

Large Scale Analysis of Web Revisitation Patterns

Eytan Adar

University of Washington, CSE
101 Paul G. Allen Center, Seattle, WA 98195
eadar@cs.washington.edu

Jaime Teevan, Susan T. Dumais

Microsoft Research
One Microsoft Way, Redmond, WA 98052
{teevan,sdumais}@microsoft.com

ABSTRACT

Our work examines Web revisitation patterns. Everybody revisits Web pages, but their reasons for doing so can differ depending on the particular Web page, their topic of interest, and their intent. To characterize how people revisit Web content, we analyzed five weeks of Web interaction logs of over 612,000 users. We supplemented these findings by a survey intended to identify the intent behind the observed revisitation. Our analysis reveals four primary revisitation patterns, each with unique behavioral, content, and structural characteristics. Through our analysis we illustrate how understanding revisitation patterns can enable Web sites to provide improved navigation, Web browsers to predict users' destinations, and search engines to better support fast, fresh, and effective finding and re-finding.

Author Keywords

Query log analysis, Web behavior, revisitation, re-finding.

ACM Classification Keywords

H3.3. Information storage and retrieval: Information search and retrieval; H5.4: Hypertext/Hypermedia. – User issues.

INTRODUCTION

Revisiting Web pages is common, but an individual's reasons for returning to different pages can be diverse. For example, a person may revisit a shopping site's homepage every couple of weeks to check for sales. That same person may also revisit the site's listing of fantasy books many times in just a few minutes while trying to find a new book.

To better understand Web page revisitation, we analyzed the interaction logs of a large number of opt-in users over a period of five weeks. The logs contained a history of all of the Web page visits made by more than 612,000 users. Thanks to the large quantity of data, it was possible for us to explore the revisitation patterns for specific Web pages along a number of dimensions not previously feasible. We observed that even very similar pages can have very

different revisitation patterns. Our analysis suggests such differences can be due to a number of factors, including a person's intent, page content, and site structure.

The log data enables us to characterize the differences between revisitation patterns, but it is difficult to deduce revisitation intent without a more detailed survey or observational data. For example, someone may revisit a Web page frequently during a short interval of time because they are monitoring changing content contained in the page, or they may do so because they are using the site as a hub to navigate to linked pages. For this reason, we supplemented the log data with a survey in which we asked participants to describe their reasons for revisiting specific pages.

By combining revisitation patterns and survey data, we find four main patterns of revisitation that are characterized by various factors including usage, content, structure, and intent. This analysis can enable Web sites, Web browsers, and search engines to better support revisitation, and we discuss the design implications of our findings.

RELATED WORK

Early research aimed at understanding general Web navigation and surfing behavior first reported a large amount of re-access of information (e.g., [4]). To better understand and quantify this phenomenon, subsequent studies were designed to specifically address revisitation behavior. These studies come primarily in two main flavors (though some, like our own, mix elements):

- *Log studies*, where browsing patterns are monitored either through proxies or instrumented browsers ([6, 9, 14, 17, 23]).
- *Survey/interview studies*, in which a questionnaire or interview is constructed to understand specific behaviors ([1, 13, 21]).

Studies of revisitation have demonstrated that 50% [9, 23] to 80% [6] of all Web surfing behavior involves pages that users have previously visited. While many revisitations occur shortly after a page's first visit (e.g., during the same session using the back button), a significant number occur after a considerable amount of time has elapsed [17].

Two significant outcomes of earlier efforts have been the creation of taxonomies of browsing behavior in relationship to revisitation [14] and suggestions for improving the design of Web browsers [1, 17], search engines [1, 25, 26], and personal information management systems [10].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2008, April 5–10, 2008, Florence, Italy.

Copyright 2008 ACM 978-1-60558-011-1/08/04...\$5.00

Table 1. Summary of the data collected for analysis.

Class	Data type	Collection Method
Usage information		
Bin	# of unique visitors	Log analysis
	Time between visits	
	# of visits per visitor	
Patterns	Revisitation curve	Log analysis, clustering
Session	Previous URL visited	Log analysis of URLs visited prior to page
	Accessed via search?	
Self-reported intent		
Survey	Reason for revisitation	Survey, monitoring
Web page content		
URL	Length	Analysis of URL text
	Domain	
	Text substrings	
Content	Terms	Analysis of content
	ODP Category	SVM classifier
	Genre	Product classifier
Change	# of changes	Regular crawl
Structure	Outlinks	HTML parsing

Taxonomies in this space have sought to define revisitation in terms of user intent, goals, or strategies [16]. New browser designs have concentrated on better back buttons [8, 15], history or bookmarks [2, 11, 22], and monitoring and notification features [13]. Furthermore, several products and browser plug-ins have emerged to assist users in revisitation and monitoring activities (e.g. [5]).

The work we present here distinguishes itself from previous studies in a number of significant ways. First, to our knowledge, this is by far the largest study of Web revisitation behavior. While other studies have observed at most a few hundred participants in their population, the cohort in our work exceeds 612,000 people. Second, while other studies have pursued the longitudinal tracking of a small number of individuals in their complete, unbounded, interactions with the Web, we attempt to capture the interaction of hundreds of thousands of people with a specific set of Web pages. By broadly sampling pages and people we can begin to understand the characteristics of a page that are associated with specific revisitation behaviors.

A novel aspect of our methodology is that we are able to find diverse Web pages that are (re-)accessed in many different ways. With a small user population it is difficult to find Web pages that are not considered “popular” but are nonetheless visited sufficiently for statistical analysis. That is, the likelihood of two people in a small population both visiting anything but very popular pages is low. With the large user population in our study we are able to find multiple access patterns for all but the most unpopular pages. By understanding the variety of revisitation patterns we can provide design insights to better support the range of observed behaviors.

