Resonance on the Web: Web Dynamics and Revisitation Patterns

Eytan Adar
University of Washington
Seattle, WA, USA
eadar@washington.edu

Jaime Teevan, Susan T. Dumais
Microsoft Research
Redmond, WA, USA
{teevan,sdumais}@microsoft.com

ABSTRACT
The Web is a dynamic, ever-changing collection of information accessed in a dynamic way. This paper explores the relationship between Web page content change (obtained from an hourly crawl of over 40K pages) and people's revisitation to those pages (collected via a large scale log analysis of 2.3M users). We identify the relationship, or *resonance*, between revisitation behavior and the amount and type of changes on those pages. By coupling our large scale log analysis with a complementary user study we explore the intent behind the revisitation behavior we observed. Using the notion of resonance to identify the likely content of interest, we describe a number of ways interaction with changing and revisited information can be better supported. We illustrate how understanding the association between change and revisitation might improve browser, crawler, and search engine design, and present a specific example of how knowledge of both can enable relevant content to be highlighted.

Categories and Subject Descriptors
H.5.4 [Information Systems]: Information Interfaces and Presentation (e.g., HCI) – Hypertext/Hypermedia: User issues.

General Terms: Human factors, measurement.

Keywords: Web log analysis, Web page dynamics, change, Web behavior, revisitation, re-finding, resonance.

1. INTRODUCTION
Revisiting Web pages is common [8], but people's reasons for revisiting can be diverse [1][18]. For example, a person may return to the SIGCHI website, pictured in Figure 1, to be reminded of the group's officers, or to catch up on the latest news. While most content on the Web does not change, pages that are revisited change frequently [2]. As revisitation is often motivated by monitoring intent [17], some relationship between change and revisitation is to be expected. However, the subtleties of change and revisitation behavior mean the two do not necessarily have a direct correlation. For example, while new content may be added to a blog hourly, individual posts may move slowly down the page and not vanish until 24 hours later when the daily content is archived. Although the page changes hourly, a user monitoring it may only need to revisit daily to catch up on the content.

Pages are often composed of many sub-pieces, each changing at different rates, illustrated in Figure 1. Advertisements on a page may change on every visit, while navigation content may almost never change. A particular user's revisitation rate can sometimes indicate the part of the page that user is interested in. Someone who returns to the SIGCHI page at a frequency similar news updates may be interested in catching up with the latest news, while someone who returns after a long interval may be interested in revisiting the page's static or navigational content (e.g., "membership"). Aggregate revisitation behavior of *many* users can indicate the most common purpose of that page.

Web content change can be beneficial to the Web user looking for new information, but can also interfere with the re-finding of previously viewed content [31]. Understanding what a person is interested in when revisiting a page can enable us to build systems that better satisfy those interests by, for example, highlighting changed content when the change is interesting, actively monitoring changes of particular interest, and providing cached information when changes interfere with re-finding.

In this paper, we characterize the relationship between revisitation and change by analyzing a large scale Web log trace for 2.3M users and a five week hourly crawl of over 40K Web pages. We begin our analysis by exploring how *revisitation behavior* relates to *change*. The analysis validates a number of hypotheses (e.g., popularity and change are correlated), but also uncovers some surprising results. For example, we find that certain measures of page change (e.g., the *amount* of change) are not linearly related to measures of revisitation (e.g., inter-arrival times)—a result with consequences to monitoring tools that use these simple measures to watch for page changes to identify events of likely importance.

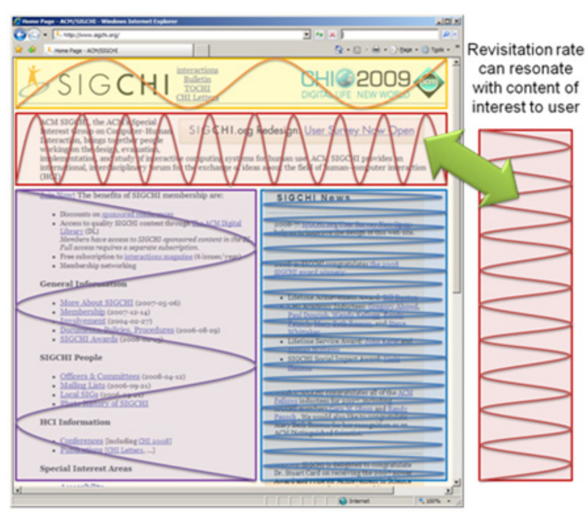

Figure 1. Web pages—and their components—change at different rates. People revisit pages for many reasons, and their reasons for revisiting can be affected by content changes. In this paper we show that revisitation rates can resonate with the rate of change of interesting content.

We develop a deeper insight into when and how changing content is important by combining more refined measures of change and revisitation. An evolutionary view of the page, represented by a two-segment *change curve*, in concert with the distribution of revisitation intervals (*revisitation curve*), allows us to identify the type of information targeted by revisitation: *static* (previously viewed and unchanged), or *dynamic* (newly available). These results are supported by a smaller user study used to find the underlying intent and expectations of change of individual users.

Using peak revisitation for a given page, in conjunction with fine-grained changes, we can also identify the interesting *content* on pages. We introduce an automated algorithm for DOM-level analysis that uses the *resonance* between revisitation and change to break pages into components that are more-, and less-, likely to be of interest. Such analysis has previously been impossible without the use of expensive surveys or other un-scalable research instruments such as eye-trackers.

We conclude with a discussion of how understanding the association between change and revisitation might improve browser, crawler, and search engine design, focusing on an example application that automatically highlights potentially interesting portions of a page identified according to our analysis.

2. RELATED WORK

Prior work has focused on either the study of change or the study of revisitation, but rarely both together. By tracking change and revisitation behavior concurrently, on a very large scale and with very fine granularity, we are able to offer a novel perspective on content change in revisited pages, revisitation patterns for dynamic and static pages, and the relationship between the two.

2.1 Changes in Content over Time

Characterizing the amount of change on the Web has been of considerable interest to researchers ([2], [7], [10], [19], [20], [21], [24], [27]). Numerous previous studies (e.g., [10], [24]) have found that Web content change occurs relatively infrequently, and identified trends in the change that does occur. For example, Fetterly et al. [10] found that past change was a good predictor of future change, that page length was correlated with change, and that the top-level domain of a page was correlated with change with edu pages changing more slowly than com pages. Koehler [20] found that page change levels off as a page ages. These studies have provided insights into crawler and search engine design, but have generally ignored actual page use.

In previous work [2] we explored how content changed in pages that we knew had been visited. We found that revisited pages change more often than pages selected using other sampling techniques, and explore the relationship in greater detail here.

