We explore additional features of re-finding (e.g., the rank of a re-found result, the order of results clicked, the re-finding query's place in a session, the text of the result page, and the trail followed from the result page) and provide a rich picture of how elapsed time affects these features.

Many re-finding queries occur following very short time intervals. We explore the differences in how people re-find previously viewed Web pages both across multiple sessions and within an individual session. Some researchers have analyzed the queries issued over short periods of time by the same individual [10; 12]. This research has given insight into how queries in sessions evolve; however they have focused on the finding of new information and not previously viewed results.

3. DEFINITIONS

It is useful to begin our discussion of re-finding behavior defining some terminology. To make the definitions as clear as possible, we reference a hypothetical query log for a user in Table 2.

Re-finding When an individual clicks a URL following a search, and then later clicks on the same URL via another search, we call it *re-finding*. There are several examples of re-finding in Table 2. For example, the CDC Swine Flu Website found on Monday (C13) via the searcher's first query is re-found again later that same day (C_{32}) and on subsequent days (C_{41}, C_{61}, C_{71}).

Note that the query used to re-find a URL may differ from the query used to previously find it. In some cases the queries will be the same (Q_6 and Q_7, both *cdc swine flu*), while in other cases they will be very different (Q_4, *swine flu*, and Q_5, *h1n1*). The URL may not be the only clicked URL for a given query; in keeping with previous work [21], we consider the instance to be a re-finding if there is any click overlap.

Previous query, **re-finding query** In this paper, we focus on sequential pairs of re-finding queries. So Q_1 and Q_3 are a re-finding query pair, and Q_3 and Q_4 a re-finding query pair, but not Q_1 and Q_4. There can be intervening queries between a re-finding query pair that do not result in a click on the re-found URL (e.g., Q_2). The first query in a re-finding query pair we call the *previous query*; the second is the *re-finding query*.

Table 2. An example query sequence. The user finds the CDC H1N1 page (marked with a **) multiple times.

	Label	Query	Click
Monday	Q_1	*swine flu incidence*	
	C_{11}		healthmap.org/swineflu
	C_{12}		www.swine-flu-map-animation.com
	C_{13}		www.cdc.gov/H1N1Flu **
	Q_2	*swine flu deaths*	
	Q_3	*h1n1*	
	C_{31}		en.wikipedia.org/wiki/H1N1
	C_{32}		www.cdc.gov/H1N1Flu **
Tues.	Q_4	*h1n1*	
	C_{41}		www.cdc.gov/H1N1Flu **
	C_{42}		h1n1.nejm.org
Wed.	Q_5	*swine flu*	
	Q_6	*cdc swine flu*	
	C_{61}		www.cdc.gov/H1N1Flu **
Sat.	Q_7	*cdc swine flu*	
	C_{71}		www.cdc.gov/H1N1Flu **

We choose to only consider sequential pairs of re-finding queries to explore how re-finding behavior evolves, as each intermediate query represents the user's most recent experience finding the URL. Additionally, if we were to consider every re-finding pair in our analysis, the quadratic number of re-finding instances relative to URL clicks would overemphasize commonly re-found URLs.

Re-finding chain The previous query is not always the first query that resulted in a click on the re-found URL, as is the case with Q_3 in the Q_3, Q_4 query pair. In fact, even though Q_1 is the first query we observe to lead to the CDC Swine Flu page, it may also not be the first query that has ever lead to the re-found URL; it is only the first we observe. Instances of multiple re-findings of the same URL by a given user are referred to as *re-finding chains*. The queries $\{Q_1, Q_3, Q_4, Q_6, Q_7\}$ are a re-finding chain.

New-finding When a query is not used for re-finding, we call it a *new finding* query. Previous queries that are not also re-finding queries (e.g., Q_1, but not Q_3) are considered new-finding queries.

Minimal change When the previous query and re-finding query are very similar, we say there has been only a *minimal change* in the query used to find a particular URL. Minimal changes include instances where the two queries are exactly the same, have differences in capitalization, white space, alpha-numerics, stop words used, or word order. To capture misspellings, queries with a normalized edit distance of less than 0.05 or an absolute edit distance less than 2 are considered minimally changed. To capture instances where the searcher intends to type the same URL into the search box, URL fragments like ".com", "www", and "http://" are ignored.

Substantial change In other instances, the previous query and the re-finding query are quite different. A word may be added, removed, or swapped, or the query may be entirely different (e.g., Q_4 *h1n1* and Q_6 *cdc swine flu*). Queries are considered to have undergone a *substantial change* in any of these cases. Queries with substantial changes are interesting because they often reflect the fact that the searcher has developed a significantly different way of expressing their information target.

Session The queries and associated result clicks that occur within a short time window of search activity are considered to be part of a *session*. We use a 30 minute time out as a traditional and simple means of estimating sessions [7]. Unlike Teevan et al. [21], we treat identical queries issued by a user in a session as different query instances. We refer to the session surrounding a previous query as a *previous session*, and the session surrounding a re-finding query as a *re-finding session.*

Trail, hop After a person has clicked on a URL from a search result page, they may continue to follow links before moving on to their next action with the search engine. We call the links they follow their *trail*, and each link in the trail a *hop.* A trail starts at a search result click, and ends when the user does not click on a link for 30 minutes, uses a bookmark, closes their browser, enters an address on the address bar, or enters a new query in the search engine [26]. Note that if a trail is longer than 30 minutes, subsequent queries will be considered part of a new session, even if very little time elapses between end of the trail and the query.

