

An Empirical Analysis of Web Page Revisitation

Bruce McKenzie and Andy Cockburn

Department of Computer Science
University of Canterbury
Christchurch, New Zealand

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

Abstract

There is a surprising lack of empirical research into user interaction with the web. This paper reports the results of an analysis of four months of logged data describing web use. The results update and extend earlier studies carried out in 1994 and 1995. We found that web page revisitation is a much more prevalent activity than previously reported (approximately 80% of pages have been previously visited by the user), that most pages are visited for a surprisingly short period of time, and that users maintain large (and possibly overwhelming) bookmark collections.

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.

Two prior studies provide the empirical foundation for our current understanding of user interaction with the web: Catledge & Pitkow (1995) and Tauscher & Greenberg (1997). Though excellent studies, there are several reasons for suspecting that their findings may no longer reflect current use of the web:

1. *Age of the studies.* The 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. *Web-browser studied.* Both studies analysed use of NCSA's XMosaic browser. Netscape Navigator and Microsoft Internet Explorer are now the dominant web-browsers. Netscape had an estimated 45% share of web-browser use in 1998, and it has been projected that Microsoft Internet Explorer will have a 65% share by 2001 (Schmalensee 1999). 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.

3. *Browsers of preference.* Tauscher & Greenberg state that none of the subjects in their study used XMosaic as their normal web-browser. Similarly, Catledge & Pitkow indicate that subjects may have chosen to use a browser other than their specially equipped version of XMosaic. Clearly, the subjects' behaviour could have been influenced by the use of a non-favoured browser.

4. *Duration of the evaluation.* 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 in even these fairly long term analyses.

This paper aims to update and overcome some of the limitations of prior empirical investigations into how the web is used. The study presents the results of an analysis of four months of daily client-side log files. The files showed the pages each user visited, the number of times

they visited them, the timing of page visits, and changes to the user's bookmark collection.

Our primary motivation for analysing web use is to provide contextual information for the design of next-generation web-browsing interfaces. There are, however, several other areas that may benefit from an improved understanding of web browsing activities. These include the design of caching proxy servers, search engines, collaborative information systems, and web-pages.

The structure of the paper is as follows. Section 2 reviews the findings of prior research on web navigation. Section 3 details our experimental method, and Section 4 provides the analysis and results of the study. Limitations of this study and implications of our findings for browser and web-page design are discussed in Section 5. Section 6 concludes the paper.

2 Prior Work

Catledge & Pitkow's (1995) study of web use involved one hundred and seven users who were staff, faculty and students in Georgia Institute of Technology's Computing Department. In three weeks, 31134 navigation acts 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 twenty three subjects who were also staff, faculty and students in a Computer Science department. As in Catledge & Pitkow's study, the web-browser used was XMosaic. Approximately 19000 navigation acts were logged during a five to six week period, giving a mean page visit rate of around twenty one 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 pre-

vious 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 two percent 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 11 bookmarks. Indeed, Pitkow (1996) reported from a survey of 6619 users that managing bookmark collections is one of the top three usability problems of the web.

3 Method

Under the Unix operating system, Netscape Navigator and Communicator maintain 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 of 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.

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 the incremental backup.

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 Ef-

fect” modifications to subject behaviour due to their awareness that their actions were being logged (Mayo 1933).

3.1 Data Extraction

We wrote a C program to extract the data from the “history.dat” file (a Berkeley DB 1.85 Hash file). The available fields in the file are as follows:

1. *URL* — the URL of the page;
2. *Title* — the HTML Title tag of the page (if any);
3. *First* — the time and date of the first page visit;
4. *Last* — the time and date of the most recent page visit;
5. *Count* — a count of how many times this URL has been visited;
6. *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).

To aid repeatability of our study, it is necessary to state the normalisations and assumptions that we made in our data analysis program.

Firstly, 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 javascript classes 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.

Secondly, we removed pages where the URL had the following suffixes: *.xbm*, *.pfr*, *.class*, *.tmp*, *.js*, *.rdf*, *.mcf*, and *.mco*, *.gif*, *.jpg* and *.jpeg*.

Thirdly, we truncated URLs involving search queries 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 charac-

terisations of web use resulting from the experiments with the “unclean” URLs were similar to those resulting from “clean” URLs.

