Web Page Revisitation Revisited: Implications of a Long-term Click-stream Study of Browser Usage

Hartmut Obendorf[1], Harald Weinreich[2], Eelco Herder[3], Matthias Mayer[1]
[1]AGIS and [2]VIKS Centers, University of Hamburg; [3]L3S, University of Hannover

ABSTRACT
This paper presents results of an extensive long-term click-stream study of Web browser usage. Focusing on character and challenges of page revisitation, previous findings from seven to thirteen years ago are updated. The term *page revisit* had to be differentiated, since the recurrence rate—the key measure for the share of page revisits—turns out to strongly depend on interpretation. We identify different types of revisitation that allow assessing the quality of current user support and developing concepts for new tools.

Individual navigation strategies differ dramatically and are strongly influenced by personal habits and type of site visited. Based on user action logs and interviews, we distinguished short-term revisits (*backtrack* or *undo*) from medium-term (*re-utilize* or *observe*) and long-term revisits (*rediscover*). We analyze current problems and provide suggestions for improving support for different revisitation types.

Author Keywords
WWW, Web browsing, Hypertext, Navigation, History, Revisitation, Recurrence rate, Web browser interfaces

ACM Classification Keywords
H.5.4 Hypertext/Hypermedia: User issues.

INTRODUCTION
The World Wide Web has become the most successful hypertext system ever, making Web browsers one of the most frequented user interfaces. Despite this indisputable importance, their interfaces still closely resemble the first 1989 prototypes [3]: Browser UIs, based on the hypertext document metaphor [11] do not match the current Web of applications and interactive Web pages (AJAX, Web 2.0).

Historically, hypertext is based on the vision of managing a constantly growing amount of information, not only providing more natural ways to access new information, but specifically introducing a concept to *revisit* information read before by following self-created trails [5]. The Web as a read-only medium lacks this revisitation concept: users can add neither links nor comments to Web documents. Hence, other browser mechanisms are needed to revisit Web pages.

How people try to find information on the Web has been subject of several studies [21]. Search engines have become the most important means to find new information, yet hyperlinks are vital to find related or more detailed information. Such navigation behavior can be investigated in short-term studies and by analyzing search engine logs.

However, only few studies have examined the revisitation behavior of Web users, and most of these focused on short-term revisitation. Knowledge about Web page revisitation is mainly based upon only three studies that range in age from seven to thirteen years. Specifically, long-term revisitation behavior is hard to analyze, requiring detailed long-term recording of user actions in their natural environment. Thus, research has mainly focused on the usability of existing tools, e.g. the use of bookmarks [13, 23] or the use of the back button [9, 25].

Furthermore, the Web has changed significantly during the past decade. Not only the number of domains and users has grown [17, 20] also its character has changed dramatically. The once static Hypertext has evolved into a dynamic medium with Web applications, interactive information resources and communication platforms. The Web—once the preserve of 'computer enthusiasts and scientists'—has become a medium for the broad public [4], delivering e-commerce, news and entertainment [10].

Little is known about the impact of these changes on users' interaction with Web browsers, their contemporary revisitation behavior, and on usability problems. We therefore felt the need to conduct a new study investigating navigation behavior not only for short-term revisits, but also for long-term revisitation, to provide new insights, and to analyze to what extent results and premises from earlier studies still hold.

RELATED WORK
While the number of studies analyzing user behavior in the Web is large, they are often limited in scope [29]. *Server logs* are easily accessible and play an important role in site usability evaluation, but they are limited to a single site and cannot report on other user activities and detailed browser interactions. *Laboratory studies* provide controlled environments, yielding detailed data (including e.g. eye move-

ments), yet the results are strongly determined by the assigned tasks and their potential to model the daily work of the user; the risk of distortion caused by the artificial environment cannot be dispelled easily. *Observational studies* deliver rich contextual data necessary to interpret user behavior and to draw conclusions regarding working conditions, workflow procedures, and user interface requirements. However, they cover only brief periods of Web use, making it difficult to study recurring patterns, rare problems, and usually do not provide sufficient quantitative data for statistical analysis.

This leaves the method of automatically capturing user interface and navigation events in *client-side logs*. Such '*click-stream studies*' provide descriptive statistics on the behavior of individual users in the Web and allow for long-term observation of user interaction and page revisitation.