When a trail is followed from a URL found via a previous query, we call it the *previous trail*, and when it is followed from a URL re-found via a re-finding query, we call it a *re-finding trail*. The re-finding trail, however, may or may not involve additional re-finding; it can be very different from the previous trail.

4. DATA SETS

We explore search engine re-finding behavior via analysis of three different datasets: 1) one which gives insight into the search engine-related behavior (search engine query logs), 2) another which gives insight into a searcher's behavior after leaving the search engine (Web browser logs), and 3) one which gives insight into the content of the found pages (a large-scale, daily Web crawl). All of these data sets, discussed in greater detail below, were collected during the month of January 2009.

4.1 Search Engine Query Logs

To understand the search engine's view of re-finding behavior, we studied the query logs from Live Search (now Bing), a major internet search engine. From the logs, we sampled information related to approximately 900 million search result clicks gathered from 106 million users. Similar to the example shown in Table 2, the sample included queries and clicked results, as well as time stamp information and the rank position of the clicked results. The sample was filtered to remove spam and processed so that pagination and back button clicks were treated as the same query. Only queries with at least one click were considered, in keeping with previous work [21].

Users were identified by an anonymous ID associated with a user account on a particular computer. As is the case with most log analyses, if a user has more than one computer, that user will have multiple IDs. Conversely, if more than one person uses the same account on a computer, they are amalgamated into a single user.

4.2 Web Browser Logs

Information about the trails people followed after running a search was collected via Web browser logs gathered from opt-in users of the Windows Live Toolbar. The toolbar provides augmented search features and reports anonymous Web usage behavior to a central server. Our analysis of the Web browser logs makes use of data from a sample of 4 million users and includes hundreds of millions of pages visits.

In addition to containing other URLs, the browser logs contain query URLs associated with multiple search engines, including Live Search, Google, and Yahoo. We used these search engine

URLs to identify search trails by extracting the queries from the URLs and analyzing where people went following a result click.

We also used the toolbar data to confirm that our findings using the Live Search query logs (Section 4.1) were consistent across a variety of different Web search engines. However, the analysis reported here uses the query logs instead of the toolbar logs whenever possible because that data is cleaner and more plentiful, applies to a broader number of users, and contains information about the results presented (order, etc.) in addition to just clicks.

4.3 Large Scale Web Crawl

To better understand the result pages people re-found, we also looked at the text content of the pages, captured via a large scale crawl of a sample of Web pages. To understand how the page content changed during the study periods, we crawled each page in our sample daily. At the onset of the data collection period, we did not yet know which Web pages would be re-found. Instead, we crawled pages that were sampled based on three different visitation-based attributes: the number of unique visitors to the page, the median time between user's visits, and the median number of visits per user. In total, 55,000 different pages were sampled. Additional information about the sampling process can be described in earlier work [2].

4.4 Relating the Three Datasets

The three datasets were related in that they covered the same time and referred to, in many cases, the same queries and URLs. Two URLs were considered the same if, based on their text, they appeared likely to refer to the same page. For example, it is common practice (although not always the case) for a primary domain and the subdomain of "www." to point to the same content, and thus the initial "www." was ignored. Additionally, a trailing slash usually does not alter the page content, and was thus ignored. We did not remove URL parameters as they can often lead to different page content.

4.5 User Study

Although the large-scale log and Web crawl data described above give a realistic picture of real world behavior, they do not provide insight into what the individual's intent is when a previously found result is found again. In order to get a better picture of whether a re-finding query was actually intended to re-find a particular URL, we conducted a small-scale critical incident user study of 9 individuals (7 males, 2 females).

Participants installed a Web browser plug-in on their primary work computer, and ran the plugin for several weeks. The plug-in logged the subject's search engine queries and result clicks, and occasionally popped up a survey following a result click to ask whether the subject had intended to find that particular URL with the issued query. The survey appeared following all re-finding clicks, and following 12.5% of all new-finding clicks. In total, we collected 159 responses.

5. FINDINGS ABOUT RE-FINDING

Using these datasets, we examine how people use search engines to re-find previously viewed results. We start our discussion by looking at re-finding in general, giving an overview of how prevalent re-finding is and what basic re-finding queries look like. We then explore how re-finding queries change, and show that when there are changes the re-finding query appears to be a better query than the previous query. We find that for multiple instances of re-finding by the same user for the same URL, the query used tends to converge to a single high quality query. We observe that people follow consistent trails from re-found results. We then investigate what may motivate the observed session-level differences, and show that re-finding may sometimes be a means of carrying tasks across sessions.

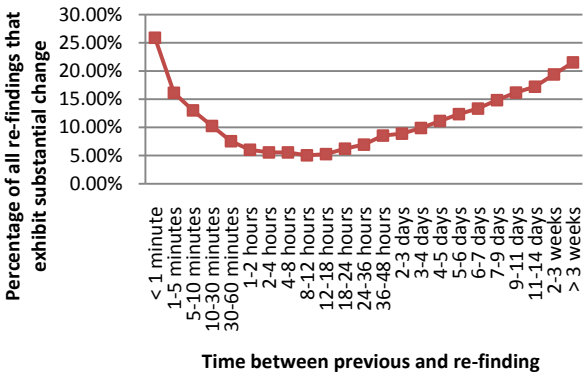

Figure 2. The percentage of re-finding queries that are substantially different from the associated previous query, as a function of the interval between the two queries. Queries are more likely to differ substantially when there is a very short or very long time interval between the re-finding and previous queries.

