Int. J. Human-Computer Studies (2000) 000–000

What Do Web Users Do? An Empirical Analysis of Web Use

Andy Cockburn and Bruce McKenzie
Department of Computer Science
University of Canterbury
Christchurch, New Zealand

{*andy,bruce*}*@cosc.canterbury.ac.nz*

This paper provides an empirical characterisation of user actions at the web browser. The study is based on an analysis of four months of logged client-side data that describes user actions with recent versions of Netscape Navigator. In particular, the logged data allows us to determine the title, URL and time of each page visit, how often they visited each page, how long they spent at each page, the growth and content of bookmark collections, as well as a variety of other aspects of user interaction with the web. The results update and extend prior empirical characterisations of web use. Among the results we show that web page revisitation is a much more prevalent activity than previously reported (approximately 81% of pages have been previously visited by the user), that most pages are visited for a surprisingly short period of time, that users maintain large (and possibly overwhelming) bookmark collections, and that there is a marked lack of commonality in the pages visited by different users. These results have implications for a wide range of web-based tools including the interface features provided by web-browsers, the design of caching proxy servers, and the design of efficient web-sites.

1. Introduction

The World Wide Web, and the web-browsers used to access it, are inextricably linked with most people's computing experience. Given the predominance of the WWW in everyday computing, there is a surprising lack of research into how the web is used.

This study aims to update and extend the understanding of web-use. In particular, we wish to empirically characterise people's navigation patterns in visiting and revisiting pages. Improved understanding of web navigation patterns has many practical applications. Our primary objective is to provide information that can guide the design of next-generation web-browsing interfaces so that they better support common navigational activities. Beyond improving browser interfaces, there are several other areas that should benefit from an improved empirical foundation of web use. These include the design of caching proxy servers, search engines, collaborative information systems, and web-pages.

There are three main categories of techniques that can be used to investigate user interaction with the web. Firstly, query techniques such as interviews and question-

naires can be used to retrieve background demographics on user groups as well as subjective data on aspects of the user's conscious interaction with the web. Since 1994, the twice-yearly "WWW User Surveys" run at the Georgia Institute of Technology's Graphics, Visualization, and Usability Center [1] have provided the definitive source for this type of information: for example, see Kehoe & Pitkow (1996).

The main attraction of questionnaires is the relative ease of gathering a large set of responses. Their primary limitations, however, are their narrow scope and their separation from the user's task: they report on the user's perceived, rather than actual, interaction.

Secondly, dynamic observation techniques such as controlled experiments, think-aloud studies and ethnographic studies observe and record the user's actions *while* they occur. In understanding web use, controlled experiments have been used to investigate 'micro' issues, such as the investigation into user models of the 'Back' button by Cockburn & Jones (1996). Think-aloud studies, in contrast, are generally used to gain insights into 'macro' web usability issues such as users' high-level tasks and goals: for example, Byrne, John, Wehrle & Crow (1999). Ethnographic studies broaden the scope of investigation further, placing a heavy emphasis on the context of the users' actions within, and around, the web. Clearly this is a long-term and time-consuming experimental technique. Bellotti & Rogers's (1997) six-month study of the publication industry provides the best approximation of web-based ethnographic methods that we are aware of. However, Bellotti & Rogers do not use the term 'ethnography' to describe their method, which relied heavily on interviews rather than 'immersion' in the environment under study.

Thirdly, static observation techniques review traces of the user's actions *after* they have occurred. In web research, the main sources of these traces are logs captured at either the client or the server. Client-side logs reveal the history of the users' actions with the browser: they allow insights into how frequently various interface components of the browser are used (such as the 'Back' button), and how frequently users navigate to particular pages. Server-side logs show data on the number of requests for particular pages within a site, together with the sources of those requests (in terms of an IP address). Server-side logs are generally used to provide statistics to site designers, but they have also been used to construct navigational aids such as site-maps that show the common trails through the information space (Wexelblat & Maes 1999). Pirolli, Pitkow & Rao (1996) and, more recently, Chi, Pirolli & Pitkow (2000) discuss the problems and uses of extracting meaningful information from web-server logs.

The analysis reported in this paper uses a static observation technique that draws on client-side logs of user interaction with the browser. The aim is to update prior studies of web-use that were based on client-side logs (Catledge & Pitkow 1995, Tauscher & Greenberg 1997), and to overcome some of their limitations (see Section 2). The analysis method (Section 3) draws on four months of daily log files for seventeen users. Each user's daily log file shows details of the pages they visited, the number of times they visited them, the timing of page visits, and changes to the user's bookmark collection. Section 4 provides the analysis and results of the study. The implications

[1] http://www.cc.gatech.edu/user_surveys/

of the results for browser and web-page design are discussed in Section 5, and the limitations of the study are presented in Section 6. Section 7 concludes the paper.

2. Prior Empirical Analyses of Web Use

Two prior studies provide the empirical foundation for our current understanding of user interaction with the web: Catledge & Pitkow (1995) and Tauscher & Greenberg (1997). Both of these studies were based on client-side log files.

Catledge & Pitkow's (1995) study involved 107 users who were staff, faculty and students in Georgia Institute of Technology's Computing Department. In three weeks, 31134 navigation acts with the XMosaic (version 2.4) browser were logged, giving a mean page visit rate of approximately fourteen pages for each user per day. Their study revealed that the dominant user interface techniques for visiting pages were clicking on hypertext anchors (52%) and on the 'Back' button (41%). Navigating to pages by typing the URL, by clicking 'Forward', or by selecting from 'Bookmarks' (also termed 'Hotlist' or 'Favourites') were all lightly used, each accounting for about 2% of navigational actions.

Tauscher & Greenberg's (1997) study involved 23 subjects, selected from a pool of nineteen staff, faculty and students in a Computer Science department and nine programmers and software engineers in a telecommunications company. As in Catledge & Pitkow's study, the web-browser used was XMosaic (version 2.6). Approximately 19000 navigation acts were logged during a five to six week period, giving a mean page visit rate of around 21 pages for each user per day. Their study confirmed that link selection (clicking on an anchor in the page) and 'Back' are the dominant navigation mechanisms, accounting for approximately 50% and 30% of navigation acts.