Fourthly, URLs that are identical except for the trailing slash were treated as identical pages: for example, <http://...xxx> and <http:...xxx/> were treated as the same page. This modification is easy to justify: the user sees the same page in either case.

Finally, 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.

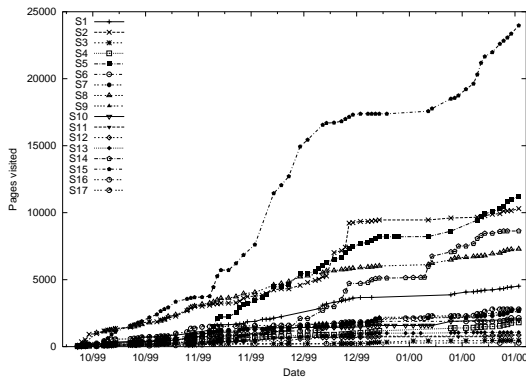
3.2 Subjects

The seventeen unpaid 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.

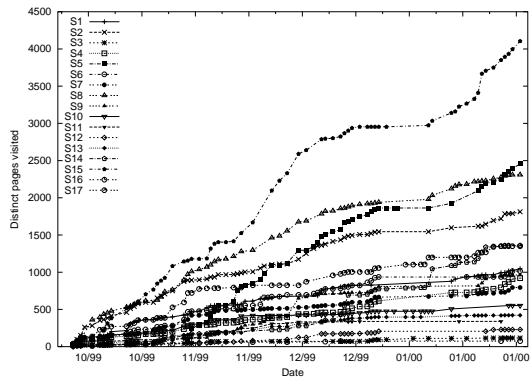
Subject 15 was employed as a web-master during the analysis period. In several of the data analyses reported below his patterns of web use are significantly different from the other subjects. These outlying data points are reported when they occur, but he was not excluded from the study because of the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

4 Results

Over the 119 days of the study the seventeen subjects visited a total of 83411 pages at 16290 different URL addresses. The mean daily page visit rate was approximately forty one pages for each user per day. Although this number does not accurately characterise each individual’s web use, it provides a strong indication of the growth of web use since the earlier studies, which had approximate daily page rate means of fourteen (Catledge & Pitkow 1995) and twenty one (Tauscher & Greenberg 1997).



(a) Total visits by time.



(b) Total vocabulary by time.

Figure 1: Total number of pages visited and total vocabulary over time for each user.

4.1 Vocabulary sizes and visit counts

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 indicates that a backup file for that day was available, meaning that the web had been used that day. The mean total number of page visits by each subject during the study is 4906 (s.d.=6032), with a minimum visit count of 281 (subject 16) and a maximum of 23973 (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 3714 (s.d.=3615).

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 1169 (s.d.=1035), with a range from 74 to 4105. 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.

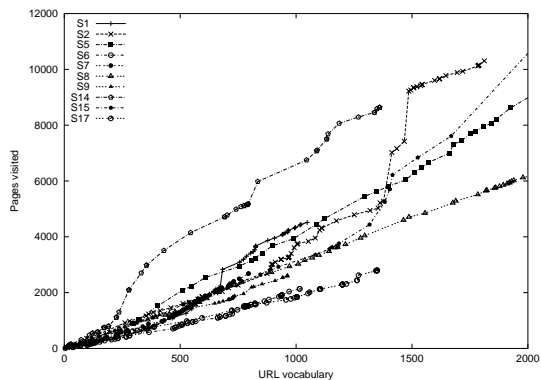


Figure 2: Truncated plot of total pages visited against URL vocabulary.

4.2 Vocabulary sizes and visit counts

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 size for the subjects that visited more than 2000 pages). Although the observed periods of rapid vocabulary growth might have implied that vocabulary and visit counts would grow relatively independently of each other, linear correlation and regression show a close relationship between visit count and vocabulary: linear regression R-squared values 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.

4.3 Revisitation rates




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


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

One factor that is likely to have contributed to the increase in revisitation rate since Tauscher & Greenberg’s 1995 study is the prevalence of highly polished commercial sites such as **cnn.com** that offer daily modifications to the information that they offer.

4.4 Dominance of favoured pages

One factor contributing to the high revisitation rate is the fact that almost all of the subjects had one or two pages that they visited far more often than any other page. 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.