Previous studies

In 1994, Catledge and Pitkow conducted the first extensive client-side Web usage study [6], analyzing the interaction of 107 students with their Web browser. They observed a frequent use of the back button, second only to hyperlinks. Personally maintained 'home pages' were used as 'indexes to interesting places' and 'hub-and-spoke' navigation was identified as a common navigation pattern: users rarely traversed more than two hierarchical levels before returning to a hub page to explore other links.

The subsequent study of Tauscher and Greenberg in 1995 focused on page revisitation behavior [32, 33]. They introduced the 'recurrence rate' as the probability that any page visit is a revisit to a previously seen page. They found a mean recurrence rate of 58% and concluded the Web was a 'recurrent system'. They differentiated browsing activities according to the URI growth rate and found that the majority of revisits were targeted on a small set of Web pages and sites. Furthermore, revisits showed considerable recency, mostly triggered using the back button. The recurrence rate was considered a key measure for the requirements of better revisitation support, and has been motivation for the development of a multitude of history tools [7].

In 1999, McKenzie and Cockburn analyzed the log files of Netscape users at their department. They reported a remarkable increase in daily visits [24], even though their study period included holidays. They also reported a rise of the recurrence rate to 81% and stated 'four out of five pages' have been seen before [24]. Finally, they found large and rather unmanaged, growing bookmark collections.

Although each of these studies is a result of excellent work, there are a number of reasons to believe their findings may not represent current Web use. Considering the age of the studies, ranging from seven to thirteen years, it is surprising that no updates are available: the Web has changed dramatically, so effects on the interaction with the Web browser are very likely. Moreover, the datasets have limitations; it is for instance probable that the duration of the first two studies (table 1) was too short to capture enough data on infrequent revisits [6, 33]. Also, the Web browser used (XMosaic) was outdated even in 1995 and, as Tauscher and Greenberg report, participants had to change their client for the duration of the study. Although McKenzie and Cockburn's participants kept their favorite browser and data was obtained retroactively from backup tapes, these history logs provided no details on the users' interaction with the browser, as only visits to URIs were recorded, and for revisits on the same day, the time was logged only for the last visit [8]. Thus, duration and sequence of frequent revisits were not available. Taken together, the reasons for an update study able to overcome these shortcomings were strong.

THE WEB BROWSING STUDY

While observing users within a laboratory setup is well understood and frequently done in usability studies [30], capturing data about the activities of users in their daily work environment holds many challenges. From our experience, people have become very sensitive about *privacy*; it was challenging to find participants with the high degree of trust required to allow recording all browsing activities. Also, changes in the Web and the work environment make it increasingly difficult to get a consistent and coherent sample of Web use: while Catledge and Pitkow were able to install the same browser for everyone, and control their participants' only means to access the Web, today computing machinery is both diverse and increasingly mobile: different browsers would be used, and not all browsing activity could be observed. To run within real work environments, the logging tool had to run reliably on a number of platforms.

25 Web users contributed logging data to the study presented here. Technically, they were all equipped with an intermediary based on the Java Scone Framework [27] and WBI [2], which logged all page requests, the triggering user actions, and central page characteristics. 15 participants agreed on using an instrumented Firefox 1.0 browser [26], while the remaining 10 users preferred to use their familiar browser. The instrumented Firefox produced a detailed log on the use of the 76 most important user interface widgets. It allowed us to improve the interpretation of all users' logs, e.g. to identify page requests that were not related to user actions, and to analyze UI events that did not lead to page requests for these 15 users.[1]

Earlier studies (e.g. [18]) do not mention any preprocessing of the recorded data. However, we found this step to be vital in order to obtain logs that actually represent single, user-initiated page visits. We found a large number of 'artifacts' in the untreated intermediary logs, events indicating the loading of inline frames, sub-frames that were loaded sequentially into a frameset, advertisements, pop-up windows and automatically refreshed Web pages [35]. Advertisements—mainly iFrames—made up to 33% of all page

[1] 13 users also used an unlogged browser for Web access at home; impossible to avoid in a naturalistic long-term study.

Table 1: Major Web usage studies and their main measures.