As well as analysing user actions at the web browser, Tauscher & Greenberg focused on the *recurrence rate* of page visits: "the probability that any URL visited is a repeat of a previous visit, expressed as a percentage". They found that the recurrence rate for the subjects participating in their study was 58%, and by re-analysing the data from 55 of Catledge & Pitkow's subjects they found a recurrence rate of 61%. This result shows that users had previously seen approximately three out of five pages visited.

Although both studies showed low use of bookmarking techniques (less than 2% of user actions), a 1996 survey (Abrams, Baecker & Chignell 1998) indicated that bookmarks were becoming more heavily used, with 84% of subjects having more than eleven bookmarks. Indeed, Pitkow (1996) reported from a survey of 6619 users that "organizing retrieved information" is one of the top three usability problems of using the web, reported by 34% of the participants.

2.1. LIMITATIONS

Though excellent studies, there are four main reasons for suspecting that these findings may no longer reflect current use of the web: the growth of the web, the evolution of web-navigation aids, the fact that the subjects were not using their 'normal' browser, the relatively crude interface of the browser studied, and the duration of the evaluations.

2.1.1. *Growth of the web*

As Kehoe & Pitkow (1996) state "Five years is not very long on most historical scales, but for the World Wide Web (WWW) it constitutes a lifetime." The two main studies were carried out in 1994 (Catledge & Pitkow) and 1995 (Tauscher & Greenberg). Given the relative youth of the web at this time, and its continued exponential growth, it seems reasonable to suspect that usage patterns may have evolved and matured.

2.1.2. *Evolution of navigation aids*

Web navigation aids are now a fundamental part of web use: web search engines, meta search engines, web directories and search agents are all commonly used to access web information. These tools, however, were either in their infancy or did not exist at the time of the prior studies.

2.1.3. *Browsers of preference*

Both studies analysed use of a specially equipped 'logging' version of NCSA's XMosaic browser. Tauscher & Greenberg state that none of their subjects used XMosaic as their normal everyday web-browser. Similarly, Catledge & Pitkow indicate that subjects may have chosen to use a browser other than XMosaic. Clearly, the subjects' behaviour could have been influenced by the use of a non-favoured browser.

2.1.4. *Web-browser studied*

Netscape Navigator and Microsoft Internet Explorer are now the dominant web-browsers. Netscape Navigator and Microsoft Internet Explorer had an estimated 10% and 84% share of web browser use in January 2001 (W3Schools.com 2001). The user interfaces to the current generation of web browsers have gone through several iterative refinements, and have been the topic of research-level scrutiny: for example, see Au & Li (1998). It is reasonable to suspect that the improved interfaces may have changed browser usage.

2.1.5. *Duration of the evaluations*

Catledge & Pitkow analysed three weeks of user interaction logs with XMosaic, and Tauscher & Greenberg analysed between five and six weeks. It is possible that long term web-page revisitation patterns will be missed even in these fairly long term analyses.

3. Method

The study reported in this paper aims to overcome the problems of the prior studies described in the previous section: it updates the results of the previous work; the subjects used their preferred browsers as normal while doing their everyday work; the browsers were current versions; and the study lasted a total of four months. We should foreshadow, however, that in overcoming these limitations we introduced new ones. In particular, unlike the prior studies we were unable to log which user interface elements were used to navigate to each page (link selection, Back button, history list, etc.) This and other limitations of the study are further reported in Section 6.

3.1. DATA SOURCE FILES

Under the Unix operating system, Netscape Navigator maintains a history file `history.dat` and a bookmark file `bookmarks.html` in a directory `.netscape` under the user's home directory. The history file keeps a list of the URLs the user has visited, the time of their last and first visit, the number of visits, and the title of each page. The history file is updated by Netscape whenever the user visits a page. The bookmark file holds all the bookmarked pages, an identifying label for each (which is extracted from the page's HTML Title tag, but can be replaced by the user), and the times at which the bookmark was added, last visited, and most recently changed. The structure of the bookmark file reflects the organisation of bookmarks into folders. The bookmark file is modified whenever the user accesses a bookmarked page, adds a page to the bookmarks, or modifies the bookmark structure using the "Edit Bookmarks" window.

3.2. PERMISSION AND PERIOD

We obtained permission from seventeen users to retrieve copies of their history and bookmark files from incremental backups. At our institution any file that is modified during the day is copied into an incremental backup. Therefore, if the user visited any pages using Netscape within a 24 hour period, there would be a copy of their `history.dat` file on the backup. Similarly, if they accessed a page in their bookmark collection, or if they modified their bookmark collection, then the backup would contain a copy of their `bookmarks.html` file.

Copies of the history and bookmark files were retrieved for a four month period (119 days), from early October 1999 to late January 2000. We asked for permission to gather the data *after* the terminating date of the study. There were, therefore, no dangers of "Hawthorne Effect" modifications to subject behaviour due to their awareness that their actions were being logged (Mayo 1933).

3.3. DATA EXTRACTION AND ASSUMPTIONS

The `history.dat` file is a Berkeley DB 1.85 Hash file (Olson, Bostic & Seltzer 1999) that contains six fields for each web page, as follows:

URL	the URL of the page;
Title	the HTML Title tag of the page (if any);
First	the time and date of the first page visit;
Last	the time and date of the most recent page visit;
Count	a count of how many times this URL has been visited;
Flag	a flag that shows whether the page was explicitly requested by the user rather than being part of another page (such as an image file that is part of another page).

A C program was used to extract the data. To aid repeatability of the study, it is necessary to state four normalisations and assumptions made in the data analysis

program [2] .

(1) Only pages with the *Flag* field set to 1 were included in the study. The history file includes data on many pages that the user has not explicitly requested. For instance, image files and java applets can be loaded by the browser as part of the page the user has requested. In terms of the user's action at the browser, we believed it would be incorrect to include these files within the set requested by the user.
(2) URLs involving search queries were truncated to remove the suffixes of the form `?name=value&name=value...` Thus, search queries were counted as visits to the same page: for example, separate searches for "cats" (`www.google.com/search?q=cats`) and "dogs" (`www.google.com/search?q=dogs`) using the google search engine would count as two visits to google. To confirm that this "cleaning" of URLs did not distort our results, we also ran our experiments with uncleaned URLs. The characterisations of web use resulting from the experiments with the "unclean" URLs were similar to those resulting from "clean" URLs.
(3) The *Count* fields for all pages were normalised to a zero value for the start of the study. Thus, we only counted page visits that occurred during the period of the study.
(4) Each page within a frameset is independently counted. If the user enters a frameset page `foo.bar/frameset.html` which refers to framed pages `foo.html` and `bar.html`, the log file will register one visit to each of the pages `frameset.html`, `foo.html`, and `bar.html` even through the user sees only one page in the browser. Implications of this limitation are discussed where appropriate in the results.