METHODOLOGY

To understand how and why a person revisits a Web page, we analyzed data from a variety of sources. The three main sources, *usage information*, *self-reported intent*, and *Web page content*, are summarized in Table 1. Below, we discuss how each type of information was collected.

Usage Information

We collected Web page visitation information by analyzing the logs collected from opt-in users of the Windows Live Toolbar. The toolbar provides augmented search features and reports anonymized usage behavior to a central server. Our analysis makes use of data from a sample of 612,000 users for a five week period starting August 1, 2006. The five week period is sufficient to capture a wide variety of revisitation patterns, although we may seasonal or yearly patterns. However, as part of a different study of data from May and June of 2007, we took the opportunity to validate our findings and identified the same general patterns described below.

User Selection

Users were identified by an anonymized ID associated with each toolbar. As is the case with large-scale log analyses, if more than one person uses the same computer, they will have the same ID. However, we believe that, with the exception of some very popular URLs, revisits will be user specific. To simplify the discussion we refer to a toolbar instance as a “user.”

For simplicity, we restricted our sample to data gathered from English speaking users in the United States. We also filtered to eliminate users for which there was limited data. Only users for which we had data for over 7 days were considered. In order to eliminate the “long-tail” of non-representative instances, users’ activities were characterized on a number of dimensions including: total number of URLs visited, unique URLs visited, number of URLs revisited, and number of URLs visited per day. The lowest ten percent of users in any of these dimensions were removed from the study population. Finally, users in the top one percent were also excluded (to eliminate robots and other badly behaving clients). The result of this filtering, for the time period of the study is over 612,000 valid users.

URL Selection

Because we were interested not only in pages with different popularity but pages with different revisitation patterns we defined a number of attributes for each page which we then used to systematically sample pages for further analysis. Specifically, we consider the number of unique visitors to a page (*unique-visitors*), the average and median inter-arrival (i.e. revisit) times for a page (*inter-arrival time*), and the average and median number of revisits per user for a page (*per-user revisits*). Per-user revisits for a page represents the median number of times a user will revisit that page.

Figure 1a shows a scatter plot of pages along these three dimensions, as well as histograms for each of the three criteria individually (log-scaled). Figure 1c highlights that

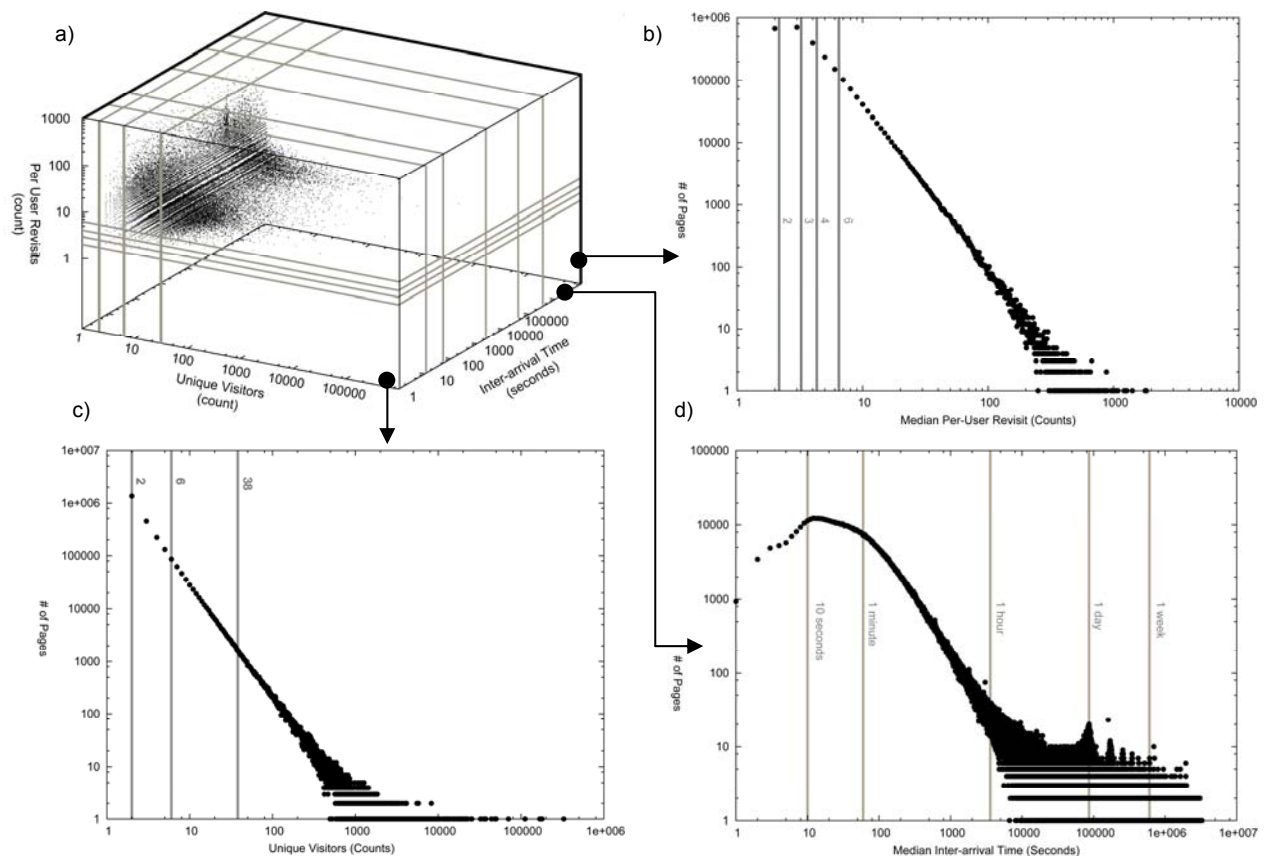


Figure 1a. A (log-scaled) scatter plot of pages (each point represents a single page) along the three dimensions of interest. 1b-d are the histograms for per-user visit, unique visitors, and inter-arrival time respectively (each point is a set of pages). The 55k pages were sampled from these distributions.

most pages were visited by very few people overall (the top leftmost point indicating that over one million pages only had one unique visitor), with a small number of revisits per person (Figure 1b), and at low (fast) inter-arrival times (Figure 1d). To sample broadly from this space without over-selecting from this “long-tail” we applied a pseudo-exponential binning to find thresholds for the unique-visitor (with a minimum of two) and per-user revisit criteria. The specific thresholds are noted as vertical lines in Figure 1(b-d). For the inter-arrival time criteria we opted to use thresholds that correspond to well understood time periods (e.g. 10 seconds, 1 minute, 1 hour, 1 day, 1 week).