	Catledge & Pitkow [6]	Tauscher & Greenberg [32,33]	McKenzie & Cockburn [8,24]	This Study [36,37]
Date of Study	1994	1995-1996	1999-2000	2004-2005
Method	Instrumented XMosaic	Instrumented XMosaic	Daily Netscape History & Bookmark File Backups	Web Proxy & Instrumented Firefox
Data Captured	Visits & User Actions (34 Types)	Visits & User Actions (32 Types)	Visits & Bookmark History	Page Stats, Visits & User Actions (76 Types)
Length (days)	21	35-42	119	52-195 (ø=105, ñ=109)
# Users	107	23	17	25
Link Events	45.7%	43.4%	–	43.5%
Back Events	35.7%	31.7%	–	14.3%
Form Submit	–	4.4%	–	15.3%
∑ Direct Access	12.6%	13.2%	–	9.4%
Recurrence Rate	61% (reported in [35])	58%	81%	45.6% (43.7%, see text)
Type of Users	100% CS	100% CS	100% CS students	64% CS, 36% other academics
Visits	31,134	19,000	84,841	137,272
URIs			17,242	65,643
Visits / User		>300	281 – 23,973	912 – 30,756
Ø Visits / Day	14	21	41	89.8 (per active day)

requests for users without an ad blocker and had a significant effect on measurements.

Study Setup

All 25 participants were unpaid volunteers. Apart from using our logging software, they took part in two 90-minutes interviews at the beginning and end of the study. Six participants (24%) were female. Ages ranged from 24 to 52 years (mean: 30.5). All participants were experienced Web users (3 to 12 years, ø=8). The study took place mainly in Germany, and in the Netherlands (two Germans worked in Ireland, one in New Zealand). All interviews were conducted in the participants' native language. While 16 participants (64%) were affiliated with computer science, 9 participants (36%) had different backgrounds: two worked in psychology, and one each in sociology, geology, electrical engineering, trading, coaching, history, and photography. Seven additional candidates were unable to complete the study due to personal or technical reasons and were not considered in the evaluation.

User actions were logged during a period of 52 to 195 days, resulting in 137,272 events corresponding to individual, user-initiated page requests. On average 89.8 pages were visited per active day (days with at least one event logged). The individual average usage varied widely from 24.9 to 283.6 page visits per active day. Although this indicates a rise in average Web use (compare Table 1: even if weekends and holidays are considered in the calculations for previous studies, the number of visited pages seems to be steadily rising), we think that such conclusions should be drawn with care; rather, we think the numbers emphasize large personal differences in kind and intensity of Web use.

To investigate revisitation behavior and personal habits of our participants in more detail and to capture more aspects of their use contexts, we conducted two interviews at the beginning and the end of the study. The first interview focused mainly on demographical data, general problems and browsing habits. The second interview aimed at the interpretation of actual situations during the study to reveal personal revisitation strategies and preferences of browser tools. We asked the users to recall and comment on long-term revisitation actions during the study, using graphical presentations of several navigation sequences that we assumed to be related to revisiting important information.

Limitations of the Study

A click-stream study inherently holds certain difficulties of interpretation [19]. For example, a log of user interactions with the Web does not exhibit all aspects of the user context and the underlying motivations for user behavior. In order to overcome this issue, we carried out two interviews (as described above) at beginning and end of the study. This qualitative data turned out to be crucial for interpreting several quantitative results, but still could only deliver limited data for a detailed qualitative analysis.

Although we tried to recruit participants with different backgrounds, all were frequent computer users with long Web experience. Still, the variance in the captured data was fairly large for almost all aspects of interaction with the Web (Table 3). Whereas this large variance prevents drawing conclusions on the 'average use of the Web', it also shows that Web browsers are used with various personal preferences and that individual users have particular de-

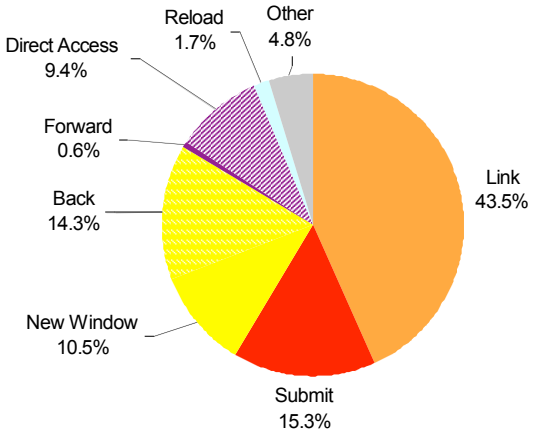

Figure 1: User actions leading to page visits.