3.4. SUBJECTS

The seventeen unpaid volunteer subjects were all faculty (7), programming staff (3), tutors (3) or graduate students (4) in the Computer Science department. Although there is an obvious bias towards high levels of computing skills in our subject pool, this is consistent with the prior studies by Catledge & Pitkow and Tauscher & Greenberg. All subjects used their normal web browser (Netscape versions in the range from 4.5 to 4.7), running under either the Sun Solaris or Linux operating systems.

Subject 15 was employed as a web-master during the analysis period. In several of the data analyses reported below his patterns of web use are significantly different from the other subjects. These outlying data points are reported when they occur, but he was not excluded from the study because of the insights the data provides into "high-end" web use.

4. Results

In the results we use the term 'visit' to describe the act of displaying a page in the web browser, regardless of the technique used to navigate to the page: link selection,

[2] In an earlier version of this paper (McKenzie & Cockburn 2001), we incorrectly excluded pages with a variety of suffixes, including `.jpg` and `.gif`. All the data-analyses have been re-run in this paper, and the plots re-generated to remove this flaw.

'Back' button, bookmarks, etc. 'Visit count', then, describes the number of pages that a user (or group of users) has displayed in the browser. If a user navigates from page a to b and back to a, then this path results in a total visit count of three: two visits to a and one visit to b. We use the term 'vocabulary size' to describe the number of distinct URLs that a user (or group of users) has visited. Consequently, the a, b, a path gives a vocabulary size of two.

Note that revisiting a page does not necessarily mean that the user sees the same information displayed. Many popular pages, such as `cnn.com` are frequently updated, and a user who visits this page (and no others) once each day for fifty days will have a visit count of fifty, but a vocabulary size of one.

Over the 119 days of the study the seventeen subjects visited a total of 84841 pages, with a total vocabulary size of 17242 different URL addresses. The mean daily page visit count was approximately 42 pages for each user per day. This number provides a strong indication of the growth of web use since the earlier studies, which had approximate daily visit count means of fourteen (Catledge & Pitkow 1995) and twenty one (Tauscher & Greenberg 1997).

4.1. VISIT COUNTS AND VOCABULARY SIZES

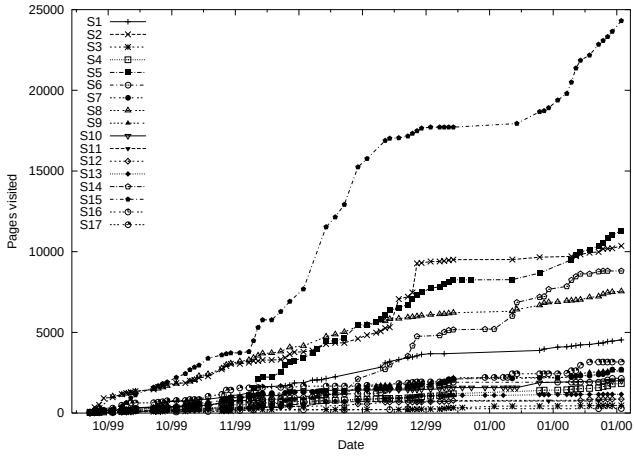
Figure 1(a) shows, for each subject, the increase in the total number of pages visited during the 119 days of the study. Each data point on a user's line in the graph indicates that a backup file for that day was available, meaning that the web had been used that day. The mean total visit count by each subject during the study is 4991 (σ =6106), with a minimum visit count of 281 (subject 16) and a maximum of 24309 (subject 15). Per-subject visit counts are shown on row 1 of Table 1. Subject 15's visitation rate is a clear outlier (more than two standard deviations from the mean), and removing his data from the analysis produces a mean visit count of 3783 (σ =3652).

Figure 1(b) shows the increase in each user's URL vocabulary (the number of distinct URLs visited) over the study. The mean per-subject final vocabulary size is 1227 (σ =1086), with a range from 74 to 4251. Each subject's vocabulary size is shown on row 2 of Table 1. The figure shows that the rate of increase in vocabulary size is fairly constant over the period of the study. Closer inspection of the graph, however, shows periods of "exploration" where the vocabulary rapidly increases, and periods where the vocabulary grows little despite regular web use.

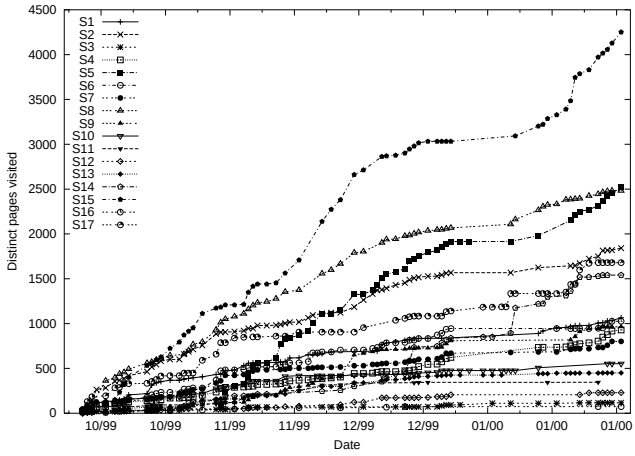
The visit counts and vocabulary sizes reported here will be larger than the actual number of pages displayed in each user's browser. This over-estimation is caused by separately counting each frame-page within a frameset (see Section 3.3).

4.2. CORRELATING VISIT COUNTS WITH VOCABULARY SIZES

We analysed the relationship between the growth of each user's visit count and their vocabulary size (see Figure 2, which plots visit counts against vocabulary sizes for the subjects that visited more than 2000 pages). Although the observed periods of rapid vocabulary growth might have implied that visit counts and vocabulary sizes would grow relatively independently of each other, linear correlation and regression show a close relationship between the variables: per-subject linear regression R-squared values



(a) Total visit counts by time.



(b) Total vocabulary size by time.