5.1 Overview of Re-Finding

In general, 21.9% of all of the queries we studied were observed instances of re-finding. This is somewhat lower than the 38.8% reported by Teevan et al. [21]. The difference almost certainly reflects the shorter time period studied (there is less opportunity to re-find with only one month of history versus a year) and the fact that we did not filter users to ensure a baseline amount of activity with the search engine per user (users who only appear in the logs for one query cannot re-find). Our value is a lower bound on the true incidence of re-finding during this time period.

Searchers appear to be targeting a particular URL more often during re-finding than new-finding. Participants in the user study reported intentionally seeking the clicked URL 48% of the time during re-finding and 30% of the time during new-finding. One participant was an outlier, and reported intentionally searching for the URL only 5% of the time. Excluding this participant, the difference is even more striking, with 72% of re-finding instances, and still only 30% of new-finding instances being intentional.

Single-click queries are particularly likely to involve re-finding; 29.6% of all single-click queries are re-finding queries. In contrast, the probability that a click during a multi-click query involves re-finding is only 5.3%. Although the first click following a query is always more likely to involve a previously found result than subsequent clicks, no click position has higher than a 7.2% probability of re-finding, regardless of click count for multiple click queries.

URLs that are re-found once are likely to be re-found again. On average, 66.1% of re-finding queries are also previous queries for a later re-finding. And if a re-finding query is minimally different from the previous query, the result is even more likely to be found again (69.2%). Query chains are discussed further in Section 5.3.

About half (48%) of all re-finding instances occur within a single session; the rest occur across sessions. The number of sessions between a re-finding query pair follows a long tail distribution, and averages 3.51. Re-finding is bursty, with re-finding queries appearing in groups. In a session, the query immediately after a re-finding query involves re-finding 59.3% of the time. Over half

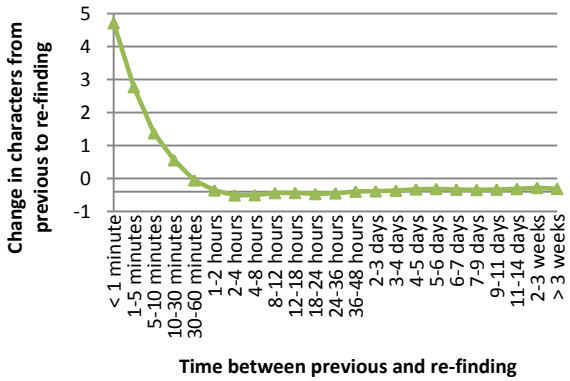

Figure 3. The change in query character length and query specificity for substantially changed queries, as a function of the time between the previous and re-finding queries. Within a session, re-finding queries are typically longer than their previous query counterpart, whereas across session they are typically shorter.

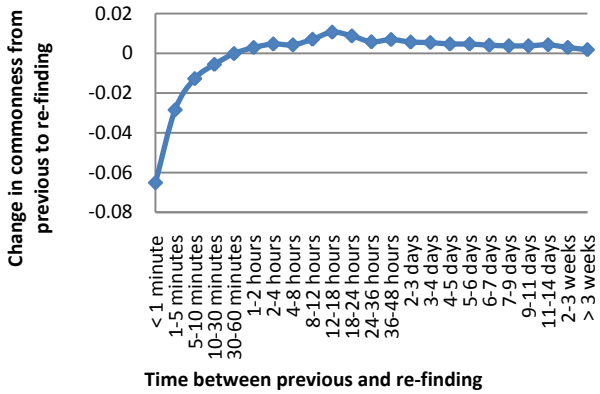

Figure 4. The change in how common the query is for substantially changed queries, as a function of the time between the previous and re-finding queries. Within a session, re-finding queries are typically less common than their pervious query counterpart, whereas across session they are typically more common.

(51.1%) of the subsequent three queries are likely to be re-finding, as are 46.6% of all remaining queries in the session, all much higher than the probability of a random query involving re-finding. Thus if a search engine observes a single instance of re-finding, it is likely to observe many more.

Most (79.2%) of the time when a result is re-found, the query used to re-find is exactly the same as the previous query, and an additional 11.4% involve only minimal changes. These findings are consistent with previous work [21]. The remaining 9.4% of re-finding queries are ones that undergo substantial changes between the previous query and the re-finding query. The data collected via the user study suggests substantial changes are more likely to occur when the re-finding query was not specifically intended to lead to a particular URL. When participants reported that their query was intended to find the re-found URL, the query changed substantially 25% of the time; in contrast, when the URL was not being sought in particular, it changed substantially 48% of the time. Because a substantial change can indicate that the searcher has a new way of expressing their information need based on previous information interactions, we look more closely at this subset of re-finding queries in Sections 5.2 and 5.3.

The percentage of re-finding queries that are substantially different from the associated previous query are shown in Figure 2 as a function of the interval between the two queries. Re-finding queries are least likely to change at intervals of about a day; revisitation of popular pages commonly follows a cyclical daily pattern [2], and this behavior may reflect re-finding using oft repeated, well learned query "bookmarks". Substantial query changes happen more often after short (less than an hour) or long revisitation intervals (a day or greater). These differences may reflect a qualitative difference in re-finding within a session as compared to across multiple sessions. How people use the search results they re-find is explored in greater detail in Section 5.4, and the differences between session-level re-finding and cross-session re-finding are discussed in Section 5.5.

5.2 Re-Finding Queries Are Better Queries

In this section we dive deeper into substantially changed re-finding queries. These queries provide a picture of how users modify their queries when the way they refer to their information target changes. The evidence suggests that searchers sometimes learn information about what they are looking for after the previous query that allows them to better express what they are looking for in the re-finding query. Our analysis (discussed below) shows that re-finding queries tend to be better queries than their corresponding previous queries; the queries become shorter, more common, rank the re-found result higher, and relate more directly to the text of the result.