2.2 Revisitation Patterns over Time

Research on revisitation stems from early Web navigation studies which reported a large amount of re-access of information (e.g., [5]). Subsequent studies were designed to specifically address revisitation behavior. These studies come primarily in two main flavors (though some mix elements [1]): Log studies, where browsing patterns are monitored either through proxies or instrumented browsers ([8], [14], [18], [26], [30]); and survey/interview studies, in which a questionnaire or interview is constructed to understand specific behaviors ([3], [17], [28]). These studies have found that 50% ([14], [30]) to 80% [8] of all Web surfing behavior involves previously visited pages. Revisitation taxonomies ([18],[23]) have sought to define revisitation in terms of user intent, goals, or strategies and we make use of these in our discussion.

Studies in this area have typically been small-scale and concentrated on tracking behavior by users rather than website. In previous work [1] we were able to gather enough data for particular Web pages to understand how those pages by studying a very large population. We found that pages with certain revisitation patterns were more likely to change, and those initial findings motivate the deeper analysis presented here.

Understanding revisitation has contributed to new browser designs ([3], [4], [11], [16], [22], [26], [29]), monitoring and notification features ([6], [17]), search engine design ([3], [31], [32]), and personal information management systems [15].

2.3 Relating Revisitation and Change

The analysis presented in this paper explores how revisitation patterns relate to changes in content over time. Douglis et al. [9] studied how the number of visitors to a page relates to change, finding that frequency of change was higher with increasing access. Pitkow and Pirolli [27] looked at a different facet of Web evolution, in particular they found that "desirable" pages were less likely to disappear (i.e., were more likely to survive). The notion of monitoring pages for changes in content is discussed in some revisitation literature ([17], [26]). Obendorf et al. [26] recorded a hash of each page downloaded in their study of revisitation, and observed that revisited content frequently changed. They additionally found changes often interfered with long-term re-finding. Similarly, Teevan et al. [31] found changes in re-finding behavior as a result of changes to a search engine's result pages.

The work reported in this paper expands on previous work to give a much richer understanding of how the revisitation and content change relate. Specifically, we study both content change and revisitation patterns for pages in a large sample of pages. We develop new metrics and methods to characterize amount and type of Web page change and revisitation patterns, and use them to identify complex relationships between the two.

3. METHODOLOGY

We studied the change and revisitation patterns of over 40,000 Web pages. In this section we discuss how the studied pages were selected, how changes to their content were observed, and how their revisitation patterns were analyzed.

3.1 Data Sample

We sampled pages for study with diverse revisitation patterns using URL visitation data collected from the logs of opt-in users of the Live Seach Toolbar. The toolbar provides augmented search features and reports anonymized usage behavior to a server. Our previous work [1] describes the sampling process in more detail, and we only summarize the selection process here.

We defined three visitation-based page attributes to use for sampling: the number of unique visitors (*unique-visitors*), the median time between a user's visits (*inter-arrival time*), and the median number of visits per user (*per-user revisits*). URLs were tagged with the three attributes using the log data of 612,000 English speaking, non-robot users in the United States from a five week period starting August 1, 2006.

Pages were sampled evenly from pseudo-exponential bins for each attribute. We used 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 x 5 x 6 = 120 bins. Some oversampling of popular pages was added by explicitly including the 5000 most visited pages. Sampled pages were crawled to

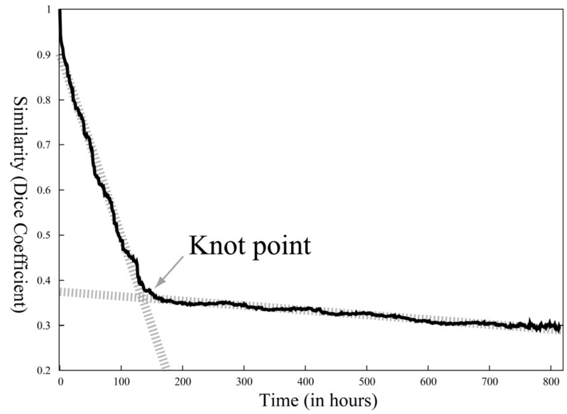

Figure 2. Change curve for http://www.youtube.com (a popular video browsing Web site).

ensure that they were still publicly available (in conformance with the robots.txt), and those that were not were removed. Pages were automatically labeled with high level categories such as news, sports, pornography, etc. URLs were crawled hourly for 5 weeks, starting May 24, 2007, and the pages' HTML stored.

Following the crawl, we returned to the toolbar logs to find the revisitation behavior of the crawled pages that exactly corresponded to the time of the crawl (May – June 2007). Using the behavior of 2.3 million English speaking, US based, non-robot users, 40,817 pages were considered in our final analysis (~25% of the 54,788 pages selected in [1] were not visited sufficiently during the May/June period and were removed from consideration).

3.2 Characterizing Change

We characterize change in two different ways. First, we compute three general change measures—how often a page changes, when it changes, and how different the page is from the previous instance. The difference between successive pages is measured using the Dice coefficient on the textual content of the page. The Dice coefficient measures the overlap in terms between pairs of pages (i.e., $2 * |X \cap Y| / (|X| + |Y|)$ where X and Y are the words that appear in two versions of the page). If the page changes at time 1, 5, and 7—the number of changes is 3, the average time between changes is 3, and the Dice coefficient is computed for the page between for the pairs: times 1 and 5 and times 5 and 7 (0 to 1 is ignored since 0 is the time of the first crawl and not the time the page changed state).

A second measure of change is how much a page evolves from some fixed point in time. In our previous example, if we pick time 0 as our fixed point, we would calculate the difference between time 0 and time 1, time 0 and time 5, and time 0 and time 7. To quantify the change over time of each Web page we use the notion of a *change curve*, introduced in previous work [2]. A change curve represents the amount of textual change (as measured by the Dice coefficient) from a fixed point in the document's history. For each page we select, at random (biased to the first week of samples), up to *n* starting points (generally 5). We define D_t to represent the Web page content at time *t*, and D_{r1} to be the content at the first randomly selected time. Content, for this analysis, is defined to be the page stripped of markup. The value of the change curve at each time point, *t*, is calculated as the average Dice coefficient from each of the randomly selected starting points to the Web page content *t* time steps in future:

$$change(t) = \frac{\sum_{s}^{r1 \ldots rn} dice(D_s . D_{s+t})}{n}$$

Change curves allow us to quickly understand a Web page's evolution over time. An example can been seen in Figure 2. The general form of the change curves is that of a "hockey stick." In other words, most documents rapidly change from the initial starting point as content shifts off the page or is changed during the first few days. For example, in a blog homepage, specific posts move off the page at a certain rate as new posts are made causing a rapid falloff in Dice similarity. At the inflection point (the location at which the change curve flattens) the similarity to all subsequent versions is approximately equal. This is not to say that the document is unchanged past this point, but simply that these pages are *equally similar* to the original page.