FIGURE 1. Total number of pages visited and total vocabulary over time (mm/yy) for each user.

ranged from 0.9 to 0.999, and all p values $< .0001$ (see rows 7 to 9 of Table 1 for a summary). Slopes for the linear regression "line of best fit" range from 2.0 (subject 4) to 6.5 (subject 2). Linear regression over all subjects gives a slope of 5.083 and an R-squared value of 0.8837 ($F(1,940) = 7140$, $p < 0.0001$). This overall slope value reflects the revisitation rate for the subject pool: for each new URL added to the overall vocabulary, four pages are revisited.

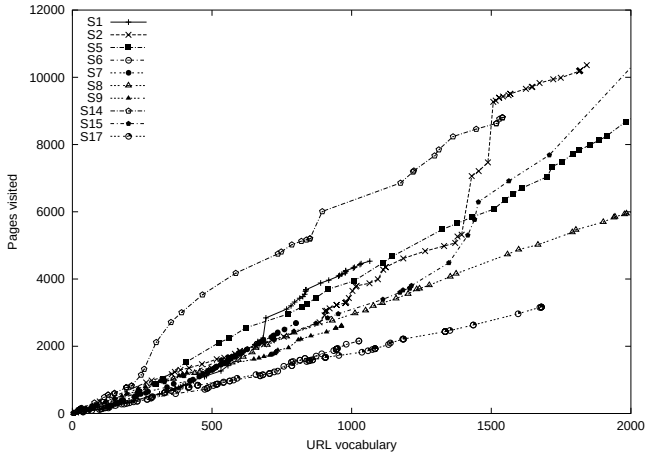


FIGURE 2. Truncated plot of total pages visited against URL vocabulary.

4.3. REVISITATION RATES

Previous studies have shown that revisitation (navigating to a previously visited page) accounts for 58% and 61% of all page visits. Our study shows that page revisitation is now even more prevalent, accounting for 81% of page visits when calculated across all users. This revisitation rate is calculated according to the formula used in Tauscher & Greenberg (1997):

$$R = \frac{total\ visit\ count - total\ vocabulary\ size}{total\ visit\ count} * 100$$

Per-subject revisitation rates ranged from a minimum of 61% (subject 4) to a maximum of 92% (subject 15), with a mean of 81% (σ =10). Row 4 of Table 1 shows per-subject revisitation rates. Interestingly, even our subject with the lowest revisitation rate was revisiting pages more frequently than the overall mean of the prior studies.

4.4. DOMINANCE OF FAVOURED PAGES

Almost all the subjects had one or two pages that they visited far more often than any other page. This pattern of behaviour substantially contributed to the high revisitation rates reported in the previous section. Subject 2, for instance, had visit counts of 4352, 384, 199, and 117 for his top four pages. Rows 10 to 19 of Table 1 show the visit counts for the top five pages for each user. The top three pages accounted for a minimum of 8.9% of page visits (subject 17), and a maximum of 48% (subject 2), with a mean of 24% (σ =13.7).

Given the extremely high visit counts to each user's favourite pages, it is useful to investigate the interface techniques that allow them to visit those pages. Netscape Navigator supports a variety of techniques that allow short-cut access to pages. These include the configurable 'Home' button, bookmarks, and the 'personal toolbar'. Rows

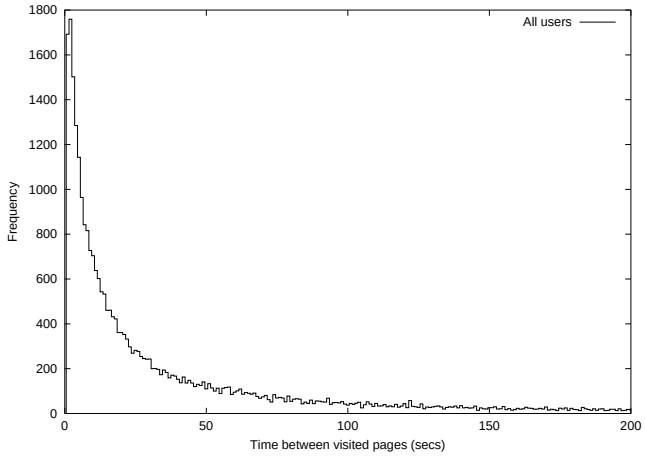


FIGURE 3. The number of pairs of pages visited within certain time gaps (number of pairs on the y-axis, time gaps on the x-axis).

11, 13, 15, 17 and 19 of Table 1 use the following symbols to encode shortcut techniques that the user could use to access the page (efficiency, here, is defined in terms of visibility and availability of interface activators that will cause the browser to navigate to the page): ❁—the page is set as the user's home page; ☛—the page is in the user's bookmark collection; ➢—the page is in the user's 'Personal Toolbar'; ✗—the page is not in any of the above categories.

Although most users had shortcuts to their top two pages (rows 11 and 13), few had a shortcut scheme for reaching their third, fourth and fifth most frequently visited pages (rows 15, 17 and 19). Of course, users may have created other shortcut schemes that we were unable to detect from the logs. In particular, one common practice that we are unable to analyse is the use of 'handy links' HTML pages that many users create to help them quickly access pages.

4.5. TEMPORAL ASPECTS OF PAGE VISITS

The results show that browsing is rapidly interactive. Users often visit several pages within very short periods of time, implying that many (or most) pages are only displayed in the browser for a short period of time. Figure 3 shows that the most frequently occurring time gap between subsequent page visits was approximately one second, and that gaps of more than ten seconds were relatively rare.

This result was calculated by taking the set of URLs in each user's daily history file and sorting them by the *Last* field: for example, the sorted file might show that at 3pm exactly (15:00:00) the user visited page x, followed by *page* y one second later (15:00:01), and page z at 15:01:05.