5.2.1 Re-Finding Queries Shorter

Queries associated with re-finding are substantially shorter than queries not associated with re-finding. On average, a re-finding query is 12.1 characters long, and its associated previous query is 11.7 characters long. In contrast, queries used to find new results are 18.9 characters long.

Re-finding queries that change substantially from the previous query are much more likely to be longer queries. They have an average length of 18.6 characters, similar to that of new-finding queries. This may be a reflection of intent. As discussed earlier, our user study suggests substantial changes tend to occur when the searcher is not seeking a specific URL. In contrast minimal change queries have an average length of 11.4 characters.

The way length changes between queries varies as a function of the time interval between queries, as can be seen in Figure 3. We observe that queries get longer within a session, and shorter across sessions. When a re-finding query occurs within an hour of the previous query, it is 173% more likely that a word will be added to the query rather than a word being removed from the query. After an hour has elapsed, it is 106% more likely a word will be removed than added.

We hypothesize that the change in length reflects a fundamental difference between intra- and inter-session re-finding queries. For within session re-finding, people sometimes continue searching after a previous visit to the URL because the result does not initially appear to meet their need. When the same result is later returned for a longer, more targeted query, that can prompt a revisit to re-access the result's potential relevance. In contrast, across sessions users may be more likely to want to re-find a specific URL. In these cases, the shorter query reflects the user's ability to better express the target result based on information learned during previous interactions. In Section 5.5 we discuss these hypotheses and the evidence for them in greater detail.

Regardless of whether the re-finding query is longer or shorter than the previous query, it is very likely to substantially overlap with the previous query. In 52.8% of all re-finding instances with

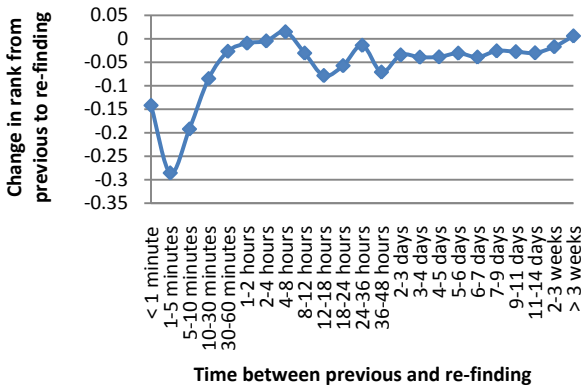

Figure 5. The change in position of the re-found result from the previous query to the re-finding query for substantially changed queries. The result almost always moves up in the result list.

substantial change, either the previous query is a proper subset of the re-finding query, or vice versa.

5.2.2 Re-Finding Queries More Common

We also explore how common the query used to re-find a result is. A URL can be found using a query that is the same as what most people would use to find it, or it can be found using a query that is not typically used to find that particular page. For example, the queries "free music" and "pandora" both return the result http://www.pandora.com, but people more commonly search for the site using the latter query. To measure how common a query is for a URL, we look at the set of all queries which result in a click by any user on the URL, and measure the percentage of time we observe the query in question in that set. The measure is query and result specific, but not user specific; a user may always click the result in question following the query, but if others search for the result using a different query, the query is not very common.

As with our earlier analysis, the commonness of the query used to re-find a result is a function of whether it is being used to re-find a result found in the same session or in a previous session. The difference between how common the previous query is and the re-finding query is can be seen in Figure 4, as a function of the time interval between the two queries. For intra-session re-findings, the re-finding query is 2% less common than the previous query, whereas for inter-session re-finding it is 0.5% more common.

5.2.3 Re-Found Results Rank Higher

When a re-finding query differs substantially from the corresponding previous query, we find the rank of the re-found result also differs. On average, the result is initially found via the previous query at rank 1.65 (i.e., it is the 1.65th result from the top of the list). When it is later re-found via a different re-finding query, it is ranked 1.57, or closer to the top of the list.

The change in position of the result between the previous query and the re-finding query as a function of the time interval can be seen in Figure 5. We observe that in 21 of our 23 time buckets (consisting of 95.8% of all instances of re-finding) the average position of the result during the previous query is further from the top of the result list than the position of the re-finding query. Again, we observe somewhat different behavior when re-finding happens within a session versus across sessions. For intra-session re-findings, the position is decreasing by 0.19 ranks, whereas inter-session re-finding is only decreasing by 0.03 ranks. It may be that the significant change in rank of a previously found result within a session inspires the searcher to return to the result to see if what they are looking for can indeed be found there.

Table 3. The percentage of time the query terms are present on the page for each time slice. Re-finding queries occur more often in the static content of the page.

	% of time slices query in page content				
Query Type	**100%**	**75-99%**	**50-76%**	**1-49%**	**0%**
Re-finding	85.8%	3.1%	0.5%	0.7%	9.9%
New-finding	78.3%	2.5%	1.0%	1.1%	17.0%

5.2.4 Re-Finding Queries More Related to Page Text

We also looked at how closely the query used to re-find a page matched the text content of the page, in order to understand how well the query reflected a consistent picture of the page. Our hypothesis was that while queries used to find content initially might reflect transient content on a Web page, query terms used for re-finding would reflect the static page content. Such queries would be more likely to consistently return the page in the result list, even as the page content is re-crawled by the search engine.

We measured how often re-finding and new-finding queries pointed to the static portion of the found result page using the percentage of time slices in the Web crawl which contained the given query words. For example, the query "times" might be in 100% of the crawled versions of the New York Times homepage, where as the query "obama" might be in only 80%, and the query "banana" in less than 1%.