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. Rows 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 (accessible from the “Home” button on the browser’s toolbar); —the page is in the user’s bookmark collection (accessible from the bookmark menu); —the page is in the user’s “Personal Folder” in the bookmark collection (accessible from the “Personal Folder Toolbar”);

—the page is not in any of the above categories. It is interesting to note that 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). Interviews with the subjects revealed a variety of techniques used to access these pages, including links encoded in the HTML of their home pages, and hand-editing of URLs in Netscape’s “Location” text entry widget.

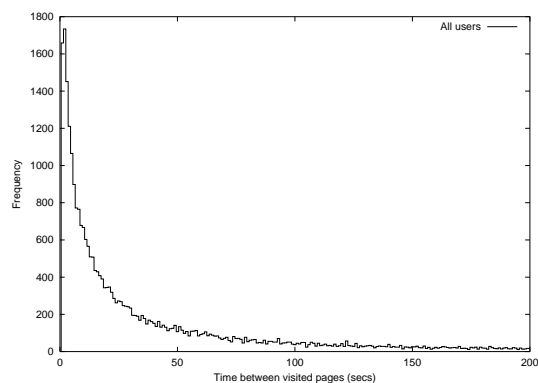


Figure 3: The number of pairs of pages visited within certain time gaps (number of pairs on the y-axis, time gaps on the x-axis).

4.5 Temporal aspects of page visits

Most pages are visited for only 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 between pages were relatively rare.

This result was calculated by taking the set of URLs in each user’s history file and sorting them by the *Last* field. The difference between the *Last* times for each pair of successive entries shows the time between page visits. This technique overestimates time gaps because the history files were collected on a daily basis, meaning that only one *Last* entry will be recorded for pages that are visited many times in the day.

This result shows that browsing is a rapidly interactive activity. It also implies that many (or most) pages are simply used as routes to other pages.

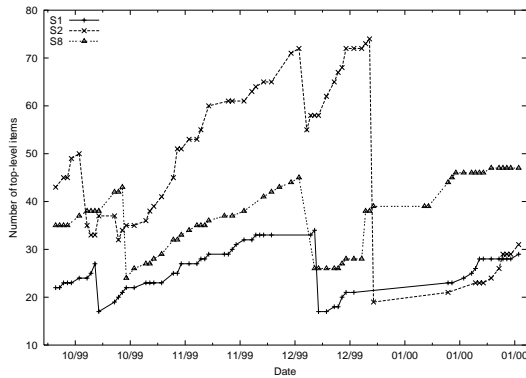


Figure 4: URLs and folders in the top-level bookmark file by time.

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 (s.d.=166.15). The mean number of folders used to store bookmarks was 18.1 (s.d.=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 (s.d.= 29.7) and 3.7 (s.d.=5.2).

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.

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 de-

pends 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.

4.7 A Community of Users?

Our subjects all worked or studied within the same department. We found a high degree of uniformity across subjects in the percentage of web-page accesses made within the department (mean 58%, s.d.=19%), within the parent institution (mean 6.5%, s.d.= 5.2%), and international (mean 30%, s.d.=18.4%). These values are summarised on rows 28 to 30 of Table 1.

The relative similarity of these values might imply that the subjects were visiting similar areas in the web. Closer analysis, however, reveals that this was not the case. For each page in the total URL vocabulary of 16290 distinct URLs visited by the subjects, we counted how many subjects had visited it. Ninety percent of the URLs (14734) had been visited by at most one of the subjects: that is, only 10% had been seen by more than one subject. Only one page (the University’s home page) had been visited by all of the subjects. A total of 828 pages had been visited by two or more subjects, and only 30 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.

5 Discussion and Implications for Design

5.1 Limitations of our study

Logs versus events. The analysis in this paper updates and extends the work of prior studies by Catledge & Pitkow (1995) and Tauscher & Greenberg (1997). However, the technique we used to gather the data—file analysis from incremental backups—is different to that of the prior studies, which used logs of user events 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.

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 of 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.

Cart leading the donkey. 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. Observational studies, such as that by Byrne, John, Wehrle & Crow (1999), are necessary to clarify the mapping between the users’ tasks and the support provided by browsers.

5.2 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.