There are two weaknesses in our technique for establishing temporal gaps between pages. The first weakness tends to over-estimate time gaps (providing an upper-bound for gaps) because only one *Last* entry is recorded for each page that is visited many times in the day. Thus, if the user visits page a at 11:45:00, page b at 11:45:10, page c at 11:45:20, then revisits page b at 11:46:00, the log will only record one visit to

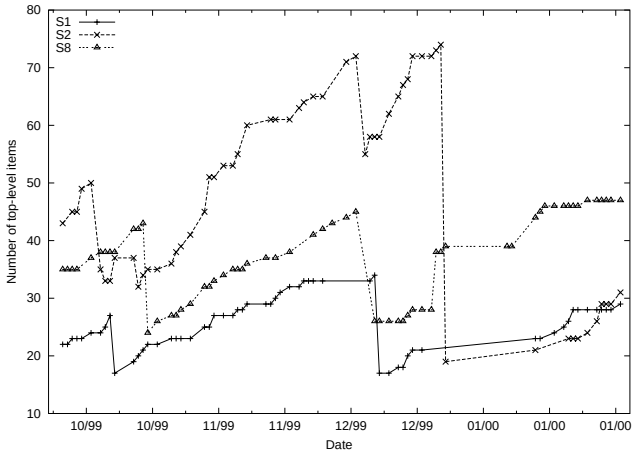


FIGURE 4. URLs and folders in the top-level bookmark file by time.

page b at 11:46:00. This will lead to the incorrect result of a gap of twenty seconds between pages a and c, rather than two ten second gaps between a and b and then b and c. We logged a total of 36892 gaps across the seventeen subjects, meaning that we detected 44% of the 84824 gaps that actually occurred between the pages visited by the subject pool (the total visit count of 84841 minus one page for each subject).

The second weakness tends to under-estimate time-gaps. It is caused by separately registering each frame-page within a frameset (see Section 3.3). For each frameset page in the log file, the logs will show that each of the associated frame-pages were visited almost instantaneously.

The important result is that browsing is a *rapidly* interactive activity—even more so than our conservative results indicate—as demonstrated by Figure 3. There are two possible browsing strategies that would explain the short gaps between pages. The first is that users rapidly follow links when navigating between pages within one browser. This would imply that many (or most) pages are simply used as routes to other pages; the users are following known trails through a series of links that are displayed at known locations on the pages. The second explanation is that users frequently use secondary windows to pop-up a series of 'spoke' pages that are linked off a single 'hub' page; clicking on a link with the middle mouse-button in Netscape Navigator causes the linked page to be displayed in a new window, and shift clicking with Internet Explorer has the same effect. Using this technique, a user could 'visit' several pages, each displayed in a separate window within one or two seconds. We suspect that both of these explanations are valid, but that the first (rapid navigation within one window) is the primary cause of the effect shown in the data.

4.6. BOOKMARKS

We analysed the contents of each subject's bookmarks file. There was a wide range of bookmark usage patterns, from subject 16 who did not use them at all, through to subject 8 who had a maximum of 587 bookmarks: see row 20 of Table 1. The mean

maximum size of the subjects' bookmarks collection was 184 (σ =166.15). The mean number of folders used to store bookmarks was 18.1 (σ =16.5): see row 21 of Table 1.

On analysing the changes in each subject's bookmark collection over time, we found that the rate of bookmark addition heavily outweighed the rate of deletion. Rows 24 and 25 of Table 1 show the number of bookmarks added and deleted for each user, giving means of 27.6 (σ = 29.7) and 3.7 (σ =5.2).

The imbalance between the rates of bookmark addition and deletion implies that users have (or will have) problems managing the size and organisation of their bookmark collections. Bookmarks in Netscape are normally selected via a pop-up cascading menu, the length of which depends on the number of "top level" items in the bookmark file. Row 23 of Table 1 shows that our subjects had up to 130 items in this top-level; a number that is certain to produce an unwieldy menu. Figure 4 shows the number of items in the top-level for three of the subjects, plotted over time. The obvious steps show how the users would periodically re-organise their bookmark file structure to overcome the problem of the menu growing too long (this effect was also noted by Abrams et al. (1998)). Our analysis shows that when re-organising bookmarks, rather than deleting items, subjects would typically relocate them to new folders (see row 25 of Table 1). We also found that twelve of the subjects had duplicate bookmark entries that they were presumably unaware of. On average, approximately 5% of bookmarks were duplicates, with subject 2 having 28 duplicates.

Web sites and pages are relatively transient, yet the low rate of deletion indicates that bookmark collections continually grow. Two months after collecting the bookmark data, we ran scripts that attempted to access each page in the subjects' bookmark collections. Any page returning 404 "Not found", 301 "Moved Permanently", or 5xx (host unavailable) was deemed invalid. Over all subjects, approximately 25% of bookmarked pages were invalid. The percentage of valid bookmarks for each user are shown on row 27 of Table 1.

4.7. A COMMUNITY OF USERS?

With the growth of research interest in systems for 'collaborative information filtering' and 'recommender systems' (for example, (Goldberg, Nichols, Oki & Terry 1992, Wexelblat & Maes 1999, Shardanand & Maes 1995)) it seemed potentially useful to review the degree of overlap between the browsing patterns of the subjects. Our subjects all worked or studied within the same department.

First, we compared the percentages of each subject's web-page visits that were made within the department's web-site, within the parent institution, within the country, and overseas. We found a relatively high degree of uniformity across the subjects: the mean percentage of web-page accesses made within the department was 58% (σ =19%), within the parent institution was 6.5% (σ =5.2%), within the country was 5.3% (σ =4.5%), and international was 30% (σ =18.4%). These values are summarised on rows 28 to 30 of Table 1.

Closer analysis, however, reveals that the subjects were not visiting similar areas in the web. For each page in the total URL vocabulary of 17242 distinct URLs visited by the subjects, we counted how many subjects had visited it. Ninety one percent of the URLs (14734) had been visited by at most one of the subjects: that is, only 9.2% had

been seen by more than one subject. No page had been visited by all the subjects, but one (the University's home page) had been visited by all but one subject. A total of 732 pages had been visited by three or more subjects, and only 89 pages were visited by eight or more subjects.

These results show that there was a surprising lack of overlap in the pages visited by this fairly homogeneous community of users.

4.8. ABSENT TITLES

Five percent of the distinct URLs visited by the subjects did not have an HTML "Title" tag associated with the page.

Titles are used by Netscape and Microsoft Internet Explorer in a variety of ways, including labelling items on the "Back" pull down menu, default identification tags in the bookmark and history lists, and labelling the window-manager border. Missing, incorrect, and inconsistent titles can frustrate the user's ability to identify pages that they wish to return to (Cockburn & Greenberg 2000). Although there are alternative interface techniques that could be used to aid page identification (such as thumbnail images of the page (Cockburn, Greenberg, McKenzie, JasonSmith & Kaasten 1999, Robertson, Czerwinski, Larson, Robbins, Thiel & van Dantzich 1998)), it is likely that text titles will remain an important page identification cue. There is little that any browser can do to ensure the presence and accuracy of the text titles, but the software used to author web-pages could improve the techniques used to prompt page-designers to carefully consider the titles they assign to pages. Often, page-titling capabilities are 'hidden' under sub-menus or dialogue boxes, rather than prominently displayed for each newly authored page.