We found, as expected, that re-finding queries were more likely than new-finding queries to refer to content that was consistently present in the page, and that new-finding queries were more likely to never actually appear in the page (see Table 3).

In this section, we have seen that when queries change substantially, they become shorter, more common, more closely and consistently tied to the page content, and rank the re-found result higher. Further, the queries exhibit different patterns of behavior depending on whether the re-finding query occurs in the same session as the previous query or in a different session. These differences are discussed in greater detail in Section 5.5.

5.3 Re-finding Converges

In addition to observing that re-finding queries tend to improve over their previous query pair, we also find that for commonly re-found results, searchers tend to converge quickly on a good query to use for re-finding and stick with that good query over time. In this section we discuss re-finding chains, or instances of multiple re-findings of the same URL by a given user.

Each of the measures discussed in the previous section contains different information about the query. A large change in how common the query is, but a small change in edit distance may be indicative of a typo, where as a large query change in commonness and edit distance may indicate learning of a better query. We combined the query length (Δ_{len}), result position (Δ_{pos}), commonness (Δ_{pop}) and static page content measures (Δ_{stat}) into a single change score, or *volatility value*, for a given query pair.

We use a weighted linear combination of the absolute difference of each of these measures to calculate overall volatility, as shown in Equation 1. The weights were chosen to normalize the four quantities. If one measure is high while others are not, the re-finding query pair will have relatively low overall volatility.

Table 4. Examples of re-finding query chains, broken down by how different the second query in the chain is from the first.

Query change	Re-found URL	Query chain used for re-finding
Great	http://groups.yahoo.com	yahoo, Yahoo Groups, Yahoo Groups, Yahoo Groups, Yahoo Groups
	http://whitepages.com	people search, white pages, white pages, white pages, white pages
	http://wachovia.com	wachovia.com, bank of yourself, bank of yourself, wachovia.com, wachovia
	http://monster.com	jobs, Monster Jobs, Monster Jobs, Monster Jobs, Monster Jobs
	http://www.pandora.com	free music, pandora, pandora, pandora, pandora
Some	http://www.cnn.com	CNN News, news, news, news, cnn
	http://webmessenger.msn.com	microsoft messenger, msnmessanger, msnmessanger, msnmessanger, msnmessanger
	http://usajobs.omp.gov	jobs, us jobs, us jobs, usa jobs, usa jobs
	http://www.ebay.com	ebay, My ebay, My ebay, ebay, ebay
	http://www.yahoo.com	Yahoo Messenger, yahoo.com, yahoo.com, yahoo.com, yahoo.com
Minimal	http://www.msnbc.msn.com	news, news, news, news, news
	http://www.autoscout24.de	auto scout24, outo scout24, outo scout24, OUTO SCOUT 24
	http://google.com	google.com, google.com, google.com, google.com, google.com
	http://www.zedge.net	free ringtones, freeringtones, freeringtones, free ringtones, free ringtones
	http://www.fedex.com	fed ex, fedex, fedex, fedex, fedex

While we note that this function as well as our selection of the weights is somewhat arbitrary in selection, our intention is to use it to find re-finding instances that are exhibiting more change as opposed to relatively stationary re-finding instances.

$$\text{vol}(q_1,q_2) = \lambda_1\Delta_{\text{len}}(q_1,q_2) + \lambda_2\Delta_{\text{pos}}(q_1,q_2) + \lambda_3\Delta_{\text{pop}}(q_1,q_2) + \lambda_4\Delta_{\text{stat}}(q1; q2) \quad [1]$$

We consider three groups of query change: *minimal* change, low-volatility substantial change (*some* change), and high-volatility substantial change (*great* change), based on the first two queries in each chain. Table 4 shows several examples of query chains of length 5 from each group. For the Minimal Change group, we randomly selected five chains whose first two queries exhibit minimal change. For the Great Change group we selected the five query chains whose first two queries exhibit some of the highest volatility. And for the Some Change group we selected five queries with some of the lowest volatility.

Query chains that begin with a great change appear to often be issued by users who start out seeking information for the first time; "people search" becomes "White pages", "jobs" becomes "Monster jobs" and "free music" becomes "Pandora." In contrast, query chains where the text initially changes substantially, but with low volatility, more often appear to reflect instances where the same query is merely expressed slightly differently. Many of these queries contain site specific words that provide little additional meaning, such as "my" in "my ebay" or synonyms such as "microsoft messenger" to "msnmessanger."

The queries used in re-finding chains appear to settle quickly. The conditional probability of the next instance of re-finding in a chain involving a minimal change between the previous and re-finding queries given that the current re-finding instance involves a minimal change is 98.2%, whereas the probability of a re-finding instance involving a minimal change is, in general, only 90.6%. The conditional probability that the next re-finding instance in a chain is a substantial change given that the current re-finding instance is substantial change is only 18.9%. This shows that re-finding queries are unlikely to transition from involving minimal change to involving substantial change. Further, 17.5% of the chains start with a substantial change, which is greater than the probability of a re-finding being a substantial change in general (9.4%). Substantial change re-finding instances are more likely to occur at the beginning of chains, and chains are more likely to end with minimal change re-findings.

5.4 Need Consistent across Queries

In addition to looking at how queries change and evolve, we also tried to get an idea of how the searcher used the re-found Web page. We did this by looking at the click trails of a user following a click from a re-found search results page. If the user were attempting to re-find previously viewed information reached by the re-found page, the click trail from the same search result to other pages outside the search engine is likely to also be the same, while if the user wanted to find new information on the re-found result, the re-finding trail may be different than the previous trail.