As illustrated in Figure 1a, we defined four bins for the unique-visitor criteria, five for the per-user revisit criteria, and six for the inter-arrival time criteria (for a combination of $4 \times 5 \times 6 = 120$ possible bins). Each URL was annotated with these three bin numbers, and we sampled from this space. Some oversampling of popular pages was added by explicitly including the top 5000 most visited pages. We crawled the pages to ensure that they were still publicly available (in conformance with the robots.txt protocol), and removed from further analysis those that were not. The final sample included 54,788 URLs with an average of 468.3 (median 650) URLs in each of the 120 bins.

Self-Reported Intent

The interaction logs allow us to observe what people did, and led us to categorize revisitation in a way that emphasizes observable behaviors. However, the logs do not tell us why people behaved the way we observed. For this reason we conducted a complimentary user survey to gather information about user intent. In addition to helping us better understand revisitation patterns, the results of the user study also serve to relate our behavioral observations to previous taxonomies of revisitation that focus on intent (e.g., *monitoring* or *communication* [14]).

Twenty volunteers participated in the study (all daily WWW users; 7 were in their 20s, 9 in their 30s, and four were 40 or older; 19 were male). Participants were employees of Microsoft in a range of roles (developers, researchers, interns, program managers, etc.) Each participant installed software to log Web page visits, which they used for one to two months. We recorded visits only for the 55,000 URLs in our log study as well as a random subset from the user’s cache and Web history. At the end of the logging period, participants were asked to complete a survey to gather greater detail about ten of the Web pages they had revisited during the observation period. Of the survey responses, 49% (~80 URLs) overlapped with the 55,000 pages in the log study. The remaining URLs

Table 2. Relationship between visits and revisitation curves

Visits (time →)	Curve	Description
.....	→ 	Rapid repeat visits
I...I...I...I.....	→ 	Slower repeat visits
II.....II.....	→ 	Mix of fast and slow repeats
I..I.....I.....I..	→ 	Variable time between repeats

included a number of intranet pages which were used for additional qualitative analysis.

For each Web page in the survey, participants were asked whether they remembered visiting and revisiting the page. If they remembered the page, they were asked to indicate their intent when visiting from a list of options (e.g., to check for new information, to purchase, to communicate, etc.). If they recalled visiting the page more than once, they were further asked to describe how often they visited the page, and whether they visited it at regular intervals.

Web Page Content

The content of a Web page and its URL can also contain clues as to why the page was revisited. A custom crawler collected content information by downloading and indexing each Web page. Each page was crawled every hour for a month to examine the changes in page content over time. Although the times of the crawls do not correspond exactly with the revisitation observed in the logs, the crawled content is likely similar. We classified each Web page into topical categories using standard techniques [20]. The topical categories used are similar to those found in the first two levels of the Open Directory [18]. Additionally, a second classifier identified various generic genres including sites for children, mail, pornography, etc.

REVISITATION CURVES

Construction

To compare and evaluate revisitation behavior for different URLs we introduce the concept of a *revisitation curve*. A revisitation curve is a normalized histogram of inter-visit (i.e., revisit) times for all users visiting a specific Web page, and characterizes the page's revisitation pattern.

Table 2 illustrates the relationship between page visits and revisitation curves. For each row, four page visits are represented as four tall bars along a time line. The revisitation curve is a histogram of the inter-visit times. The x-axis represents the inter-visit time interval, and the y-axis represents a count of the number of visits to the page separated by that interval. The specific density of visits determines the shape of the revisitation curve. For example, the page in the first row shows four visits in rapid succession, and none at longer intervals, so the revisitation curve shows a high number of revisitations in the smallest

interval bin. In contrast, visits in the second row are spread out which shifts the peak of the revisitation to the right (corresponding to a higher inter-arrival time bin). Note that the number of visits in each example is the same. That is, the *same* number of visits per user can result in very *different* revisitation curves.

Revisitation curves are generated first by calculating all inter-arrival times between consecutive pairs of revisits. Deltas, in minutes, are computed for all users revisiting a specific URL ($\Delta_{\text{minutes}} = \text{floor}(\Delta_{\text{seconds}} / 60)$). Because binning by minute results in very sparse data for slow revisits, 16 exponential bins were used to smooth the histograms and to provide a finer grained resolution than the six bins used in URL sampling. Some manual tuning of the bin boundaries was used to generate more descriptive timescales. The boundaries were 1, 5.5, 11, 32, 64 (~hour), 98, 136, 212, 424, 848, 1696 (~day), 3392, 6784, 13568 (~week), 27136 (~2 weeks), and 55000 minutes (~month).

Because histograms are count based, URLs that had many more visitors or more revisits per visitor had higher counts. In order to compare revisitation patterns between URLs we normalize each individual curve by the centroid (i.e. average) of all curves. This centroid is depicted in Figure 2 (which also illustrates the effects of binning and bin location). This bottom curve shows the normalized histogram using the raw (un-binned) data. The periodicity in the figure corresponds to a 24 hour cycle and is likely due to the daily patterns of Internet usage. If a person has a “usual” time for Web surfing, this kind of periodic behavior will be seen in revisitation curves (similar to periodicities observed in re-finding behavior [19]). The top portion of the figure shows the smoother centroid curve that is obtained by summing over all points in each of the 16 inter-visit bins, which are shown as vertical lines.