To compare different Web pages to each other, and determine the relationship between revisitation behavior and document change, we abstract the change curves. As change curves are generally hockey stick shaped, we do this by identifying the curve's inflection point, or *knot*, and fit two linear regressions to the curve, one up to the knot, and the other following it. The knot point reflects the time at which the amount of change slows (Time) and the overlap in content at the steady state (Dice).

3.3 Characterizing Revisitation

In addition to characterizing Web page change, we looked for patterns in how each page was revisited with the hypothesis that there is a relationship between revisitation and change.

For each page, we looked at the average number of times a URL was visited, the number of unique visitors it received, and the average inter-arrival time. To further compare and evaluate revisitation behavior for different URLs we used the concept of a *revisitation curve* [1]. 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. Curves are generated by calculating all inter-arrival times between consecutive pairs of revisits and binning them, generally into exponential bins. Because histograms are count based, pages that were visited more had higher counts, so we normalized each individual curve by the average of all curves. For example, the Amazon homepage (http://www.amazon.com) revisitation curve () peaks to the right, indicating more revisits happen over a day or longer.

We consider each revisitation curve to be a signature of user behavior in accessing a given Web page. Depending on their shape, revisitation curves were classified into four groups: *fast*, *medium*, *slow*, and *hybrid*. For *fast* revisitation patterns (), people revisited the member Web pages many times over a short interval but rarely revisited over longer intervals. *Slow* revisitation patterns (), with people revisiting the member pages mostly at intervals of a week or more (the Amazon home page above is in this category). *Hybrid* revisitation () is a combination of fast and slow, and displays a bimodal revisitation pattern. Finally, *medium* revisitations () are primarily at intervals between an hour and a day.

3.4 Characterizing Intent

While the toolbar logs enabled us to associate observable revisitation behavior and change, they do not reveal real user intent. For this reason we conducted a complimentary user study to gather information about people's revisitation intent as a function of change. Twenty volunteers (employees of Microsoft) participated in the study (described in [1]). Each participant

Table 1. Several measures of change broken down by revisitation bins. The first set of measures represent the mean number of changes for pages in the bin, the mean time between each change, and the mean amount of change. The second set of measures represent the location of the knot point of the change curve.

Revisitation bin		Change Summary			Change Curve Knot	
		Num.	Time	Dice	Time	Dice
Unique visitors	2	**184.93**	**138.23**	**0.80**	146.10	0.76
	3-6	**211.86**	125.78	**0.83**	143.23	0.77
	7-36	**232.44**	**106.86**	0.83	144.51	0.75
	36+	**254.65**	**102.55**	**0.82**	139.25	0.73
Per-user revisits	2	**172.91**	**133.26**	0.82	**157.38**	**0.78**
	3	**200.51**	**119.24**	0.82	154.53	0.77
	4	**234.32**	**109.59**	0.81	**142.98**	**0.74**
	5-6	**269.63**	**94.54**	0.82	**132.13**	**0.71**
	6+	**341.43**	**81.80**	0.81	**116.96**	**0.68**
Inter-arrival time	<1 day	**214.17**	**126.27**	0.82	**145.37**	**0.75**
	1 day+	**245.06**	108.01	0.82	**133.14**	0.76
	1 week+	**289.34**	**91.49**	0.82	133.32	**0.72**
	2 weeks+	**245.66**	88.06	0.82	**141.81**	**0.73**
	4 weeks+	**211.77**	100.44	0.80	**156.69**	**0.74**
Revisit Curve Bin	Fast	**182.21**	**150.10**	**0.78**	**147.03**	**0.74**
	Medium	**283.15**	**93.66**	**0.80**	**127.82**	**0.71**
	Slow	**212.66**	**111.58**	**0.81**	**153.82**	**0.75**
	Hybrid	**259.03**	**109.88**	**0.81**	**137.04**	**0.74**

installed software to log Web page visits, which they used for one to two months. We recorded visits for the 40K URLs in our crawl as well as a personalized 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 pages they had revisited during the observation period. Of the surveyed URLs, 38% (61 URLs) overlapped with the 40K pages in the log study. The "personalized" random sample, representing the remaining 62%, was also used in 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, whether they visited it at regular intervals, and how often they expected the page to change. By relating actual page change, people's expectation of change, and their stated intent behind their revisitation, we are able to better explain the behavior we observed.

4. REVISITATION & CHANGE TRENDS

In this section we explore the high level relationships between our rich behavioral data and our change data. We start with the simple hypothesis that increased user revisitation behavior correlates positively with increased change, and find that the connection between the two is not necessarily simple. For example, though we observe that generally pages that change a lot were visited more often and were revisited after shorter intervals, the *amount* of change from version to version did not correlate to any of our metrics.

Table 1 summarizes the findings discussed in this section. Both the discussion and the table are broken down by the three measures of revisitation discussed earlier: the number of unique visitors to a page (*unique-visitors*), the 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*). For each behavior measure, Table 1 presents three measures of change (number of changes, time between changes, Dice coefficient for successive changes) and the coordinates of the knot point (time and Dice). The significance of each measure is tested by applying an ANOVA to an ordered binning of the particular metric for overall significance. Bolded results are significant against the previous bin through Tukey's HSD post-hoc.

4.1 Unique Visitors and Change

We begin our analysis by looking at how the number of unique visitors to a page correlates with changes to the page's content. Although static pages can be popular, it is more likely that continued popularity is achieved through some dynamic content and maintenance. As the number of unique visitors increases, the mean number of changes we observed increases, and the time between each successive change decreases—ranging from ~138 hours between changes for pages with only 2 unique visitors to ~102 hours for those with 36 or more unique visitors.

However, the mean amount of change (as calculated by the Dice coefficient) does not have a similarly distinct trend. The most *and* least popular pages both have the biggest changes between versions (though the difference between all bins is fairly small, ranging between Dice coefficients of .8 and .83). Thus, while popular sites change more frequently, the same cannot be said about the amount by which they change. This is an indication that the *amount of change* may not be as important as *what is changing*.

The knot point does not differ significantly with changes to the number of visitors. This may be anticipated as page popularity does not tell us how often revisitation occurs in the specific page only that it occurred more than once. Recall that the knot point is measuring the approximate location when pages have "stabilized"—when every subsequent page is equally (dis)similar to the starting point. If we believe that individuals will try to synchronize their revisitation behavior to catch content before it "decays" off the page (e.g., at the knot point), we must look to the number, and interval, of *per-user* revisitations.