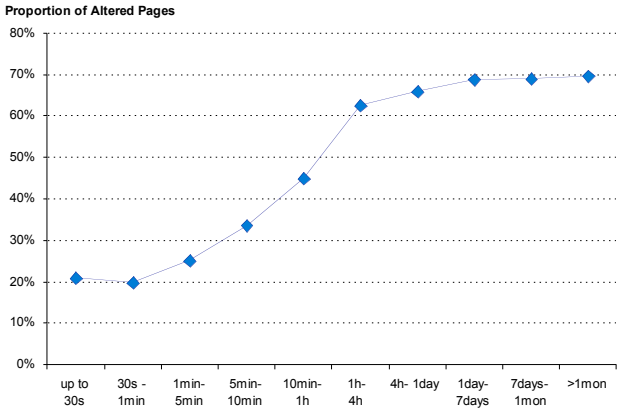

Fig. 3: Page change rates for different revisitation periods.

mands. Examining these different demands and behaviors more closely was one important goal of this study.

RESULTS OF THE WEB BROWSING STUDY

The first unexpected result we found was a considerably lower use of the back button compared to earlier studies. In Catledge and Pitkow's data [6], its use amounts to 36% of all navigation actions. Tauscher and Greenberg still found a 32% share of 'back' events [33]. Looking at the navigation actions of all users in this study (Figure 1), the back button rate decreased to 14.3% of all navigation actions. Although the large discrepancy to preceding studies may seem surprising, recent smaller studies [25,14] did also report a lower share of back button use. Submission of forms has become much more important (15%), as has opening pages in a new window or tab (11%). Following hyperlinks remained to be the most frequent activity with 44% of all user actions. Choosing a bookmark, typing a URI in the address bar and the homepage button (subsumed as 'Direct Access', 9%) were used somewhat less than in earlier studies (Table 1).

Redefining 'Revisit' and 'Recurrence Rate'

The reduced share of back button usage suggests our users returned less frequently to previously visited pages. We therefore calculated the *recurrence rate*—the probability that any page visit is a revisit, introduced by Tauscher and Greenberg. This rate seemed to grow in time—from 60% in the mid-nineties to about 81% in the end of the nineties (Table 1). Our results did not follow this upward trend: the average recurrence rate of our users was only 45.6%.

Table 2: Mean recurrence rates averaged over all users.

Truncated URI	Full URI & GET Param.	Full URI & POST Param.	Page Content
69.4%	**45.6%**	**43.7%**	**34.6%**

We found several reasons for the changed rates. Firstly, we preprocessed the log data to remove page requests that were not directly related to user actions (see page 2-3 and [35]). This cleaning process influenced the recurrence rates, as advertisements, frames and auto-reloads led to many additional revisits. Without data cleaning, individual recurrence rates were 6% to 20% higher, and the mean rate was 54.1%.

Furthermore, the notion of *revisits* varied between the different studies, as the original definition by Tauscher and Greenberg (Figure 2) allows for different interpretations.

$$R = \frac{total\ URLs\ visited - different\ URLs\ visited}{total\ URLs\ visited} \times 100\,\%$$

Figure 2: Recurrence rate by [32]

Investigating the particularly high rate of revisits reported by McKenzie and Cockburn (81%), we found that they truncated the URI and did not consider any query parameters for search pages [24]. Although they report this did not change their subsequent analyses, it increases the number of revisits: Every query to a search engine and all result pages would be regarded as the 'same'. If this were used for all sites, every visit to a dynamically generated Web page based on 'HTTP GET' parameters would be considered a revisit, even if the parameters determined different page content. page. For our participants, this would result in a mean recurrence rate of nearly 70% (Table 2).

Since dynamic Web sites and form submission actions had a much higher relevance in our study than in previous ones, we found it necessary to reconsider the definition of 'revisitation' and find more exact definitions of types of revisits.

When users revisit Web pages, they might want to access the same resource again—just as if they wanted to return to a known place in the real world. However, like the real world, the Web is in constant change. It may well be that users want to re-access a resource as they *expect* changed content, for example, new headlines on a news site.

In order to analyze to what extent the content of Web pages had changed upon revisitation, we recorded fingerprints for every page visit (hash codes calculated from the page source code). For revisits within one hour, the content of 26% of all document had changed, a rate much lower than

the average number of page requests involving parameters[2]. However, after one day or later already 69% of all revisited pages did experience a change, a rate that stays nearly constant for longer periods in our study (Figure 3).

The above numbers demonstrate the highly dynamic nature of the contemporary Web. We think that, ideally, a definition of recurrent behavior should distinguish revisits motivated by reading the *same* content from revisits motivated by reading *updated* content. If the fingerprint of the page content is considered for the calculation of the recurrence rate, the average rate of our participants would even be below 35%. However, it is difficult to automatically determine whether changes to the content of a Web page are *relevant* to the user or not (e.g. changes to embedded advertisements are usually insignificant).