5. Implications for Design

Studies such as this help us understand the scale and nature of web use. Several research and development strands can potentially benefit from this improved understanding.

5.1. HISTORY AND REVISITATION TOOLS

According to the results, approximately 81% of the URLs that a user visits are revisitations. This result substantially exceeds the previously reported values of 58% and 61%. From these high values it is clear that any minor interface inefficiency in supporting revisitation will result in massive productivity losses when multiplied across millions of users.

Current commercial browsers support a wide array of facilities for helping users return to pages. These include the 'Back' button used to return to recently visited pages, bookmarks (or 'favourites') used for frequently visited pages, history lists for more temporally distant pages, and a variety of other schemes. Each of these mechanisms would independently benefit from HCI research on the support they provide, but possibly greatest user benefits would be provided by integrating the diverse revisitation interfaces into a single interface component, thus eliminating the need for the user to learn multiple interfaces.

		s1	s2	s3	s4	s4	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17
Visits, vocabulary and revisitation																		
1	Visit count	4531	10361	471	1833	11299	2156	2694	7542	2613	2000	753	863	1163	8806	24309	281	3166
2	URL vocabulary	1065	1842	116	929	2519	1026	801	2481	964	553	346	229	450	1540	4251	74	1679
3	Pages visited once	644	980	62	715	1199	741	485	1294	561	351	258	119	272	777	1879	38	1169
4	Revisitation rate(%)	85.8	90.5	86.8	61.0	89.4	65.6	82.0	82.8	78.5	82.5	65.7	86.2	76.6	91.2	92.3	86.5	63.1
Visits to search pages																		
5	Count	186	411	6	225	432	191	91	845	281	234	50	140	42	448	721	0	309
6	Percentage of visits	4.1	4.0	1.3	12.3	3.8	8.9	3.4	11.2	10.8	11.7	6.6	16.2	3.6	5.0	3.0	0	9.8
Linear regression of visit count with vocabulary																		
7	Slope	5.03	6.46	4.26	2.03	4.61	2.115	3.45	3.01	2.57	3.80	2.17	3.60	2.70	5.93	6.36	4.33	1.87
8	R-Squared	0.96	0.90	0.98	0.997	0.997	0.98	0.98	0.999	0.99	0.98	0.98	0.97	0.996	0.99	0.99	0.97	0.998
9	p	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Visits to top five pages (☸= homepage, ☛= bookmarked, ➢= personal toolbar, ✗= no special access)																		
10	Count for # 1	820	4352	98	218	1275	166	366	493	186	660	134	128	67	434	1494	57	164
11	Access to # 1	☸	☸	☛	☸	☸	✗	☸	☸	✗	✗	✗	✗	☛	☸	✗	✗	✗
12	Count for # 2	549	384	84	74	445	153	82	199	118	95	44	125	30	420	1214	37	75
13	Access to # 2	➢	☛	☸	☛	✗	☸	➢	➢	✗	☸	☸	✗	✗	✗	☸	✗	✗
14	Count for # 3	106	199	20	36	158	43	61	172	57	77	19	102	28	220	484	23	42
15	Access to # 3	☛	✗	✗	✗	✗	✗	✗	☛	✗	✗	➢	✗	☸	✗	➢	☸	✗
16	Count for # 4	95	117	14	22	147	38	51	96	49	40	16	30	22	158	398	7	41
17	Access to # 4	☛	✗	✗	✗	✗	✗	✗	✗	☸	☛	✗	☛	✗	✗	✗	✗	✗
18	Count for # 5	58	115	10	19	144	20	51	75	31	27	14	11	18	154	218	7	35
19	Access to # 5	✗	✗	✗	✗	✗	✗	✗	✗	☛	✗	✗	✗	✗	✗	✗	✗	✗
Bookmark statistics																		
20	Max. num. bookmarks	165	565	107	272	189	94	247	587	188	76	86	112	25	89	84	0	242
21	Max. num. folders	16	58	12	19	19	3	25	52	36	13	8	12	1	6	17	0	11
22	Mean URLs / folder	9.6	9.5	8.8	13.2	10.6	30.6	9.8	11.0	4.6	5.6	11.8	8.9	20.6	16.8	4.8	0.0	19.5
23	Max. top level items	34	74	21	59	32	90	44	47	30	22	38	27	20	37	2	0	130
24	Num bookmarks added	37	108	3	33	50	3	5	48	65	6	19	7	6	10	14	0	56
25	Num bookmarks moved	21	147	0	0	58	0	0	56	8	0	38	0	0	20	2	0	3
26	Num. bookmarks deleted	11	9	0	2	10	0	0	17	6	1	0	0	0	5	2	0	0
27	Percentage valid	77.8	75.0	99.0	73.1	79.2	76.3	70.3	61.9	90.1	64.3	80.2	100.0	79.2	82.7	87.3		82.3
Page locations: percentages of page visits																		
28	Department	46.7	68.1	79.6	53.0	53.7	63.0	71.0	41.1	45.0	66.8	67.4	51.3	49.0	65.3	73.3	90.0	5.4
29	Organisation	4.8	4.1	2.9	7.1	8.5	12.7	10.2	3.8	1.7	1.4	3.3	21.6	10.1	3.7	2.8	9.3	2.2
30	International	31.1	26.1	17.5	36.9	34.0	12.3	14.8	50.0	49.9	28.7	21.9	23.4	31.8	28.7	20.5	0.0	82.4

TABLE 1Summary of the data values retrieved for each subject.

5.1.1. *The 'Back' button*

The prevalence of revisitation calls the stack-based behaviour of the 'Back' button into question. 'Back' removes recently seen pages from the set of accessible pages through its 'stack-pruning' implementation (Cockburn & Jones 1996). Greenberg & Cockburn (1999) describe a variety of alternative implementations of the 'Back' button that do not prune recently visited pages, including a true and complete temporal history of the pages that the user has seen in the browser.

In our on-going work, we will conduct a comparative evaluation of the traditional 'stack-based' behaviour of the 'Back' button and our 'temporally-based' behaviour.