4.2 Per-User Revisits and Change

Analysis of the number of times an individual revisits a Web page (or a page's "stickiness") reveals that the pages individuals revisited more times changed more frequently and at shorter intervals. For example, revisited pages changed only twice changed every 138 hours, on average, while those that were revisited 6 or more times changed every 81.8 hours. However, the amount of change does not appear to trend in a particular direction, reinforcing that the amount of change is not as crucial as the specific information that is changing.

Unlike the popularity category, we *do* find a trending in per-user revisitation when compared to knot location. The more the average user revisits a page, the earlier the knot point and the more different the eventual steady state is to the original page. The implication of this is that users may revisit more often in order to capture content that will vanish from the page. This is consistent with the model that the rate of change before the knot point (the initial, steeper, downward slope) represents the rate at which the dynamic data on the page is "lost."

Further evidence for this is provided by our smaller user study which indicated that people appeared to have a reasonable understanding of Web content change. The knot point for the pages where participants expected meaningful change upon revisiting was sooner than for pages where meaningful change was not expected (60.7 hours v. 97.0, $p < 0.05$).

4.3 Inter-Arrival Time and Change

Because the same number of revisitations can occur very quickly (e.g., 5 revisits in 2 minutes and never again) or very slowly (e.g., 1 revisit per week over 5 weeks), it is additionally worth considering the average inter-arrival time.

We might expect that the more rapidly a page changes, the lower the inter-arrival time (the faster the revisits). However, we find that as the inter-arrival time increases, the number of times a page changes increases for inter-arrival times of less than 2 weeks before going down again. This is somewhat counter intuitive as it means that pages with both the high and low inter-arrival times are those with the fewest changes. It is here that we first begin to recognize situations in which revisitation patterns are not necessarily related to frequency of change—an issue we will return when comparing revisitation curves and change curves. The mean time between changes shows a similar bowed pattern, with the longest times for pages that change slowly or rapidly.

This result may also be explained when considering our previous work [1] in which we analyzed visitor behavior in different types of revisitation (e.g. fast, slow, medium and hybrid). Recalling that the mean inter-arrival time relates to the revisitation peak, a "fast" revisitation curve corresponds to primarily low inter-arrival times. As above, pages in the very fast revisitation category, where people revisit a page a lot during a short period of time but never return after a longer interval, change slowly (see Table 1). This would again seem to contradict our hypothesis that revisitation should match change. However, as noted in previous work [1], 77% of revisitations in the fast category were preceded by a visit to a page from the same domain—indicating a "pogo-stick" browsing behavior (rather than change monitoring). Since users exhibiting this behavior are simply surfing back and forth from the origin page, they are less likely to be interested in monitoring changes on that page in the short term.

Thus, we may refine our hypothesis to exclude those fast revisitations that are more likely the result of page and link structure rather than any kind of monitoring intent. The remaining categories (e.g., medium and slow) do appear consistent with our hypothesis. The number of changes (higher for the faster revisits), the average time between change (lower for faster revisits), the location of the knot point (content vanishes more rapidly for faster revisits) and eventual stable state of the change curve (greater changes for faster revisits) all display significant differences.

With the exception of the very shortest inter-arrival time, we do find the expected relationship given the knot point. The more time it takes for the content to vanish off the page (further knot points) and the less the eventual steady-state (lower Dice), the longer the inter-arrival time is. Again we see less revisitation for content that takes more time to change.

4.3.1 *Importance of Change*

Figure 3a shows a different representation of the knot data, tracking inter-arrival times, per-user revisits and knot points simultaneously. The figure shows the number of revisits that occur as a function of the knot point. The data is further stratified into 5 different revisit intervals (i.e., the average revisit is the same day, after 2 days, after 7 days, after 14 days and after 27 days). As might be expected, the number of revisits with a shorter revisitation interval (e.g., 2 days) is higher than the number for longer revisits (e.g. 27 days). The figure also shows the best fitting linear function (on a log-linear scale) for each of the 5 revisit intervals. In agreement with our prior observations, the linear functions all have negative slopes indicating that when the knot point occurred after a long period (e.g., four weeks), there were fewer revisits than when the knot point occurred early (e.g., one day).

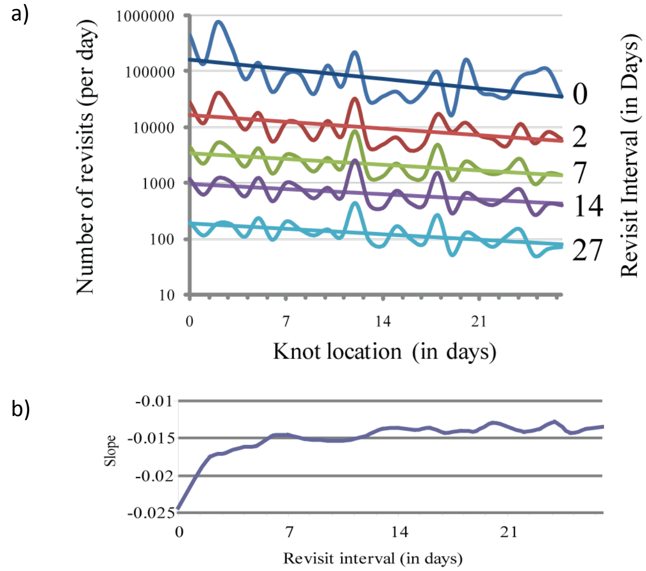

Figure 3a) A comparison of knot location to the average number of revisits stratified by different average revisit intervals (e.g., the bottom line is the number of revisits per day as a function of the knot location for pages with an average revisitation rate of 27 days). A linear regression is fit on each individual curve, and in 3b) the slope of the different interval regressions is tracked (i.e. the slope of each linear regression in 3a going from 0 to 27).

Most interestingly, the linear functions have *different* slopes. When people revisited a page quickly (e.g., within the same day, top curve), those revisitations are strongly related to how frequently the page changed (people revisited more as the page changed more frequently). On the other hand, when people revisited a page slowly (e.g., after many weeks, bottom curves), those revisitations are less related to how frequently the page changed. Or stated another way, the slower the revisits, the less the importance of change. This can be seen further in Figure 3b, which plots the slope for the linear function for revisits for the full range of intervals. The flattening of the curves as the revisit interval increases suggests that content change is more closely related to short term revisitation behavior than long term revisitation behavior.

To summarize some of our key findings thus far:

- The more popular the page, the more rapidly it changes.
- The more times a page is revisited, the more rapidly it changes, and the earlier the amount of change stabilizes.