To complete the normalization, for each URL the un-normalized bins in each revisitation curve were divided by the corresponding count in the centroid.

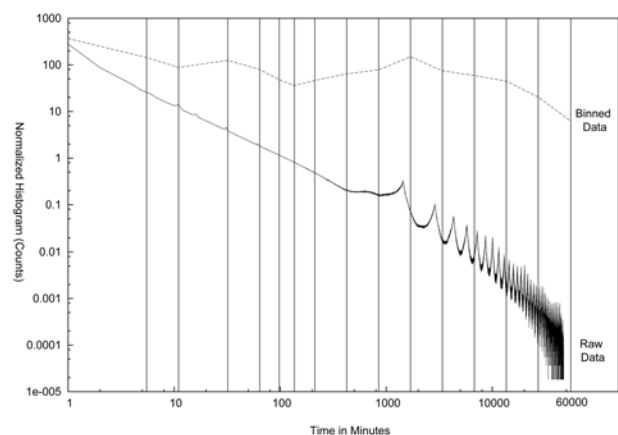




Figure 2. The centroid (average) revisitation curve for all pages in the 55,000 sample.

For every bin, i :

$$(\text{normalized}) \text{ revisit-curve}_{\text{url}}[i] = \text{count}_{\text{url}}[i] / \text{centroid}[i]$$

Roughly speaking, the normalized revisitation curve for each URL represents the percentage over, or under, revisits to that URL compared to the average revisit pattern. Though there are a number of other ways to normalize this type of data (normalize to 0-1 range, subtract out the centroid, etc.) we find that our mechanism works well in practice and allows us to simultaneously compare and group the different behavior patterns in our study.



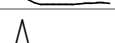
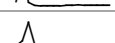
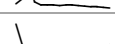


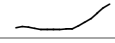




Examples of revisitation curves for two specific URLs are:

-  <http://www.amazon.com/> has a revisit curve that peaks towards the right, indicating that most revisits occur after a longer time period (over a day).
-  <http://www.cnn.com/> displays a peak on the left, perhaps driven by automatic reloads, along with a higher middle region, perhaps due to users checking for latest news.

Grouping

We consider each revisitation curve to be a signature of user behavior in accessing a given Web page. Given our representation of behavior it is natural to ask about the range of different curves. In order to group these curves we apply a clustering algorithm to the 55,000 curves to group those that have similar shape and magnitude. Specifically, we use repeated-bisection clustering [12] with a cosine similarity metric and the ratio of intra- to extra- cluster similarity as the objective function. In practice we find that clusters are fairly stable regardless of the specific clustering or similarity metric. By varying the number of clusters and testing within- and between-cluster similarity we find that the objective function levels off at around 12 clusters

Table 3. Cluster groups and descriptions.

Cluster Group	Name	Shape	Description
Fast Revisits (< hour) 23611 pages	F1		Pornography & Spam, Hub & Spoke, Shopping & Reference Web sites, Auto refresh, Fast monitoring
	F2		
	F3		
	F4		
	F5		
Medium (hour to day) 9421 pages	M1		Popular homepages, Communication, .edu domain, Browser homepages
	M2		
Slow Revisits (> day) 18422 pages	S1		Entry pages, Weekend activity, Search engines used for revisitation, Child-oriented content, Software updates
	S2		
	S3		
	S4		
Hybrid 3334 pages	H1		Popular but infrequently used, Entertainment & Hobbies, Combined Fast & Slow

(graphically represented in Table 3 and Figure 3), which we use in the analyses below.

As shown in Table 3, we further ordered, named, and manually grouped the clusters based on general trends into four groups: *fast*, *medium*, *slow*, and *hybrid*. Many revisitation patterns fell at the extremes. Five clusters represented primarily fast revisitation patterns, where people revisited the member Web pages many times over a short interval but rarely revisited over longer intervals. On the other hand, four clusters represented slow revisitation patterns, with people revisiting the member pages mostly at intervals of a week or more. Between these two extremes are two groups of clusters. One is a hybrid combination of fast and slow revisitations, and displays a bimodal revisitation pattern. The other type consists of two medium clusters comprised of URLs that are revisited primarily at intervals between an hour and a day. The clusters in this group are less peaked and show more variability in revisitation intervals than the fast or slow groups.

The self-reported revisitation reinforces the selection of our grouping criteria as patterns from the surveys were fairly consistent, not only with the participant's observed page interactions, but with patterns in the aggregate log data. Participants tended to report hourly or daily visits to pages that were clustered as fast or medium-term revisitation, and weekly, monthly or longer revisits those pages with slow revisitation patterns. The self-reported regularity of access decreased as the visitation interval increased. Participants reported visiting early and medium pages at regular intervals, and slow pages at irregular intervals.

In our discussion of the clusters, we focus primarily on large trends between these groups of clusters. Where appropriate, we also discuss details of the specific clusters.

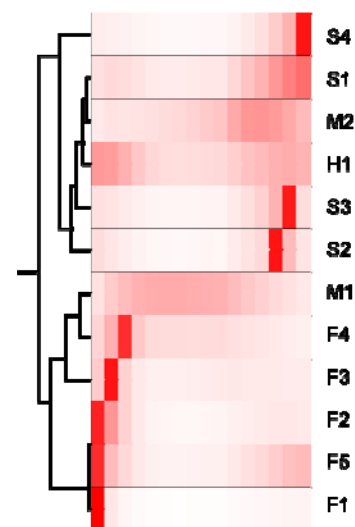


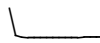
Figure 3. The hierarchical relationship of the different clusters plotted by Cluto [12]. Each row is a cluster, a column is a time bin, and darker colors represent a higher magnitude (imagine viewing a time series from above).