Accordingly, we argue for a notion of page revisits that comprises both same-content and updated-content revisits. This definition should be able to distinguish resources that are not only determined by the full URI (including the query part, i.e. all HTTP GET parameters) but also by POST parameters, as many dynamic Web resources require these values to identify the content. In order to consider all parameters in this definition, an address in this sense is the concatenation of the full URI string[3] with the string of POST parameters (Figure 4).

$$R = \frac{total\ visits - |Full\ URI + POST|}{total\ visits} \times 100\,\%$$

Figure 4: An updated definition of the recurrence rate.

Following this definition, the average recurrence rate was only 43.7% (Table 2) compared to 45.6% without POST parameters and 69.4% with neither GET nor POST being considered. This variance demonstrates that the definition of a revisitation is vital for all following statistics, and also points towards the highly dynamic nature of the Web [36].

The Influence of User Habits and Site Types on Revisits

Analyzing our data we found that two aspects of revisitation behavior deserve more attention: the influence of personal user habits and the character of visited sites.

We measured a high intra-individual variation of the recurrence rates: Calculated based on our definition, rates ranged from 17.4% to 61.4% (see Table 3). This suggests that drawing extensive conclusions for user requirements based on *mean* recurrence rates is potentially misleading—personal behavior seems to differ dramatically. Although not all data from earlier studies is available for analysis, both Tauscher (42.5-74.3% [33]) and McKenzie and Cockburn (60.7-92.6% [24]) reported a rather large variance.

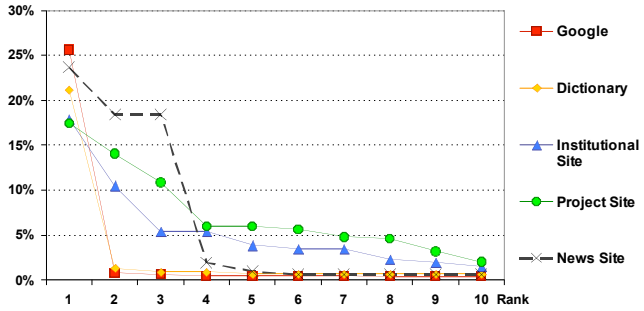

Figure 5: Revisitation distribution for different sites.

These observations suggest that revisitation rates reported in this and previous studies only *illustrate possible ranges*. Individual behavior is often more important for an analysis of user requirements than looking at averages. It might be helpful to discern different types of users, or even tasks.

To explain the observed differences in user behavior, we tried to identify different user groups in our population. We found no supportable effects of profession, gender or nationality. Individual differences were mainly caused by user tasks—that also differed significantly between members of the same department or firm—personal habits, private interests and, accordingly, the sites visited.

The influence of site type on page visitation was quite high. When a site was visited more frequently, also more different pages within this site were visited ($r=0.903$, $p<0.01$). Revisits to some site types entail many different revisited pages on the same site, while other site types are characterized by only one revisited page (Figure 5).

Search engines and dictionaries provide a single portal page as access point; from this page a query is issued, which leads to various result pages. Hence, by their very nature, these sites have only one 'popular' page and a long tail of pages that are visited only once or twice. By contrast, institutional and project Web sites often have a portal page which is accessed quite often, but also a range of other pages that are revisited regularly; these pages may offer information on a certain topic or department, or may provide applications which are used on a regular basis. Finally, several news sites provide a few frequently visited pages; they relate to overview pages of certain news categories the user was interested in. Future revisitation tools could consider these site characteristics. For example, when bookmarking a news site with three popular categories, the single bookmark could automatically generate three sub options (say: local news, sports and entertainment) based on user habits. For institutional Web sites the bookmark could provide an appropriate hierarchy, either determined by the site's structure or derived from frequent visits.

[2] 44.1% of all page requests were parameterized; as many links encoded parameters in the URI, this number is higher than the number of form submissions alone (15.3%).

[3] The 'fragment' reference was ignored as it typically only addresses a location within the resource, i.e. the browser scrolls to a certain position. In our study, almost no links (<1%) included this element of the URI.