To explore the overlap in re-finding trails, we measured the percentage of time a given hop was the same across re-finding query trails with the same initial result click for a given user. For comparison, we computed a comparable value for the same URL using data collected from people who found it using a new-finding query. We looked at a number different types of hops, including: the second hop from the result page (the first hop being the click through to the result page), third hop from the result page, the first hop the user dwelled on for more than 30 seconds, and the final hop in the trail that started at the query result page.

As can be seen in Table 5, we find that trails are specific to users; the overlap between trails taken when a result is re-found is much higher than the overlap between trails taken from the same result by different users. When a person re-finds a result, they do the same thing more often than might be expected. Further, when the re-finding query is only minimally changed from the previous query, we find users are even more likely to follow the same path. The user tasks in these cases may be highly repetitive.

The trails people follow after a re-finding result click varies as a function of whether the re-finding occurs within the same session

Table 5: Percent of hops that are repeat hops in search trails following (a) re-finding for a given user, and (b) new-finding across users, given the first hop is the same.

(a) Trails following re-finding clicks

Change to query	Hop			
	Second	Third	Dwell	Final
Substantial	26.30%	21.27%	13.00%	18.44%
Minimal	43.93%	30.54%	21.07%	26.96%
All Re-finding	38.80%	26.87%	19.35%	19.67%

(b) Trails following new-finding clicks

Hop			
Second	Third	Dwell	Final
10.43%	3.97%	5.41%	5.01%

or within a different session. As shown in Table 6, users are at least as likely to follow a consistent path from a re-found result when it is re-found in the same session as when it is re-found in a different session. One reason for this could be that if a user intentionally wants to re-find to retrace a given trail, it is easier for the user to retrace previous steps within the same session. But all of our data taken together suggests the picture may be richer.

5.5 Session-Level Differs from Cross-Session

The analysis we have presented thus far suggests re-finding is very different when it occurs at the session-level as compared to across sessions. In Section 5.2 we saw that when re-finding occurred within a session, the re-finding query was more likely to be longer, less common, and rank the result much higher, where as when the re-finding occurred across session, the query was more likely to be shorter and more common. In Section 5.4 we saw that people were more likely to follow the same path when re-finding within a session than across a session. In this section we look at what all of these findings together tell us about how re-finding is being used at the inter- and intra-session level.

5.5.1 Intra-Session Re-Finding is Reevaluating

We hypothesize that some instances of session-level re-finding may involve the user returning to a previously found result that the user initially believed did not satisfy the user's information need, but that the user was willing to revisit to see if it now satisfies that user's need. The re-finding query within a session is typically longer, and the re-found result is typically ranked closer to the top of the list.

In contrast, we hypothesize that cross-session re-finding involves people trying to intentionally re-find the same result they have seen before as easily and directly as possible. The queries across session are less likely to change, and are likely to be short and common when they do, and rank the result somewhat higher. If the re-finding interval is at least a day, the amount of change to a cross-session re-finding query generally increases. We suspect this may reflect the user forgetting the previously used query terms, as well as changes to the search results and page content.

However, although we suspect results re-found across session are more likely to be actively sought out than results re-found within a session, we also suspect they are more likely to be visited to find new information. We saw in Table 6 that intra- and inter-session re-finding had almost the same percentage of second hop overlap;

Table 6. The percentage of hops that are the same following a re-finding query as they were following the associated previous query, broken down by whether the two queries occurred in the same session or not.

Session	Hop			
	Second	Third	Dwell	Final
Same	39.96%	33.43%	27.27%	24.95%
Different	41.76%	27.87%	17.14%	25.90%

however same session re-findings was much more likely to have the same third hop and to dwell on the same hop. In different sessions, the users may be looking for new content; for example, checking a news website and navigating to the sports page. Such repeat trails would likely have periodic but cross-session patterns. In these cases, it is also likely that the user would choose a new article to read after repeating the first step.

To better understand the validity of these hypotheses, we look to our user study. We observe that only 41.4% of the re-finding that occurs within the same session was labeled as intentional. This number is closer to what is typical of new-finding queries than is typical of re-finding queries. There are a number of examples of unintentional findings within a session. In some cases this happens after the user substantially changes their query (e.g., from "assembly programming" to "assembly tutorial"). In these cases, the user may have felt on first glance that that result did not adequately meet their need, but was more confident in the result when it appeared again for a different query. As we saw in our analysis of the log data (Section 5.2.3), in the user study the average change in result rank was the greatest in the first half hour. The large move up the result list may have influenced our participants' beliefs that the result would satisfy their information need. We also observe that there are multiple URL clicks to other pages in-between the intra-session re-finding instances. It may be that those other pages do not satisfy the information need, so the user chooses to return to one that might.

5.5.2 Inter-Session Re-Finding is Picking up a Task

When looking more closely at cross-session re-finding, we see some evidence to suggest users may sometimes be picking up a task they left off when re-finding across sessions. Figure 6 (a) shows the probability that a query will be a previous query while Figure 6 (b) shows the probability that a query will be a re-finding query, both as a function of its position in the session. The fact that the last query in a session is much more likely to be re-found in the future could indicate sessions often represent tasks that are not yet complete. Similarly, the initial query in a session is more likely to be a re-finding query than other queries in the session, and could indicate searchers may be picking up a task that was previously abandoned. When there are at least five queries in a session, the first query is over twice as likely to be a re-finding query than the last query in the session. Similarly, the last query in five-query sessions is 1.4 times more likely to be a previous query associated with future re-finding.