ANALYSIS OF REVISITATION PATTERNS

Using the properties described in Table 1, we characterized the pages that fell into each cluster group. Our findings were further supplemented by a manual examination of the Web pages closest to each cluster's centroid. A summary of our findings can be found in Table 4. Unless noted otherwise, all results are significant as measured using ANOVA/Kruskal-Wallis with post hoc tests, or χ^2 -tests as appropriate. In most cases, a two-tailed significance level of $p < .0017$ was used. This represents a .01 experiment-wise error rate using a Bonferroni correction for six dependent pair-wise tests. All differences in Table 4 are significant except for four pair-wise comparisons—the medium and hybrid groups do not differ in the unique visitors or URL length (chars); the fast and hybrid groups do not differ in % domain .com; and the slow and hybrid groups do not differ in % domain .edu. For tests involving terms or topics, χ^2 -tests were used and the significance levels of individual tests were corrected to preserve an experiment-wise error rate of .01.

Fast Group

Pages in the fast group tended to be revisited at intervals of less than an hour. Such sites elicit quick revisitation by forced reloads or shallow exploration and receive almost no long term revisitations. There are a number of reasons why Web pages may display a fast revisitation pattern.



Pornography & Spam: A main category in fast group, and in particular in cluster F1 , were spam or porn pages. Thirty eight percent of the URLs in F1 (4746 of 12474) were categorized as Porn (compared to the average cluster which contained 923 pages of porn representing 16.5% of pages in that cluster). Pages in the fast group were also likely to contain words like “xxx”, “sex”, “photo”, and “free” in the URLs and the text of the pages.

Hub & Spoke: Many of the pages in the fast group appeared to exhibit a hub-and-spoke revisitation pattern. For example, a person may start at a list of all products, such as a table of blouses, visit an individual product description pages and then rapidly return to the original page to explore more options. As evidence of this “shopping page” behavior, fast pages were more likely than pages in other groups to contain the words, “buy”, “catalog”, or “shop” in the URL, and to belong to shopping-related categories. The two most popular categories for this group (seen in Table 4) are *Home & Garden* and *Clothes & Accessories*, both of which are comprised mostly of shopping pages.

We hypothesized that visits to hub-and-spoke pages would be particularly likely to be preceded by a visit to a page in the same domain, and we found that this was indeed the case. Seventy seven percent of all revisits in the fast group were from the same domain. This is significantly more than the percent of visits from the same domain to pages in the medium (43.8%) or slow (56.5%) groups. Additionally, the fast group had the highest number of links on the page pointing back at pages in the same site (87%). The longer



URLs for pages in this group is also consistent with pages reached by internal site navigation.

Shopping & Reference Web sites: We found that not only specific Web pages but entire *Web sites* may display an over-representation of URLs in the fast group. These include a number of shopping/classifieds Web sites (e.g. Craigslist and Kelly's Blue Book) as well as reference pages. For example, pages belonging to Wikipedia and the Internet Movie Database (IMDB) appeared more often in this group (though the homepages themselves were in the slow group). We may hypothesize that certain information needs, such as specific encyclopedia details, are localized temporally. User satisfying such needs may revisit frequently for the current information need, but do not generally return again for an extended period of time.

Auto refresh: Another reason for frequent fast revisits was that pages were auto-refreshing. For example, CNN's homepage causes a browser refresh every 30 minutes, Fox New's homepage every 10, Bloomberg every 5, and Sports Illustrated refreshes game scores every 60 seconds. In particular, pages in clusters F3  and F4 , with peaks at five and ten minutes respectively, may be pages where much of the revisitation occurs from auto-refresh. Both clusters display an over-representation of the terms “meta refresh” and “window.location” corresponding to the HTML and Javascript reload mechanisms. Further supporting this hypothesis is the fact that the visited page immediately preceding pages in these clusters are exactly the same 29% of the time, compared with 12% of the time for pages in the medium group and 10% of the time for pages in the slow group.

Fast monitoring: From our user survey we found that monitoring applications in this domain tended to be for quickly changing information that required attention for a short period. For example, a page with local, up-to-the-minute traffic conditions was one such page.

Medium Group

Pages that fell in the medium-term revisitation group were visited hourly (M1 ) or daily (M2 ). Pages revisited hourly (M1) appeared to be portals, Web mail, and chat, as evidenced by keywords in the URL and topic classification. Pages in M2 were bank pages (bank and credit union URLs were very common), news pages, and sports pages.

Popular homepages: Pages in the medium group were among the most easily recognizable, with many being very popular. The URLs were significantly shorter at a mean of 38.3 characters (versus 48.5 and 43.2 for fast and slow respectively) and a mean directory depth of 2.4 (versus 3.2 and 2.7) indicating that they were more likely to be top level homepages. There were many more unique visitors to these Web sites than to others, with a mean number of 254.3 visitors, compared with 40 for the fast group and 75 for the slow group. Interestingly, however, the hybrid

Table 4. Properties of the clusters, based on analysis of the cluster URLs, content, usage, sessions, and changes.