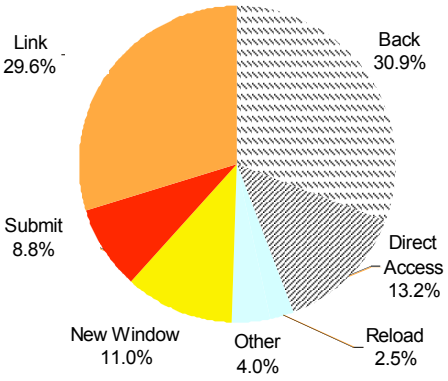

Figure 6: User actions leading to revisits.

A further finding of this study may well be important for future history tools: the personal dominance of a few sites seems to grow for many users. The share of visits to the personal top ten sites ranges from 37.9% to 89.8% in this study (Table 3). The overall most dominant site was Google search with 15% of all page requests. It was the most frequently visited site for 11 participants and within the top 4 sites for all other users.

ANALYZING USER REVISITATION BEHAVIOR

Users do not only show different recurrence rates and site-specific behavior, they also have many different intentions for revisiting a page. As a first measure towards connecting activities and classes of revisits, we distinguish revisits by the type of actions that were used to access a page.

For our participants, the back button caused only 31% of all revisits (see Figure 6). Another type of navigation actions leading to revisits was 'direct access' (bookmarks, the homepage button, the history list and typed URIs); they were only responsible for 13.2% of all revisits. Over 50% of page revisits were triggered by other navigation actions, mainly link following. Considering the high share of short-term revisits—we found 72.6% of all revisits to occur within one hour (per user between 50.0% and 83.9%)—we had expected a higher rate of back button use for revisits.

The low back button share was not caused by an increased use of the back button pull-down menu (Figure 7) as we first assumed: only 3% of all back button events originated from this pull-down menu. Participants' explanations comprised that it is 'often simpler to just click several times on the back button', than to make the pull-down menu appear and scan its often incomprehensive list of page titles.

We found evidence that the low back button usage was caused by major changes in browsing strategies: a considerable share of 'hub-and-spoke' navigations [9] has been replaced by opening link targets in new windows or 'browser tabs'. We found that some users opened many windows or tabs to navigate to different pages from a hub page (e.g. search results and news overview pages). As the old page remains accessible, the effective need for backtracking is greatly reduced. Instead of navigating back and forth, peo-

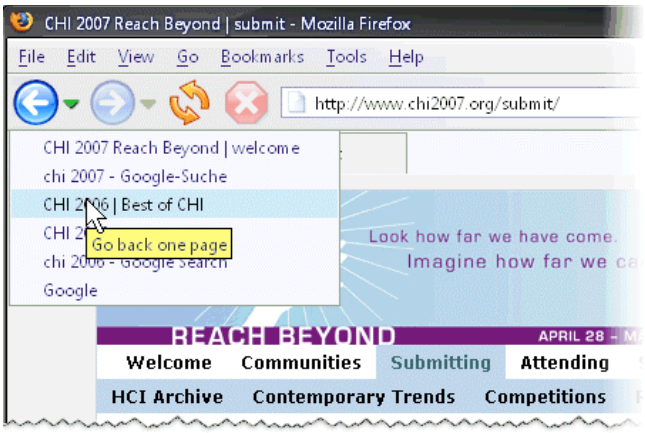

Fig. 7: Pull-down menu of back button: only of little use?

ple switch between different windows or tabs. This results in fewer page requests and fewer revisits.

Our data supports this hypothesis: the group of participants with the top third of new window events employed the back button to a lesser extent (10.2%) than the bottom third (16.4%), indicating that multiple windows are used as an alternative to backtracking ($t=2.509$, $p=0.026$).

Further, six of our 15 Firefox users reported to make frequent use of browser tabs. For them the group of tab actions (open, select, and close tab) represented in mean 19.2% of all UI activities. Consequently, they were backtracking less often (9.9%) than the remaining users (18.3%) that hardly opened any tabs ($t=2.311$, $p=0.038$). In the interviews, these tab users reported to utilize tabs as a means to compare pages or to keep important information at hand.

A second reason for the lower back button usage is related to the increased number of form submissions. We compared the backtracking usage of the top third 'form submitters' of our participants with the remaining participants. The regular users of Web forms pressed the back button less frequently (9.2%) than the remaining participants (16.2%), a difference that is marginally significant ($t=2.715$, $p=0.012$). This result characterizes a major change of the Web: the move from a hypertext information system with primarily static documents into a combination of common information source and service-oriented interactive sites. The latter are more comparable to applications than to hypertext systems. Whereas hypertext navigation involves orienteering behavior with frequent backtracking, interactive applications are mainly used for completing certain tasks that consist of different workflow steps.