- Page revisitation is not directly synchronized to the amount of time between changes or how quickly the change stabilizes. This may be because not all revisitation is motivated by monitoring.
- Quick revisits (e.g., within the same day) are more strongly related to change. Thus, short term revisitation behavior is more closely tied to change, and the relation is non-linear.

5. DYNAMIC AND STATIC INTENTS

Despite the general tendencies in our data, the precise relation between peaks in revisitation curves and knot points is much more nuanced. If individuals are revisiting with the sole intention of monitoring, we might expect that the bulk of revisits occur before a page changes so significantly that data will be lost (i.e. before or around the knot point). However, selecting those URLs in the with a fixed knot point (e.g., 56-113 hours), we find that the maximum peaks in their revisitation curves are dispersed (see Figure 4) for the pages in the medium and slow categories. Note that pages in the fast category which include, noisy, non-monitoring behaviors (e.g., "pogo-stick") are removed in this example. In this example, 16% of the URLs' revisitation curves peak before the knot, 72% peak after and only 12% peak at the knot point. Thus, while we might see general trending in revisitation—where more changes or earlier knots leads to more revisits—visitors do not appear to be synchronizing to some exact time point. We hypothesize this difference could relate to whether users are interested more in the changing content of the Web page or in the (more) stable content.

Figure 5 shows three different examples relating change to revisits with revisits peaking before (www.nytimes.com), at the same time (www.woot.com), and after (www.costco.com) the knot point. Visitors to the New York Times Web site are typically interested in finding information about current news events and are therefore interested in any changing content. Catching those stories *before* they decay off the page (i.e., before the knot point) leads to higher revisitation in the early periods. In contrast, Woot, a website that offers a new, one-time offer for electronic goods once a day (every 24 hours, corresponding to the knot point), experiences increased revisitation at the *same* time that the new deal has been posted. The Costco homepage, which provides entry to the mega-warehouse's Internet site has revisitation rates peaking far *after* the knot point. Although the page presents new deals, it also provides entry to the company's catalog, store information and other details which are likely not needed on a daily basis. The late peaking of the revisitation curve relative to the early knot point may indicate a user need for accessing the stable, unchanging aspects of the website.

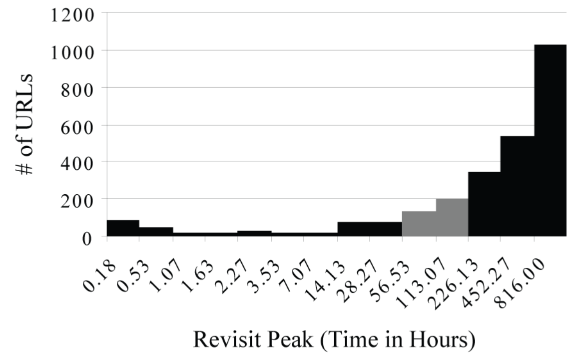

Figure 4. Distribution of revisitation peaks (for pages in the medium and slow revisitation categories) for a fixed knot point (56-113 hours, colored in gray).

5.1 Before, During, and After Knot Points

To more formally verify our observations we construct subsets of pages that have a more clearly defined purpose and relate the change curves to knot points for those pages. Specifically, we consider categories of pages where we have some expectation about the motivation of revisitation (e.g., news or shopping).

In order to evaluate whether we have significantly more revisits than expected before or after the knot point we generate a set of bins that are dependent on the position of the knot point. Because we are interested in revisitation peaks that are "just before" or "just after" (as in the Woot example), we used 2.5% - 5% of the timeline (~20 – 40 hours) to define our bins immediately before and after. In total, we look at 4 bins for each URL: far before the knot point (more than 20-40 hours before), immediately before, immediately after, and far after (more than 20-40 hours). For each bin we count the number of revisits, and normalize these counts by the *expected* number of revisits for all Web pages. The normalized bins therefore represent the percentage of expected revisits actually observed in each bin (< 1 is less than expected, > 1 is greater). To test our hypothesis we select a *group of interest* (e.g. news pages or forum pages) and a control set of pages not in this group but with a similar distribution of temporal knot locations. We compare the group of interest by the prevalence of revisits in bins before and after the knot point.

We looked at four groups of interest—news pages and forums, which are used to keep up with new information, and pornographic and shopping pages which often have rapidly changing ads. These four specific types were selected as we have some testable expectation about the relationship between change

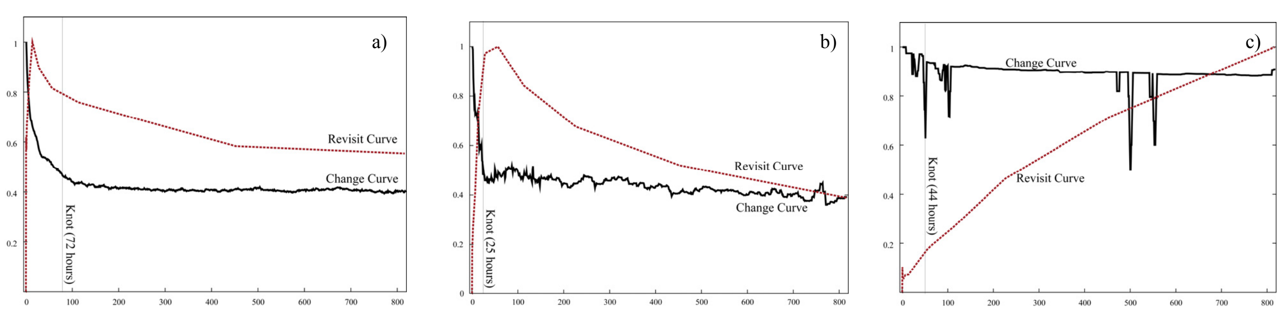

Figure 5. Revisitation and (normalized) change curves for the home pages of a) the New York Times (http://www.nytimes.com), b) Woot (http://www.woot.com), and c) Costco (http://www.costco.com).

and revisitation. Some key findings include:

- Homepages for *news organizations*, which are generally used to find new information, tended to be to more to the left of the knot (i.e. before content has vanished, Kruskal-Wallis (KW), $p < 0.00001$).
- *Forum pages* (those with "forum" in the URL) also display a revisitation tendency to the left of the knot (KW, $p < 0.001$).
- We would expect that pages with a large number of rapidly changing *ads/spam* are less likely to attract revisits that match this frequency of change. This pattern is seen for homepages classified as *pornographic*, where revisitations are more to the right than average (KW, $p < 0.01$).
- Revisits for homepages of *retailers* ("Shopping" categories), are generally to the right of the knot indicating that the rapidly changing information is less critical in driving revisits (KW, $p < 0.01$).