Property type		Fast (< hourly)	Medium (daily)	Slow (> daily)	Hybrid (<hourly, >daily)
<i>Usage information</i>					
Bins	Unique visitors	40.1	254.3	74.9	274.8
	Visits/visitor	5.7	7.4	4.0	4.9
Session	Previous URL is the same	28.6%	6.8%	7.3%	10.6%
	Previous URL same domain	77.0%	43.8%	56.5%	65.2%
	Accessed via a search	2.9%	4.0%	4.3%	3.4%
<i>Self-reported intent</i>					
Survey	Revisitation reason	Buy something, monitor live content (e.g., sports scores or stock)	Communicate, listen to music, view videos, search, play games	Interact with personal data, view previously viewed information	Visit new content or follow new links, buy something
<i>Web page content</i>					
URL	Length (chars)	48.5	38.3	43.2	38.8
	Length (pieces)	3.2	2.4	2.7	2.6
	Domain .com	77.9%	72.3%	75.5%	77.8%
	Domain .edu	1.1%	4.4%	1.7%	1.6%
	Characteristic substrings	buy, shop, photo	mail, bank	money, weather	game, music
Content	Characteristic terms	price, gallery, auction	news, search, home	pictures, movie, table	change, updated, songs
	Distinguishing topics	House & Garden Clothes & Accessories Porn	Finance & Investment World Cultures Astrology & Psychic	Business & Finance Soccer Movies	Video Games Music Internet and the Web
Change	Number of changes	222.0	315.2	283.3	344.6
Structure	Outlinks (total/unique)	74/58	86/69	85/69	112/92
	% Outlinks to same site	75.9%	67.1%	72.5%	71.2%

group displayed a similar number of unique visitors – a fact we return to in later discussion.

Consistent with the notion that popular sites appear in the medium group, in this group we find pages from search engines (e.g. www.google.com and search.yahoo.com) as well as the homepages for BBC's news site (news.bbc.co.uk) and Google news (news.google.com). Additionally, we find medium rate revisits for URLs on games Web sites such as Yahoo Games (games.yahoo.com) and Gamesville (www.gamesville.com).

Communication (Web mail and forums): Mail and forum pages are over-represented in this group. This is likely due to the timescale of human-to-human communication. That is, the turnaround in communication between two individuals through a Web mediated system appears to be more on the hourly or daily basis than significantly faster or slower. The width of the revisitation curve for this group indicates that there is considerable variability (both across and within individuals) in revisitation intervals. Survey respondents further indicated that “communication” was a reason to access a number of Web pages in this group.

.edu domain: The medium group also contained a significantly higher number of educational domain Web pages, with 35% (440 of 1245 .edu pages in the study) of the URLs coming from the .edu domain (versus 28%, 31%, and 4% for the fast, slow, and hybrid groups respectively). The high representation of educational pages may reflect the fact that student populations are particularly likely to re-access Web content at daily or weekly periods.

Browser homepages: According to the survey data, most browser homepages fell in this group, with none falling in the fast group, and few in the slow group. This is consistent with the fact that portals and search engines, both frequently used as browser homepages, appear in the medium group.

Slow Group

Pages with slow revisitation patterns were visited on average at intervals longer than a day. These pages most likely represented people's long term interests and infrequent information needs.

Entry pages: Many pages appeared to be product home pages or the entry pages for specialty search engines (e.g. for travel, jobs, or cars). These pages were not used in a hub-and-spoke manner like the pages in the fast group, but were perhaps used to get to such pages.

Weekend activity: Cluster S2 exhibits a peak at a week. We hypothesize much of the activity in this cluster is related to weekend activity. For example, the page <http://www.fetenet.com>, which is close to the cluster's centroid, contains event listings. The Movies category is the most distinguishing category for this cluster, followed by the Religion category, both of which further suggest weekend activity.

Search-engines used for revisitation: One trend we noted across groups was that as the revisitation period increased, the pages were more likely to be accessed from a search

engine. The URLs are also longer in the slow group than the medium group, which suggests that direct access typing the URL was less likely for these pages.

Child-oriented content: Interestingly, many pages for children's Web sites are over-represented in the slow group. This includes pages for the Cartoon Network (www.cartoonnetwork.com), the Barbie Web site (barbie.everythinggirl.com), Nickelodeon (www.nick.com) and Disney (disney.go.com). This finding may be due to limited availability of Internet access for children who may only have access at certain times of the day or days of the week. Additionally, many Web sites are tied to TV programs which have weekly periodicities. Children's Web sites also display hybrid behavior which appears to be due to multiple functions of the Web site, which we describe in more detail in the next section.

Software updates: Our survey data also revealed pages in the slow group were used to download software updates. For examples, pages with "update" in the URL or content included application update sites such as Microsoft Windows Update and McAfee Update.

Hybrid Group

URLs in the hybrid group bore similarities to several other groups, as can be seen in Table 4. Much of the navigation behavior was similar to what was observed for the fast group (e.g., hub-and-spoke), but the hybrid pages appeared to be of higher quality and receive more long-term revisitations. Like the medium group, hybrid pages had short URLs and a large number of unique visitors (274).

Popular, but infrequently used, site homepages: It is interesting that the most popular Web pages, in terms of unique visitors, fall into two different behavior patterns—medium and hybrid groups. Whereas the pages occurring in medium group indicate a fairly constant need, hybrid interactions reflect a rarer need that nonetheless requires many page revisits for the individual need to be met. For example, a page such as Automart's homepage (www.automart.com) provides a front end interface to search for new sales listings, in this case for cars. A user may visit the page, search for updated or new listings and check many listings in any given visit. The homepages for various classifieds listing service, such as local Craig's List homepages, fall into this group as well. Both hybrid and medium group's pages display a higher than expected number of "login[s]" in the text indicating both types are transactional or personalized in some way.

According to the survey participants, the amount of meaningful change they expected to the page decreased as the revisitation interval increased. Visits to hybrid pages were particularly likely to be to find new information (on a page that had previously been visited). Hybrid pages also displayed the greatest number of changes in content. A hybrid behavior might be a mechanism for "catching-up" with changed content. Pages in the medium group

displayed the second highest number of changes and indicate a potentially more regular monitoring activity.

Entertainment & Hobbies: Hybrid behavior is likely observed in situations when an activity requires hub-and-spoke movement around the page, but the activity itself is somewhat infrequent. For example certain shopping related pages such as shopping carts exist in this group. High-level auction pages are also in this group (e.g. computers.ebay.com, home.ebay.com, etc.). The difference between auction related pages, such as eBay, and other shopping destinations might be the requirement of monitoring ever-changing auctions, successive bids, and long-term interest in collectibles.