Our participants reported several problems with the back button caused by these changes of the Web. First, backtracking fails when *multiple windows* or *tabs* are used. For every new window or tab, a new history stack is created, barring return to the originating page via the back button. Instead, users have to handle different windows and tabs to relocate the originating document. This was considered especially problematic when multiple tabs and windows

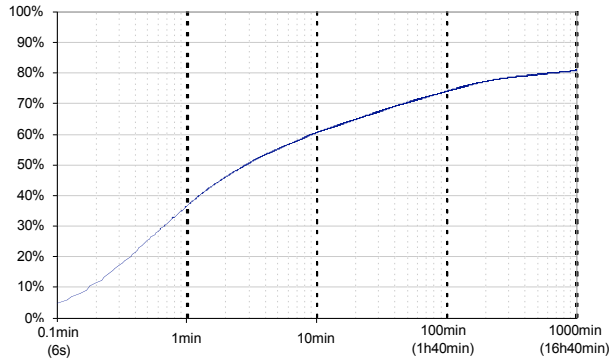

Figure 8: Cumulative page revisits in logarithmic time.

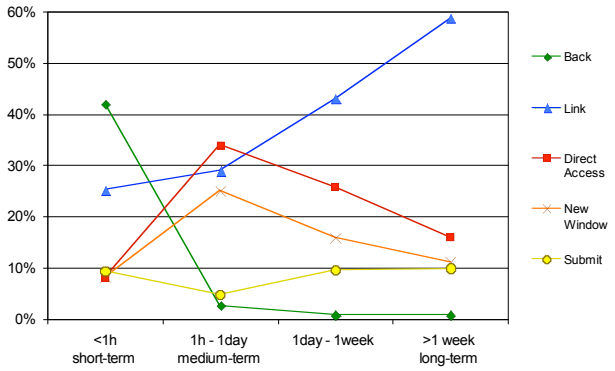

Figure 9: Comparison of methods for initiating revisits.

were used at the same time. The increased cognitive overhead related to handling multiple windows in hypertext systems were already reported in pre-Web studies [16].

Further, the back button is often unsupported by Web applications. They show unexpected effects if the user returns to the last page, e.g. when the input data from the last form is deleted and has to be retyped. Backtracking to pages created from POST parameters actually leads into a warning message of the browser and often even causes an error message of the Web application. Furthermore, such pages cannot be bookmarked at all; they are *volatile* and no browser history mechanism allows for returning to them.

Better *browser support* for *multiple windows* and *Web applications* should prevent these problems and benefit Web users and developers. Web system designers should consider that the back button is still an often-used interaction tool that users heavily depend on. It is thus dangerous to simply disable it, and deprive users of this tested tool—apart from technical difficulties that arise when users use e.g. gestures or keyboard shortcuts to trigger the back function. Instead of hiding the back button, as commonly practiced in many Web applications, it should rather support the intentions of users: when users click back in an application context, this usually means 'undo'.

Temporal and Action-Based Classification of Revisits

Previous studies did already reveal that users interact quickly with their Web browser [8] and most revisits occur after a short time [32]. Although during our study over 50% of all revisits occurred within 3 minutes, the other half took place after longer and much longer periods (compare the flat curve of Figure 8). Still a mean of 15% of all revisits occurred after a week or longer.

In order to distinguish different kinds of revisits and to group user navigation in meaningful and manageable chunks, the notion of *sessions* has been introduced. It is frequently used in server-log analysis. Owing to the low descriptiveness of common server logs, heuristics are required to define a contiguous sequence of actions of one user [12]. As a rule of thumb, many log analysis applications use a timeout of 30 minutes [31].

A similar separation of user activities is often used on client side: here, a session denotes not the visit to a single site, but rather a continuous period of browsing. Statistical analysis of user actions by Catledge and Pitkow led to the definition of a session timeout of 25.5 minutes [6]: the mean time between page requests in their study was 9.3 minutes. Adding 1.5 standard deviations, they identified a timeout of 25.5 minutes, a definition that was also used in later studies. However, this definition is problematic, as the time between two page visits in our study does not follow a Gaussian but a long tailed Zipf distribution (compare [8])—52% of navigation events followed within 12 seconds, while some lay hours or days apart [37]. Any time-out value would have been an arbitrary point on this long tail.

Consequently, we chose to follow an alternative approach based on the main time units effecting our lives: we differentiate between revisits that take place within an hour ('short-term'; 72.6% of all revisits), a day ('medium-term'; 12%), a week (7.8%), or longer ('long-term'; 7.6%). Using this naturalistic classification, we were able to identify different user strategies to revisit Web pages.