In general, these results confirm for us a relationship between knot point and "intent" as measured by revisits. However, further evidence can be found by directly asking users in our smaller scale user study.

5.2 Revisitation Intent and Change

In our user survey we asked participants about their intent in revisiting Web pages. Intents included finding or monitoring new information, re-finding previously viewed information, form filling, communication, shopping, and homepage (i.e., accessing the browser startup page). Participants were more likely to be interested in finding or monitoring new information in pages that changed rapidly, and more likely to re-find previously viewed information in pages that changed less frequently. For 19 of the URLs, participants explicitly responded they were looking for new information when they visited the page. For 11 of the pages they responded that they were monitoring information, and for nine they were interested in previously viewed information. The mean/median knot point for each was 59.8/46 (new information), 53.3/46 (monitoring), and 88.9/87 hours (previously viewed). We find suggestive evidence that the knot point was sooner (weakly significantly at $p < 0.1$) for monitoring and finding new information compared to visiting old information. All four instances where the page was used to communicate with other people (via email or message boards) involved very quick knot points (the longest being 23 hours) and medium revisitation patterns.

Looking more closely at the specific reasons people gave for revisiting certain URLs, the most common reason (given 29 times) was to use a search engine or enter data in a form. Pages marked with this revisitation reason changed less frequently, with a knot point of 85.4 hours instead of 61.9 hours. These are pages where the participants appeared to not be interested in change. As one person stated, "I am pretty sure the page changes regularly, but as I am interested in is the search field, and it doesn't change. I don't notice anything else."

6. RESONANCE AND STRUCTURE

As we have illustrated above, resonance is not necessarily between revisitation and the overall change rate of the page. Some revisitation behavior resonates with the fastest changing content on the page, others with the more stable information. Thus, we would like to find a mechanism for separating out different portions of the page and identifying those most likely to be relevant to the visitor. In the past, identifying such content would require more expensive eye or mouse tracking studies, interviews, or surveys. Instead, we offer a simple approach that utilizes revisitation and change information to automatically identify these targets. Though ideally we might combine this algorithm with additional information, we believe that this technique represents a novel, extensible, mechanism for partitioning pages into more, and less, important information.

Abstractly, we would like to rank sub-pieces of the page—which are each changing at a different rate—by their similarity to the revisitation rate. To accomplish this we briefly introduce an algorithm for effectively labeling change rates of Document Object Model (DOM) structures. Web pages are semi-structured constructs composed of a tree of DOM objects, and we wish to label each DOM element with the rate at which it changes.

Our technique makes use of the algorithms defined in [2] which are intended to act on multiple copies of the same page in an efficient manner. The essential details for this particular application are that each version of a Web page is *serialized* to an easy to process structure. For example, if the HTML document at time t_i was simply:

```
<BODY><B>text1</B><IMG SRC="image1"></BODY>
```

we would generate a two line file:

`/body[001]/b[001]` t_i `hash(text1)`
`/body[001]/img[001]` t_i `hash(image1)`

Where the hash() is a function (e.g. MD5) which takes as input some text and produces a short, unique hash value. If the second version of the page at time t_j, was:

```
<BODY><B>text2</B><IMG SRC="image2"></BODY>
```

we would add to our file:

`/body[001]/b[001]` t_j `hash(text2)`

`/body[001]/img[001]` t_j `hash(image2)`

Taking this combined file and sorting it, we can make a single pass through the data to determine the rate of change for any DOM element (i.e., by calculating the mean or median difference between each subsequent times where the hashes of the data are *not* equal). With enough evidence, for example, we might find that the bolded text changes once every hour whereas the image only changes once a day.

We next illustrate these DOM level change patterns. Figure 6a-c shows histograms of the proportion of DOM elements that change at different points in time for three different pages. Note that individual DOM elements change at different times, which is not reflected in the page-level change measures (which change at the fastest rate of any individual element). Above each histogram are the change and revisitation curves for the page as a whole. The grey vertical bars highlight the most prevalent revisitation interval (i.e., the peak of the revisitation curve). Given the change rates associated with each element, we can apply a filter that selects those elements that are changing at approximately the same rate as revisitation. Figure 6d-f is an image corresponding to the three Web pages, each illustrating a different revisitation-to-change relationships. The Seattle Post Intelligencer page (d), for example, has very fast revisitation periods. Masking elements with slow rates of change hides navigation elements and slow changing information leaving only "breaking news", current weather, and advertisements). Masking in the Woot page (e) pulls out those elements that change approximately every 24 hours, which is information about the new product being sold. Finally, the Tribute.ca site (f), the homepage for a movie rating and showtime database for Canada, has very slow revisits. Masking fast changing elements leaves only the navigation and search