Another over-represented category within the hybrid group are "games" pages (as indicated by the page text and by the page categories). It is interesting—though perhaps unsurprising—that games are played rarely but repeatedly.

Combined Fast & Slow: Unlike the medium group, the hybrid group received significantly fewer mean visits per visitor (4.9 v. 7.4). This is consistent with the observation that most URLs in the medium group are attached to constant, more generic, information needs (e.g. communication, news, etc.) whereas page in the hybrid group represent more easily satisfied needs (e.g. games, music, or product purchase) and are more likely a combination of slow (rare information need) and fast (high "session" revisits).

Web Sites Across Groups

Because of large scale sampling of URL data we are able to also look at how pages within a Web site are distributed across groups. This analysis generates confirmation of the findings described above.

As mentioned earlier, there is an over-representation of URLs for shopping and reference Web sites in the fast group (e.g., Craigslist's pages from Seattle and New York, seattle.craigslist.com and newyork.craigslist.com, Ikea at www.ikea.com, Kelly's Blue Book, for pricing used cars, www.kbb.com, and Wikipedia at www.wikipedia.org). An interesting characteristic of such sites is that while a vast majority of the pages *inside* the website were in the fast group, their *top-level* homepages (e.g., <http://www.ikea.com/> and <http://www.wikipedia.org/>) fell into the slow group. The homepage visits probably represent the occasional need people have for the type of information (Ikea furniture, encyclopedic references, or used car pricing), while the deeper pages that fall in the fast group probably represent individual's behavior when satisfying the need. This suggests that while some *Web pages* themselves may fall into fast or slow groups, the *Web sites* may display hybrid or slow behavior.

In general, we find that Web sites that serve a single consistent role tend to be present in only one group. For example, BlinkYou, a predominantly fast site at www.blinkyou.com, provides widgets to embed in MySpace homepages and MSN Weather, a generally slow

site, serves weather at weather.msn.com. Those that serve multiple roles have URLs that fall more equally into all groups. For example, EsMas, www.esmas.com, a Spanish language portal and media site that provides shopping, classifieds, news, galleries, and many other features has a fairly even split across groups (with no significant overrepresentation in any group).

DESIGN IMPLICATIONS

Web Browser Design

Previous studies of the revisitation patterns of individual users have led to numerous browser enhancements [2, 11, 13, 15, 22]. In one study, Cockburn et al. [7] explored presenting a user's history as a series of Web page thumbnails enhanced with information about the user's page visit frequency, and allowed the user to group thumbnails to visualize "hub-and-spoke" revisitation patterns. Given the results presented here, we believe there may be value in providing awareness of, and grouping by, a broader range of revisitation patterns. For example, users may want to quickly sort previously visited pages into groups corresponding to a *working* stack (recently accessed fast pages), a *frequent* stack (medium and hybrid pages), and a *searchable* stack (slow pages).

Several history mechanisms have tried to predict whether people will access a Web page based on overall measures of last access time and frequency of access. It may be possible to better predict the future utility of a URL by taking into account the page's revisitation pattern. For example, if a page is visited weekly and the user has not visited it in almost seven days, it is very likely to be revisited soon. On the other hand, a page the user visits daily but that was just visited is unlikely to be visited in the immediate future.

More accurate predictions of the future utility of a Web page can be used in many ways. The color of the page's URL could change to indicate increased potential interest, or the address bar's completion drop-down list could favor URLs that are likely to be accessed. Alternatively, such pages could be proactively recommended and the content pre-fetched [3] for efficiency in browsing or indexing.

In order for a Web browser to take advantage of revisitation patterns for a page or site, aggregate visitation data could be made available by the site itself. If this is not feasible, it may be possible to predict the page's revisitation pattern by classifying it using non-behavioral features like text and link structure. Initial explorations indicate this approach is promising. Personalized revisitation patterns could further be combined with aggregate observations or predictions.

Search Engine Implications

Revisitation analysis also has a number of implications for search engine design and in particular to the related issue of re-finding. Prior research has demonstrated re-finding behavior is prevalent [19, 25] in search engine use. Our analysis, which further demonstrates a relationship between

search and specific kinds of revisitation behavior, suggests several ways search engines can support re-finding.

For repeat searches, there is a tension between displaying new results and preserving previously viewed results [24]. It may be particularly important for search engines to provide consistent results for queries that return many slow and hybrid pages, since these searches are likely to be used to re-find pages found a long time ago. Though consistency can be achieved by keeping the results displayed static, it may be better achieved through personalization where the results an individual is likely to revisit are emphasized.

The revisitation patterns of pages returned for a query can tell the search engine something about the searcher's information need. Depending on the particular task or distribution of results, the search engine could include pages that had consistent or diverse revisitation patterns.

The types of pages a person is interested in may also suggest how receptive that person is to new information (e.g. suggestions of related content or advertisements). For example, if a query returns results that are generally in the fast group, this could indicate that the user is looking for something new and may be particularly responsive to the suggestion of relevant content. On the other hand, if the results are primarily in the medium or slow revisit groups, the user may be more likely to have a specific intent and not respond to suggestions. Content that is somewhat orthogonal to the user's objective may be most helpful in these cases by appealing to different interests.

Search engines may also benefit by taking into account users revisitation behaviors when crawling the Web. Assuming changes to page content are correlated with people's aggregate revisitation patterns, the revisitation patterns could indicate the optimal rate for re-crawling and re-indexing different Web sites.

Web Site Design

Just as search engines may want to make it easy to find slow and hybrid pages, Web site designers may want to ensure that such pages are easy to find. This could be done by creating appropriate keyword lists for indexing, minimizing page change and movement, and providing navigational shortcuts from common search landing pages.