In the Figure 6 (a) we also see a slight increase in probability that the last query in the session is also a re-finding query, as well as an increase in the probability that the first query is the initial query in Figure 6 (b). This is because, as mentioned in the discussion of query chains, queries that are re-finding queries are also more likely to be previous queries for a later instance of re-finding than queries in general are.

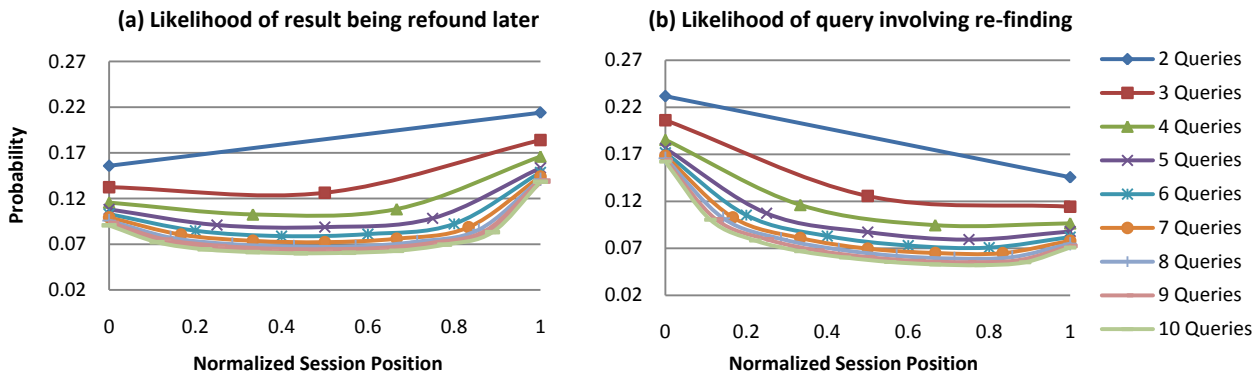

Figure 6. The probability that a query at the position in the session is (a) the previous query in a re-finding instance, or (b) the re-finding query in a re-finding instance. Each line shown represents cross-session re-finding probabilities for sessions with a given number of queries. Results found at the end of a previous session are more likely to be re-found, while results found at the beginning of a re-finding session are more likely have been seen before.

6. DESIGN IMPLICATIONS

There are a number of ways the rich understanding of re-finding developed in this paper can be used to improve the search experience. We discuss several in this section, including how we can best help the individual re-find and how re-finding behavior can be used to improve the search experience in general.

6.1 Helping the Individual

Perhaps the most obvious way that a search tool can improve the user experience given the prevalence of re-finding is for the tool to explicitly remember and expose that user's search history.

Our findings suggest that certain aspects of a person's history may be more useful to expose to the searcher than others. For example, results that are re-found often may be worth highlighting in some fashion. Similarly, almost a third of all URLs found via a one-click query are repeated, indicating these URLs comprise an important subset of the history. For multiple-click queries, the result that is clicked first following the query is more likely to be useful later, and thus should be emphasized, while results that are clicked in the middle may be worth forgetting from the search history entirely to reduce clutter. Results found at the end of a query session are more likely to be re-found, and thus may be worth emphasizing over results found earlier in a query session.

The query used to re-find a URL is often better than the query used to initially find it, and we believe the re-finding query may express how the person has come to understand this result. Re-finding queries should generally be emphasized in the history, and the previous query may even be worth forgetting to reduce clutter. In particular, as re-finding queries tend to converge, results that are frequently and recently found with a particular query may best be associated only with the recent query term.

Search history, or the important subset of the history, can be exposed on demand (e.g., via a history viewer) or in context as a user conducts a related search. A search tool may even find that URLs that are particularly likely to be re-found or their associated queries are worth exposing prior to a search, on the homepage or as part of the querying interface. When exposing previously found results, it is sometimes useful to label or name those results, particularly when those results are exposed as a set. Re-finding queries may make useful labels. A Web browser could even take these bookmark queries and make them into real bookmarks.

Search engines are particularly valuable because they use a searcher's queries to identify the right context to display related information. Identifying the right context for when to expose a result the user may want to re-find is important. Re-finding queries are more common and shorter than the previous-finding query. This suggests that if a previously found result is going to show up for a new query that is better by some measure, that result may be what the person is looking for. In some cases, even when the result is not going to be returned, a search engine may be able to identify that it should. If, for example, the user's current query is a substring of a previous query, the search engine may want to suggest the results from the history that were clicked for the longer query. In contrast, queries that overlap with but are longer than previous queries may be intended to find new results more than previously viewed results.

Session-level information can be useful for determining when to expose previously found results. At the beginning of a session, when people are more likely to be picking up a previous task, a search engine should provide access into history. In the middle of the session, it makes sense to focus on providing access to new information or new ways to explore previously viewed results. At the end of a session, a search engine may want to suggest storing any valuable information that has found for future use.

In some cases, the search engine can be very certain it knows that a person is trying to re-find a previously viewed result. For example, we saw that when a person issued the same query twice and clicked on the same result each time, a future identical search was highly predictive of a repeat click. In these cases, the search engine can treat the result specially and, for example, taking additional screen real estate to try to meet the user's information need with that result. The consistency in the trail taken following a re-finding result click means search engines could use this additional real estate to provide deep functionality like common paths and uses in the snippet. For results that are re-found across sessions, it may make sense instead to provide the user with deep links to new avenues within the result to explore.