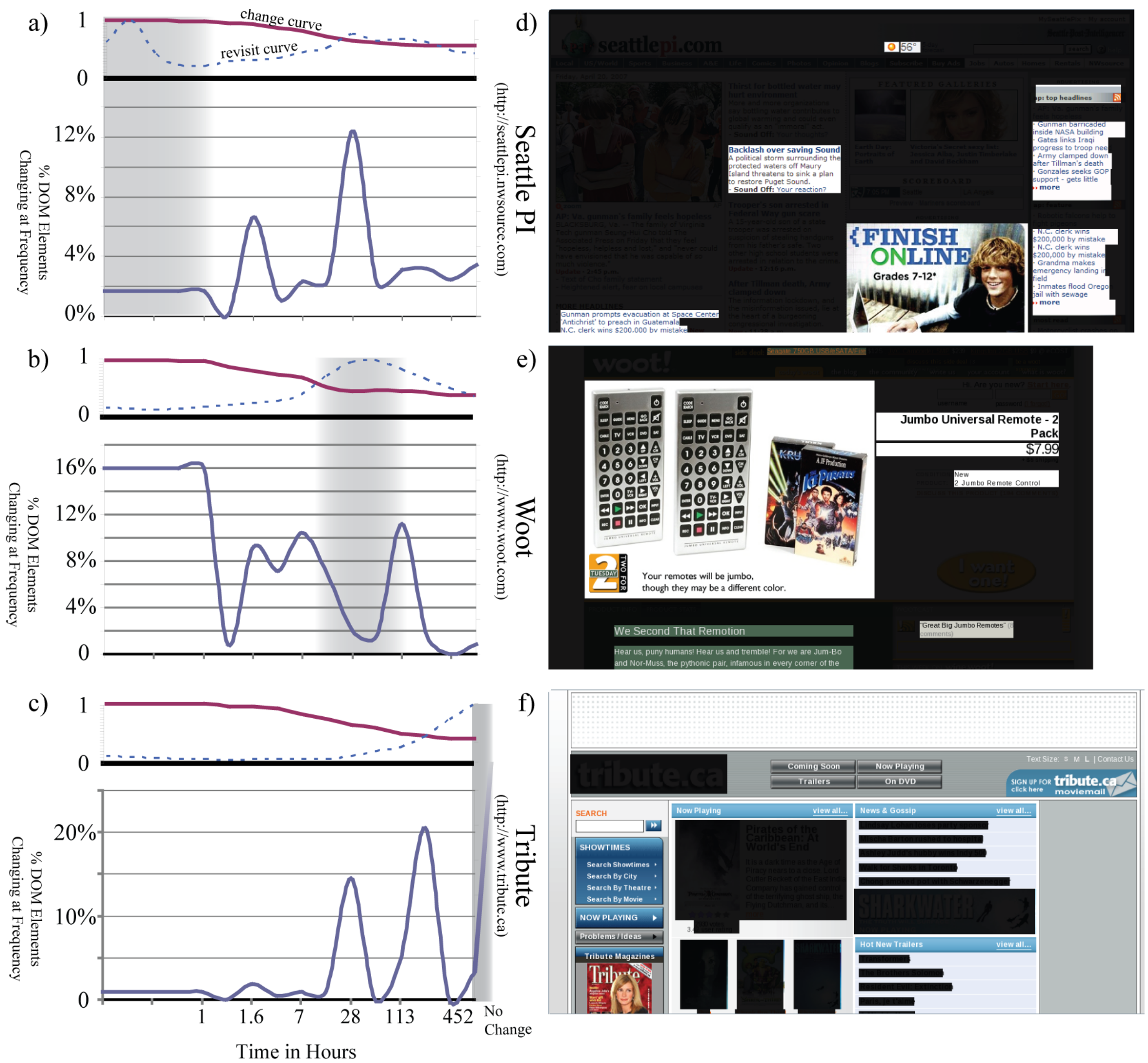

Figure 6a-c) A combined chart for each page's revisitation, change and DOM curves. The change and revisit curves are normalized to 1. Beneath each change/revisit pair is a plot of the amount of the page (as percent of DOM elements) changing at a given rate. The gray bars represent the approximate peak of revisitation behavior. Figures d-e) are the evaluated pages showing DOM elements that are changing at approximately the peak revisitation rate (for the Tribute page this includes unchanging content).

elements which are the likely targets of revisitations. A demonstration applet of the algorithm is available at http://cond.org/resonance.html.

The algorithm works particularly well when there are elements on the page are clearly differentiated. This may not always be the case as some "undesired" content may change at the same rate as content that is being monitored. For example, the stock market average on the New York Times homepage changes rapidly—likely on every visit of the crawler—as do advertisements. In this situation, both elements are displayed, but it is likely one is of more interest than the other. Additional work on grouping and clustering elements that are near each other visually or have certain shapes consistent with advertising may help. Additionally, use of of click logs and potentially mouse or eye tracking can further improve the results of this algorithm. The benefit of the approach we propose is that it can be automated and scaled to many pages very easily. Historical revisitation information can provide a unique mechanism for inferring which *portions* of the page are of interest to page visitors. This approach may of course be personalized, with the filter set to different values, as different groups or individuals may have revisitation rates that diverge from the average.

7. APPLICATIONS AND IMPLICATIONS

There are many ways the results of our analysis can be used to improve the Web experience. In this section we discuss how the relationship between change and revisitation can be used to inform Website, browser, and search engine design.

7.1 Website Implications

Website designers would like to understand why users are returning to their pages. While some inference can be made from the links that are clicked on, there are many situations when users revisit and do not click on anything. Our research illustrates a mechanism by which a site owner can gain additional insight into the content that is motivating revisitation behavior.

A more specific application for Website owners is an optimization of the "what's new" pages that visitors utilize to determine new content that is of interest. As argued in [1], because different pages, even on the same site, are revisited at different rates, "what's new" pages can be designed at different granularities. The work presented here further argues that the resonance between what is changing and the revisitation pattern may point at content that is more of interest. Thus, a website can identify the rate of change of pages or portions of the page, and highlight those that correspond to the peak revisitation rates.

7.2 Web Browser Design Implications

Change monitoring is an area of active interest for browser implementers and researchers alike (e.g., [6][17][24]). Just as a Website designer may create optimized "what's new" information for their pages, a client-side implementation may provide additional change analysis features to the user. The ability to expose and interact with meaningful change would be particularly useful within a client-side, Web browser context, where a user's history of interaction is known. Pages displayed in the browser could be annotated to provide the user with (highlighting not just any change, but those changes that are relevant given revisitation resonance). A browser could also act as a personal Web crawler, and pre-fetch pages that are likely to experience meaningful change (as measured by the revisitation/change resonance). This would allow for a faster Web experience and give users the ability to access new content in offline environments. Rather than only storing the most recent version of the page (or all versions), one could develop a caching system that only stores those pages with changed content that is likely to be of interest (or conversely displaying old versions if stability is preferable).

Other applications where resonance may be used are mobile browsers where content from the original page may be filtered (as illustrated in Figure 6) to highlight what is more likely to be of interest to the user. For example, knowing that a news site has fast revisitation would allow the mobile application to pull out the rapidly changing content and hide or reorder the display to downplay slow changing or unchanging information. Conversely, a page with slow revisitation patterns might be filtered of fast changing content to highlight navigation and search structures. Removing information that is less likely of interest might save bandwidth and screen real estate in other applications as well.

7.3 Search Engine Implications

Our analysis of revisitation and change also has a number of implications for search engine design, in particular to the related issue of re-finding. Prior research has demonstrated re-finding behavior is prevalent [31] in search engine use. Our analysis, which further demonstrates a relationship between the changes to pages and specific kinds of revisitation behavior, suggests several ways search engines can support re-finding in the dynamic environment of the Web.

Just as a browser can provide intelligent re-crawling based on the resonance between revisitation and change, a search engine can achieve the same result on a much larger scale. Optimized crawling may lead to crawling strategies that understand what information on a page is interesting and should be tracked more or less aggressively for indexing. For example, change in advertising content or change that occurs as a result of touching a page (e.g., changes to visit counters or information about page load time) should not inspire repeat crawls, while change to important content should. Furthermore, a document need not be indexed if it has not changed in a meaningful way, potentially saving server resources.

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 we suspect are interesting because they contain new content, 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 ones where we suspect the static content is interesting, 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.