5.1.2. *History Lists*

Netscape Navigator and Microsoft Internet Explorer both support history lists that allow the user to investigate the complete history of pages prior to a temporally distant point at which entries expire from the list. The techniques used by these systems are substantially different. Netscape supports a textual list that can be sorted by one of several dimensions: alphabetical list of URL or title, visit count, time of first or last visit, and so on. Internet Explorer, in contrast, uses a 'temporal chunks' mechanism that allows pages to be recalled based firstly on their temporal distance from the current time ('yesterday', 'last week', etc.), subsequently by an alphabetical list of web-sites, and finally by an alphabetical list of page titles.

It is unclear which of these techniques is more effective, and we currently have a student investigating this issue.

5.1.3. *Visual Histories*

Many systems have implemented visual histories: from early systems such as MosaicG (Ayers & Stasko 1995) and WebNet (Cockburn & Jones 1996) through to recent systems such as Footprints (Wexelblat & Maes 1999), WebView (Cockburn et al. 1999), and PadPrints (Hightower, Ring, Helfman, Bederson & Hollan 1998). With the exception of Footprints and PadPrints, the major limitation of the work on web revisitation schemes is the lack of empirical evaluation. Cockburn & Greenberg (2000) provides a review and discussion of issues in the usability of visual histories.

5.2. BOOKMARKING TOOLS

Having noted that bookmark maintenance is one of the top three usability problems on the web (Pitkow 1996), several research projects are investigating new styles of bookmarking interfaces: for example, the "WWW Dynamic Bookmark" (Takano & Winograd 1998) and the Data Mountain (Robertson et al. 1998).

Our study reveals that users build very large bookmark collections, and that the current interface schemes tend to become unwieldy (producing extremely long menus), forcing users to re-organise their bookmark structure. The results also showed that approximately a quarter of bookmarks are invalid. Finally, the results indicated that users tend to have one or two pages that are visited far more often than all other pages.

There are three clear design implications. Firstly, bookmark collection systems should be sufficiently scalable to manage large collections. The Data Mountain, for instance, has been shown to be effective for a collection of 100 pages, but may have difficulty scaling to much larger data sets. Secondly, bookmark collection systems should include tools that assist users in managing their collections, particularly in identifying invalid bookmarks. Thirdly, systems should support shortcut mechanisms (such as the "Personal Toolbar Folder") for efficiently navigating to a small set of frequently visited pages.

5.3. LOOK-AHEAD NAVIGATION

Given the high revisitation rates and rapid navigation patterns revealed by our data, it seems reasonable to suspect that users are often repeating common trails in order to reach a common destination point. The single *Last* data point for each page in the log files means that we are unable to effectively search for statistical data on the frequencies of these occurrences, but anecdotal evidence supports the claim. Many of our colleagues report paths typified by the following statement:

> I've never bookmarked the library's search page. I keep forgetting because once I'm there I start my search rather than thinking to bookmark it. Anyway, I've got a good shortcut. First, I click 'Home' which takes me to the Department's homepage, then I click on the link to the University's homepage, and from there I click on 'Departments' and then 'Libraries'. It takes quite a few clicks, but it doesn't take too long.

Through statistical analysis of each user's navigation patterns, it should be possible for next-generation browsers to predict likely destinations in a similar manner to that used by the 'Reactive Keyboard' (Darragh & Witten 1992).

To a small extent, current browsers are already moving in this direction, as demonstrated by Microsoft Internet Explorer's 'AutoComplete' feature which offers to complete partially typed URLs with those previously visited by the user.

5.4. PAGE AND SITE DESIGN

The results show that users spend a very short period of time at most pages. This rapid navigation behaviour indicates that most pages should be designed to load quickly, and to clearly present their links to the user. This property of browsing provides supporting evidence to Nielsen's web design guidelines such as "Scannability" and "Keep your texts short" (Nielsen 2000). Advanced web page features such as Macromedia Flash and java applets (which have a relatively high loading and start-up time) should be reserved for pages that the designer expects users to peruse for long periods.

An alternative interpretation of the rapid page navigation result is that web-sites should be designed to shorten the navigation paths to popular designation pages.

Good page design, with appropriate use of navigational shortcuts, should minimise the number of transient pages used.

6. Limitations of the study

6.1. LOGS VERSUS EVENTS

The analysis in this paper updates and extends prior client-side log analyses of web-use (Catledge & Pitkow 1995, Tauscher & Greenberg 1997). However, the technique we used to gather the data—file analysis from incremental backups—is different from that of the prior studies, which used low-level logs of the actual user events (button clicks, etc.) executed at the browser. Both techniques have their own strengths and weaknesses. The primary weakness of our technique is that we cannot determine which interface event caused a particular page to be accessed. The previous studies were able to report, for instance, that use of the 'Back' button accounted for approximately 40% of user actions at the browser. The primary strength of our technique, however, lies in our ability to gather data about the user's browsing activities without changing, in any way, their browsing environment. One of the primary limitations of the previous studies was that the users were not using their preferred web-browsers.

6.2. SUBJECT GROUP

Like our own study, the previous studies have relied heavily on subjects who work and study in Computer Science departments. Clearly this is not a representative sample of web users. However, the fact that all the studies have used the same subject group provides uniformity in the subject pool, and increases the likelihood that the observed changes in browsing behaviour are real rather than arising from cultural and social differences in the subject pool.

6.3. CHICKEN AND EGG PROBLEM

Our study, like those before it, characterises what users *do* with the web. It is not clear, however, to what extent their activities are determined by limitations of the browser rather than by their actual desires. Dynamic observational studies, such as that by Byrne et al. (1999), are necessary to clarify the mapping between the users' tasks and the support provided by browsers.

7. Conclusions

This study updates and extends the empirical foundation for understanding web use. The study shows both expected and unexpected results when compared to earlier studies carried out in 1994 and 1995. As expected, users are daily visiting many more pages than earlier studies: approximately three times more than 1994 and twice as many as 1995. More surprisingly, the revisitation rate has increased from approximately 60% to 81%; on average, users have previously seen four out of five pages that they visit. The results also show that web-use is rapidly interactive, with users visiting many pages within seconds of each other. Additionally, there is evidence that users keep track of large numbers of bookmarks, that they seldom delete items from these

collections, and that a relatively high percentage of bookmarks are invalid. Many other statistical characterisations of web use are also reported.