6.2 Helping Everybody

An understanding of search engine re-finding behavior can further be used to improve the search experience in general. Currently search engines consider many metrics to try to identify the most relevant and high quality results. When a user re-finds a URL,

that is a statement of quality for the URL, both as associated with the previous query and the re-finding query, and a search engine can use this information to its advantage.

For example, results that are commonly re-found may be worth boosting in search results or crawling more often. Our findings regarding click order suggest the first results clicked following a query are most likely to be re-found, and thus be high quality, so search engines may want to weight the first and last clicks particularly strongly when considering click information in general for their associated queries.

Similarly, re-finding can tell us something about query quality, since re-finding queries tend to be higher quality than previous-queries. Common previous query and re-finding query pairs can be remembered by a search engine, and when the previous query is issued to the engine, it can suggest the re-finding query or use the re-finding query in some other way (e.g., for query re-writing). Common re-finding queries may also make good generic keywords or tags for their associated results in any situation where tags are useful.

7. CONCLUSION

In this paper, we have explored how search engines are used for re-finding previously found search results. We looked at when re-finding occurred and how re-finding queries differed from traditional queries. We explored the differences between queries that had substantial changes between the previous query and the re-finding query and those that had minimal changes. When the changes were substantial, we showed that re-finding queries tended to be better queries than the previous query used to reach a given URL. We have also observed that substantial changes were likely to appear early in the re-finding chain, and that re-finding often converged on a high quality query. We explored the differences between re-finding behavior as it occurred within the same session and across multiple sessions, and saw that cross-session re-finding may be a way of bridging a task between two different sessions. In the future, we look forward to applying these insights into building a better search experience.

ACKNOWLEDGEMENTS

The authors would like to thank Susan Dumais for her suggestions of interesting research directions, and Dan Liebling, Ken Church, and Peter Bailey for their assistance with data analysis. This research was funded in part by National Science Foundation IIS-0713111 and an NSF Fellowship. Any opinions, findings, conclusions or recommendations expressed in this paper are the author's, and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Abrams, D., Baecker, R., and Chignell, M. Information archiving with bookmarks: Personal Web space construction and organization. CHI '98, 1998, 41-48.

[2] Adar, E., J. Teevan, and S. T. Dumais. Large scale analysis of Web revisitation Patterns. CHI '08, 2008, 1197-1206.

[3] Aula, A., Jhaveri, N., and Käki, M. Information search and re-access strategies of experienced Web users. WWW '05, 2005, 583-592.

[4] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. and Frieder, O. Hourly analysis of a very large topically categorized Web query log. SIGIR '04, 2004, 321-328.

[5] Bruce, H., Jones, W. and Dumais, S. Keeping and re-finding information on the Web: What do people do and what do they need? ASIST '04, 2004, 1-10.

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

[7] Catledge,L. and Pitkow, J.. Characterizing browsing strategies in the World-Wide Web. WWW '95, 1995, 1065-1073.

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

[9] Dou, Z., Song, R., and Wen, J. R. A large-scale evaluation and analysis of personalized search strategies. WWW '07, 2007, 581-590.

[10] Jones, R. and Fain, D. C. Query word deletion prediction. SIGIR '03, 2003, 435-436.

[11] Komlodi, A., Soergel, D., and Marhionini, G. Search histories for user support in user interfaces. *JASIST*, 57 (6), 2006, 803-807.

[12] Lau, T. and Horvitz, E. Patterns of search: Analyzing and modeling Web query refinement. UM '99, 1999, 119-128.

[13] Ma, Z., Pant, *G., an*d Sheng, O. R. Interest-based personalized search. *TOIS*, 25 (1), 2007, 1-38.

[14] Morris, D., Ringel Morris, M., and Venolia, G. 2008. SearchBar: A search-centric Web history for task resumption and information re-finding. CHI '08, 2008, 1207-1216.

[15] Obendorf, H., Weinreich, H., Herder, E., and Mayer, M. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. CHI '07, 2007, 597-606.

[16] Raghavan, V. and Server, H. On the reuse of past optimal queries. SIGIR '95, 1995, 344-350.

[17] Richardson, M. 2008. Learning about the world through long-term query logs. *TWEB*, 2 (4), 2008, 1-27.

[18] Sanderson, M. and Dumais, S. Examining repetition in user search behavior. ECIR '07, 2007, 597-604.

[19] Tauscher, L. and Greenberg, S. How people revisit Web pages: Empirical findings and implications for the design of history systems. *IJHCS*, 47 (1), 1997, 97-137.

[20] Teevan, J. 2008. How people recall, recognize, and reuse search results. *TOIS*, 26 (4), 2008, 1-27.

[21] Teevan, J., Adar, E., Jones, R. and Potts, M. Information re-retrieval: Repeat queries in Yahoo's logs. SIGIR '07, 2007, 151-158.

[22] Teevan, J., Dumais, S. T., and Horvitz, E. 2005. Personalizing search via automated analysis of interests and activities. SIGIR '05, 2005, 449-456

[23] Teevan, J., Dumais, S. T., and Liebling, D. J. 2008. To personalize or not to personalize: Modeling queries with variation in user intent. SIGIR '08, 2008, 163-170

[24] Wang, P., Berry, M. W. and Yang, Y. Mining longitudinal Web queries: Trends and patterns. *JASIST*, 54 (8), 2003, 743-758.

[25] Wedig S. and Madani, O. A large-scale analysis of query logs for assessing personalization opportunities. KDD '06, 2006, 742-747.

[26] White, R. W., Bilenko, M. and Cucerzan, S. Studying the use of popular destinations to enhance Web search interaction. SIGIR '07, 2007, 159-166.