A search engine with a rich understanding that the Web is a dynamic information environment could also benefit its end users in more direct, obvious ways by exposing page change in its user interface. If we can make an intelligent guess as to whether searchers who are revisiting previous information sources are interested in re-finding previously viewed content or in viewing newly available content, we can better support both behaviors. For searchers interested in new content (identified either by the query alone or by the user's query history), the search result summaries could highlight the new content

8. CONCLUSION AND FUTURE WORK

Though much research has been generated on both the evolution of the Web, and the revisitation behavior of users, little has been done to tie the two together.

The research presented here makes a significant step in understanding the association between change and access. Our study is unique among studies of Web content change in that the pages are actually used, and unique among studies of revisitation in that we focus on how content change relates to revisitation. In this paper we have taken a very fine grained crawl of 40k documents and related that to the revisitation patterns of 2.3 million users. We have identified and quantified both the non-linear relationships between behavior and change as well as the importance of changing content in different situations. Because simple assumptions and metrics of change/revisitation interaction limit our ability to understand and leverage this relationship, we introduce new metrics and techniques. Our analysis provides an alternative to other metrics and allows us to infer potential features of interest on the page, be they highly dynamic content, stable search and navigation, or something in between. Additionally, we have illustrated how different revisitation patterns resonate with different kinds of changes

The implications of the relationships between revisitation behavior and change have applicability to a wide range of services from the individual's browser to the community's search engine. For the individual user interested in monitoring content or re-accessing what was there before, it is valuable to design systems that are cognizant of how information changes. Systems that are aware of potential intent in relation to the changing information should be able to leverage this information in any situation where monitoring, revisiting, and re-finding behaviors exist. This understanding may also enable new applications. For example, by recognizing the revisitation patterns of the user, a mobile browser might filter content to only display stable information, or only render that which is changed.

In addition to sampling pages by behavior, as we do in this study, we would like to expand the page collection to include enough Web pages of a specific type (e.g., “sports scores” or “health management”) to perform additional statistical analysis of change and revisitation by type. We believe that there are also a number of future opportunities in studying revisitation/change resonance. Our approach DOM level analysis, for example, has concentrated on Boolean notions of change (i.e., is the new version of the text different than the previous version). By studying the amount of change and applying more sophisticated spectral analysis techniques, it may be possible to differentiate between the frequency of major changes (e.g., a new news story) and that of minor changes (e.g., an update to an existing story). We hope to combine our DOM-based algorithm with additional behavioral data, and additional testing, to further refine the automatic detection of content that is important when people revisit.

9. ACKNOWLEDGEMENTS

We would like to thank Dan Liebling, Ronnie Chaiken, Bill Ramsey, and Dennis Fetterly for their help in obtaining and analyzing the data. We also appreciate helpful discussions with Sara Adar and Dan Weld.

10. REFERENCES

[1] Adar, E., J. Teevan, and S. T. Dumais. Large scale analysis of Web revisitation patterns. CHI ’08, 2008.

[2] Adar, E., J. Teevan, S. T. Dumais, and J. L. Elsas. The Web changes everything: Understanding the dynamics of Web content. WSDM ’09, 2009.

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

[4] Ayers, E. and J. Stasko. Using graphic history in browsing the World Wide Web. WWW ’95, 1995.

[5] Catledge, L.D. and J.E. Pitkow. Characterizing browsing strategies in the World-Wide Web. WWW ’95, 1995.

[6] ChangeDetect (2007). ChangeDetect Web Page Monitoring

[7] Cho, J. and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. VLDB ’00, 2000.

[8] 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.

[9] Douglis, F., A. Feldmann, and B. Krishnamurthy. Rate of change and other metrics: A live study of the World Wide Web. USENIX Symp. on Internet Tech. and Systems, 1997.

[10] Fetterly, D., M. Manasse, M. Najork, and Wiener, J. A large-scale study of the evolution of Web pages. WWW ‘03, 2003.

[11] Friedman, J., T. Hastie, and R. Tibshirani, Additive logistic regression: A statistical view of boosting, *Ann. Statist.* 28(20):337-407, 2000.

[12] Grandi, F., Introducing an annotated bibliography on temporal and evolution aspects in the World Wide Web, *SIGMOD Records*, 33(2):84-86, 2004.

[13] Greenberg, S. and A. Cockburn. Getting back to back: Alternate behaviors for a Web browser's back button. 5th Annual Human Factors and the Web Conference, 1999.

[14] Herder, E. Characterizations of user Web revisit behavior. Workshop on Adaptivity and User Modeling in Interactive Systems, 2005.

[15] Jones, W., S. Dumais, and H. Bruce. Once found, what then?: A study of ‘keeping’ behaviors in the personal use of Web information. ASIST ‘02, 2002.

[16] Kaasten, S. and S. Greenberg. Designing an integrated bookmark / history system for Web browsing. Western Computer Graphics Symposium, 2000.

[17] Kellar, M., C. Watters, and K. M. Inkpen. An exploration of Web-based monitoring: Implications for design. CHI’07, 2007.

[18] Kellar, M., C. Watters, and M. Shepherd. A goal-based classification of Web information tasks. ASIST ’06, 2006.

[19] Kim, J. K., and S. H. Lee. An empirical study of the change of Web pages. APWeb ’05, 2005.

[20] Koehler, W. Web page change and persistence: A four-year longitudinal study. *JASIST*, 53(2):162-171, 2002.

[21] Kwon, S. H., S. H. Lee, and S. J. Kim. Effective criteria for Web page changes. APWeb ’06, 2006.

[22] Milic-Frayling, N., R. Jones, K. Rodden, Smyth, G., Blackwell, A., and Sommerer, R. Smartback: Supporting users in back navigation. WWW ’04, 2004.

[23] Morrison, J. B., P. Pirolli, and S, K. Card. A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions. CHI ‘01, 2001.

[24] Nadamoto, A, and K. Tanaka. A comparative Web browser (CWB) for browsing and comparing Web pages. WWW ’03, 2003.

[25] Ntoulas, A., Cho, J., and Olston, C. What’s new on the Web? The evolution of the Web from a search engine perspective. WWW ’04, 2004.

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

[27] Pitkow, J. and Pirolli, P. Life, death, and lawfulness on the electronic frontier. CHI ’97, 1997.

[28] Sellen, A. J., R. Murphy, and K.L. Shaw. How knowledge workers use the Web. CHI ’02, 2002.

[29] Takano, H. and T. Winograd. Dynamic bookmarks for the WWW. Hypertext ‘98, 1998.

[30] 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.

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

[32] White, R. W. and S. M. Drucker. Investigating behavioral variability in Web search. WWW ‘07, 2007.