Modeling users in terms of revisitation behavior might allow a Web site designer to apply our results in the design of their site. By taking into account the four potential revisitation styles, in conjunction with the site design and content, information architects can understand and simulate user behavior with their Web sites. This may make it possible for designers to understand potential interaction patterns even before the Web site is launched, and to evaluate a site's success in meeting the design objectives.

Finally, we believe there are opportunities for Web site designers to better support monitoring activities. For example, some Web sites create a list of "what's new" to aid in repeat visitors' discovery of new content. Our

research shows that different pages, even on the same site, can have different revisitation patterns. As a result, it may be in the site's interest to maintain "what's new" lists at different granularities depending on the specific page.

CONCLUSION

Our paper has presented an exploration of the diverse ways that people revisit Web pages and the reasons behind their actions. Our work, which is based on the largest study to date of Web visitation logs coupled with user surveys and Web content analysis, has allowed us to develop a unique view of people's revisitation behaviors. By analyzing tens of thousands of Web sites we have been able to identify 12 different types of revisitation behavior corresponding to four groups that are orthogonal to previous work. Our analysis of these groups in various contexts has led us to define a set of design implications for client applications, Websites, and search engines.

Looking forward, we hope to refine and test the designs we have described above, and better characterize the impact of document change on revisitation. As can be seen in Table 4, the amount of change to a Web page's content that we observed varied greatly as a function of the group to which the page belonged. We believe that revisitation patterns and intent are likely to correlate with meaningful change. For example, hybrid pages were particularly likely to change, and survey participants indicated they were particularly interested in new content from these pages. We also believe that a study over longer periods may lead to interesting new insights about seasonal effects and periodicities.

ACKNOWLEDGEMENTS

The authors would like to thank Dan Liebling, Ronnie Chaiken, and Bill Ramsey for their data analysis help.

REFERENCES

1. Aula, A., N. Jhaveri, and M. Käkik. Information search and re-access strategies of experienced Web users. In *Proceedings of WWW '05*, 2005.
2. Ayers, E. and J. Stasko. Using graphic history in browsing the World Wide Web. In *Proceedings of WWW '95*, 1995.
3. Bhide, M., P. Deolasee, A. Katkar, A. Panchbudhe, K. Ramamritham, and P. Shenoy. Adaptive Push-Pull: Disseminating dynamic Web data. *IEEE Trans. Comput.* 51(6):652-668, 2002.
4. Catledge, L.D. and J.E. Pitkow. Characterizing browsing strategies in the World-Wide Web. In *Proceedings of WWW '95*, 1995.
5. ChangeDetect. www.changedetect.com, retrieved Jan. 2008.
6. Cockburn, A. and B. McKenzie. What do Web users do? An empirical analysis of Web use. *Int. J. of Human-Computer Studies*, 54(6): 903-922, 2001.
7. Cockburn, A. and S. Greenberg. Issues of Page Representation and Organisation in Web Browser-Revisitation Tools. *Australian J. of Info. Systems*, 7(2):120-127, 2000.
8. Greenberg, S. and A. Cockburn. Getting back to back: Alternate behaviors for a Web browser's back button. In *Proceedings of the 5th Annual Human Factors and the Web Conference*, 1999.
9. Herder, E. Characterizations of user Web revisit behavior. In *Proceedings of Workshop on Adaptivity and User Modeling in Interactive Systems*, 2005.
10. Jones, W., S. Dumais, and H. Bruce. Once found, what then?: A study of 'keeping' behaviors in the personal use of Web information. In *Proceedings of ASIST '02*, 2002.
11. Kaasten, S. and S. Greenberg. Designing an integrated bookmark / history system for Web browsing. Western Computer Graphics Symposium, 2000.
12. Karypis, G. Cluto — a clustering toolkit, www.cs.umn.edu/~cluto, retrieved Jan. 2008.
13. Kellar, M., C. Watters, and K. M. Inkpen. An exploration of Web-based monitoring: Implications for design. In *Proceedings of CHI '07*, 2007.
14. Kellar, M., Watters, C., and Shepherd, M. A goal-based classification of Web information tasks. In *Proceedings of ASIST '06*, 2006.
15. Milic-Frayling, N., Jones, R., Rodden, K., Smyth, G., Blackwell, A., and Sommerer, R. Smartback: Supporting users in back navigation. In *Proceedings of WWW '04*, 2004.
16. Morrison, J. B., P. Pirulli, and S. K. Card. A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. In *Proceedings of CHI '01*, 2001.
17. Obendorf, Hartmut, H. Weinreich, E. Herder, and M. Mayer. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proceedings of CHI '07*, 2007.
18. Open Directory Project, www.dmoz.org, retrieved Jan. 2008.
19. Sanderson, M. and S. T. Dumais. Examining repetition in user search behavior. In *Proceedings of ECIR '07*, 2007.
20. Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1-47, 2002.
21. Sellen, A. J., R. Murphy, and K.L. Shaw. How knowledge workers use the Web. In *Proceedings of CHI '02*, 2002.
22. Takano, H. and T. Winograd. Dynamic bookmarks for the WWW. In *Proceedings of Hypertext '98*, 1998.
23. Tauscher, L. and S. Greenberg. How people revisit Web pages: Empirical findings and implications for the design of history systems. *Int. J. of Human-Computer Studies*, 47(1):97-137, 1997.
24. Teevan, J. The Re:Search Engine: Simultaneous support for finding and re-finding. In *Proceedings of UIST '07*, 2007.
25. Teevan, J., E. Adar, R. Jones, and M. A. Potts. Information re-retrieval: repeat queries in Yahoo's logs. In *Proceedings of SIGIR '07*, 2007.
26. White, R. W. and S. M. Drucker. Investigating behavioral variability in Web search. In *Proceedings of WWW '07*, 2007.