Our future work will continue to analyse web use. We will use this information as input to the design of web revisitation tools that integrate and extend the wide variety of schemes currently supported by commercial browsers.

Acknowledgements

This research was aided by an equipment grant-in-aid from Microsoft Research. Thanks to Steve Jones, Michael JasonSmith and the anonymous reviewers for helpful comments on the paper.

References

Abrams, D., Baecker, R. & Chignell, M. (1998), Information archiving with bookmarks: Personal web space construction and organization, *in* 'Proceedings of CHI'98 Conference on Human Factors in Computing Systems Los Angeles, April 18–23', pp. 41–48.

Au, I. & Li, S. (1998), Netscape communicator's collapsible toolbars, *in* 'Proceedings of CHI'98 Conference on Human Factors in Computing Systems Los Angeles, April 18–23', ACM Press, pp. 81–86.

Ayers, E. & Stasko, J. (1995), Using graphic history in browsing the world wide web, *in* 'Proceedings of the Fourth International World Wide Web Conference. 11–14 December, Boston'.

Bellotti, V. & Rogers, Y. (1997), From web press to web pressure: Multimedia representations and multimedia publishing, *in* 'Proceedings of the ACM SIGCHI'97 Conference on Human Factors in Computing Systems, Atlanta, Georgia, March 22-27', pp. 279–286.

Byrne, M., John, B., Wehrle, N. & Crow, D. (1999), The tangled web we wove: A taskonomy of WWW use, *in* 'Proceedings of CHI'99 Conference on Human Factors in Computing Systems Pittsburgh, May 15–20', pp. 544–551.

Catledge, L. & Pitkow, J. (1995), Characterizing browsing strategies in the world wide web, *in* 'Computer Systems and ISDN Systems: Proceedings of the Third International World Wide Web Conference. 10–14 April, Darmstadt, Germany', Vol. 27, pp. 1065–1073.

Chi, E., Pirolli, P. & Pitkow, J. (2000), The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site, *in* 'Proceedings of CHI'2000 Conference on Human Factors in Computing Systems The Hague, The Netherlands, April 1–6', pp. 161–168.

Cockburn, A. & Greenberg, S. (2000), 'Issues of page representation and organisation in web browser's revisitation tools', *Australian Journal of Information Systems* 7(2), 120–127.

Cockburn, A., Greenberg, S., McKenzie, B., JasonSmith, M. & Kaasten, S. (1999), Webview: A graphical aid for revisiting web pages, *in* 'Proceedings of the 1999 Computer Human Interaction Specialist Interest Group of the Ergonomics Society of Australia (OzCHI'99). November 28–30 Wagga Wagga.', pp. 15–22.

Cockburn, A. & Jones, S. (1996), 'Which way now? Analysing and easing inadequacies in WWW navigation', *International Journal of Human-Computer Studies* 45.(1), 105–129.

Darragh, J. & Witten, I. (1992), *The Reactive Keyboard*, Cambridge University Press.

Goldberg, D., Nichols, D., Oki, B. & Terry, D. (1992), 'Using collaborative filtering to weave an information tapestry', *Communications of the ACM* 35(12), 61–70.

Greenberg, S. & Cockburn, A. (1999), Getting back to back: Alternate behaviors for a web browser's back button, *in* '5th Conference on Human Factors and the Web, Gaithersburg, Maryland. June 3.'. `http://zing.ncsl.nist.gov/hfweb/`

Hightower, R., Ring, L., Helfman, J., Bederson, B. & Hollan, J. (1998), Graphical multiscale web histories: A study of padprints, *in* 'Proceedings of the 1998 ACM Conference on Hypertext, June 20–24. Pittsburgh, Pennsylvania.', ACM Press, pp. 58–65.

Kehoe, C. & Pitkow, J. (1996), 'Surveying the territory: Gvu's five www user surveys', *The World Wide Web Journal* 1(3), 77–84.

Mayo, E. (1933), *The Human Problems of an Industrial Civilization*, Cambridge, MA: Harvard University Press.

McKenzie, B. & Cockburn, A. (2001), An empirical analysis of web page revisitation, *in* 'Proceedings of the 34th Hawaiian International Conference on System Sciences, HICSS34, Maui, Hawaiii, (CD ROM)', IEEE Computer Society Press.

Nielsen, J. (2000), *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing.

Olson, M., Bostic, K. & Seltzer, M. (1999), Berkeley db, *in* 'Proceedings of the FREENIX Track: 1999 USENIX Annual Technical Conference. Monterey, California. 6–11 June'. `http://www.usenix.org/events/usenix99/full_papers/olson/olson.pdf`

Pirolli, P., Pitkow, J. & Rao, R. (1996), Silk from a sow's ear: Extracting usable structures from the web, *in* 'Proceedings of CHI'96 Conference on Human Factors in Computing Systems Vancouver, April 13–18', pp. 118–125.

Pitkow, J. (1996), 'Gvu's www user surveys', WWW page: `http://www.cc.gatech.edu/gvu/user_surveys/survey-04-1996/`

Robertson, G., Czerwinski, M., Larson, K., Robbins, D., Thiel, D. & van Dantzich, M. (1998), Data mountain: Using spatial memory for document management, *in* 'Proceedings of the 1998 ACM Conference on User Interface Software and Technology, November 1–4. San Francisco, California.', ACM Press, pp. 153–162.

Shardanand, U. & Maes, P. (1995), Social information filtering: Algorithms for automating 'word of mouth', *in* 'Proceedings of CHI'95 Conference on Human Factors in Computing Systems Denver, May 7–11', pp. 210–217.

Takano, H. & Winograd, T. (1998), Dynamic bookmarks for the WWW, *in* 'Proceedings of the 1998 ACM Conference on Hypertext, June 20–24. Pittsburgh, Pennsylvania.', ACM Press, pp. 297–298.

Tauscher, L. & Greenberg, S. (1997), 'How people revisit web pages: Empirical findings and implications for the design of history systems', *International Journal of Human Computer Studies, Special issue on World Wide Web Usability* 47(1), 97–138.

W3Schools.com (2001), 'Browser statistics', `http://www.w3schools.com/browsers/browsers_stats.asp`

Wexelblat, A. & Maes, P. (1999), Footprints: History-rich tools for information foraging, *in* 'Proceedings of CHI'99 Conference on Human Factors in Computing Systems Pittsburgh, May 15–20', pp. 270–277.