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, Czerwinski, Larson, Robbins, Thiel & van Dantzich 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. First, 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. Second, bookmark collection systems should include tools that assist users in managing their collections, particularly in identifying invalid bookmarks. Third, systems should support shortcut mechanisms (such as the “Personal Toolbar Folder”) for efficiently navigating to a small set of frequently visited pages.

History and revisitation tools. Approximately 80% of the URLs that a user visits are revisitations. This result substantially exceeds the previously reported value of 60%.

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, but these techniques have not yet

	s1	s2	s3	s4	s4	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17
Visits, vocabulary and revisitation																	
1 Visit count	4514	10304	471	1808	11236	2135	2683	7291	2612	1999	747	863	1067	8633	23973	281	2794
2 URL vocabulary	1048	1812	116	921	2455	1015	795	2309	963	552	340	229	421	1360	4105	74	1349
3 Pages visited once	627	956	62	710	1134	735	479	1165	560	350	252	119	255	625	1785	38	867
4 Revisitation rate(%)	86.1	90.7	86.8	60.7	89.9	65.6	82.1	84.0	78.6	82.5	66.3	86.2	76.1	92.8	92.6	86.5	69.0
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	12.4	3.8	8.9	3.4	11.6	10.7	11.8	6.7	16.2	3.9	5.2	3.0	0	11.0
Linear regression of visit count with vocabulary																	
7 Slope	5.07	6.54	4.26	2.02	4.72	2.11	3.46	3.13	2.57	3.81	2.20	3.60	2.64	6.60	6.44	4.33	1.97
8 R-Squared	.96	.89	.98	.99	.99	.98	.98	.99	.98	.98	.97	.99	.99	.99	.99	.97	.99
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, X= 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	⊗	⊗	⊗	⊗	⊗	X	⊗	⊗	X	X	X	X	⊗	⊗	X	X	X
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	>	⊗	⊗	⊗	X	⊗	>	X	⊗	⊗	X	X	X	X	⊗	X	X
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	⊗	X	X	X	X	X	X	⊗	X	X	>	⊗	⊗	X	>	⊗	X
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	⊗	X	X	X	X	X	X	X	X	⊗	⊗	X	X	X	X	X	X
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	X	X	X	X	X	X	X	X	⊗	X	X	X	X	X	X	X	X
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	150
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	5	2	0	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	46.9	68.3	79.6	52.6	54.0	63.2	71.0	41.4	45.0	66.9	67.1	51.3	45.3	66.9	73.8	90.0	5.9
29 Organisation	4.7	4.2	2.9	7.2	8.6	12.2	10.2	3.8	1.7	1.4	3.3	21.6	11.0	3.8	2.5	9.3	2.5
30 International	31.0	25.9	17.5	37.3	33.7	12.5	14.8	49.5	49.9	28.7	22.1	23.4	34.2	26.9	20.4	0.0	82.1

Table 1: Summary of data values retrieved for each subject.

been evaluated.

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) and Web-View (Cockburn, Greenberg, McKenzie, Jason-Smith & Kaasten 1999). Version 5 of Microsoft Internet Explorer also includes a “temporal segment” technique for reviewing recently visited pages. With the exception of Footprints, the major limitation of the work on web revisitation schemes is the lack of empirical evaluation. Cockburn & Greenberg (1999) provides a review and discussion of issues in the design of history schemes for web revisitation.

Page 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 javascript 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.

6 Conclusions

This study updates the empirical foundation for understanding web use. The study shows both expected and unexpected results when compared to 1994 and 1995 studies. 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 60% to approximately 80%: four out of five pages visited have been seen before. The results also show 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 reported.

Our future work will continue to analyse web use, and 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 Michael JasonSmith 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'.
- 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.
- Cockburn, A. & Greenberg, S. (1999), Issues of page representation and organisation in web browser's revisitiation tools, *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. 7–14.
- 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.
- 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/>
- Mayo, E. (1933), *The Human Problems of an Industrial Civilization*, Cambridge, MA: Harvard University Press.
- Nielsen, J. (2000), *Designing Web Usability: The Practice of Simplicity*, New Riders Publishing.
- 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.
- Schmalensee, R. (1999), 'Microsoft presspass: Testimony on behalf of microsoft', WWW page: <http://www.microsoft.com/presspass/trial/schmal/I.htm>
- 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.
- 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.