Expectedly, the back button was the preferred means of returning to pages after an hour or less (short-term), closely followed by links, which probably relate to the many structural links modern sites provide to return to landmark pages. While problems concerning short-term revisits were already discussed in the previous section, revisitation behavior for medium-term and long-term revisits showed different patterns and problems.

Browser Support for Medium-Term Revisitation

Looking at revisits between one hour and a day, another pattern emerges: 'direct access' events (URL-entry, bookmark selection) were most frequent for such page revisits. We found that these events mainly related to pages visited on a regular basis, the most prominent members of this category of resources are query pages (e.g. search engines and dictionaries), overview pages of frequently updated sites (like news services) and personal pages of different online services (shopping sites, online auctions or blackboards).

Individual revisitation strategies for such regularly accessed pages differed a lot. Some users only used the bookmark

Table 3: Overview of descriptive statistics for the participants of this study.

User	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Visit Count	7373	5069	5694	5961	30756	1784	912	9757	19570	1506	2241	1315	997	2786
URL Vocabulary	3675	2678	2534	2248	16869	1404	404	6141	9525	914	1267	703	643	1477
Site Vocabulary	610	436	269	523	2127	318	132	1258	1647	137	180	140	136	307
Pages Visited Once	2639	1863	1729	1360	12640	1190	227	4744	6604	646	956	462	512	985
Recurrence Rate	49.8%	44.0%	55.1%	54.2%	44.0%	17.4%	50.4%	33.6%	47.3%	37.4%	35.3%	43.9%	34.8%	41.0%
Google Search Use	15.5%	29.9%	4.0%	12.7%	17.4%	18.9%	22.7%	17.0%	5.9%	8.7%	13.3%	29.9%	16.1%	6.5%
Visits In Top 10 Sites	60.6%	53.4%	75.9%	42.9%	61.9%	47.2%	62.7%	43.8%	37.9%	60.5%	69.7%	72.1%	65.0%	50.3%
User	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**Mean**	**StdDev**	**Median**
Visit Count	2025	1036	3603	7752	2442	3784	5228	3884	1809	4865	5123	5490.9	6564.8	3784.0
URL Vocabulary	1365	640	1270	3018	1381	2080	2223	2002	968	2142	3441	2840.5	3515.7	2002.0
Site Vocabulary	153	143	84	495	263	465	460	450	218	322	655	477.1	496.4	318.0
Pages Visited Once	1076	539	986	2068	1042	1458	1302	1390	664	1449	2933	2058.6	2615.2	1302.0
Recurrence Rate	28.6%	37.2%	55.5%	61.4%	42.5%	39.1%	58.8%	49.9%	45.3%	53.2%	32.9%	43.7%	10.3%	44.0%
Google Search Use	8.4%	16.2%	13.4%	13.8%	13.4%	24.8%	26.3%	38.5%	20.6%	4.6%	17.1%	16.6%	8.6%	16.1%
Visits In Top 10 Sites	66.9%	59.7%	89.8%	69.5%	62.3%	68.3%	64.6%	66.5%	58.7%	68.4%	54.5%	61.3%	11.4%	62.3%

menu, others only the bookmark toolbar, and a few participants had the habit to type in the URI into the address bar using its auto-completion feature. Some participants also used icons on their desktop to open frequently used pages in a separate window.

Relating this to Tauscher and Greenberg's argument of Web navigation as a *recurrent system* [6], a major share of revisits does not concern 'content pages', but resources that either provide access to different Web applications, or supply a list of links to content pages.

The emerging importance of such dynamic resources becomes manifest in two recent developments. Firstly, the frequent ***re-utilization*** of such query-based pages is partly replaced by small appliances, such as Apple's 'Dashboard', 'Yahoo! Widgets', and browser extensions that provide an integrated toolbar for formulating and submitting queries without the need to load an HTML page (such as the Google toolbar). However, a flexible and direct integration of Web appliances in common office applications, like online dictionaries in word processors, is still not commonly supported.

Secondly, if users frequently return to known places on the Web to check for updates, i.e. on a news site or a forum, they ***observe*** it for interesting changes. Lists of frequently updated pages are increasingly provided as RSS feeds and can be integrated into the browser sidebar using dynamic bookmarks, and special RSS feed aggregators are becoming more widely used. If this trend continues, revisitation rates are likely to drop in the future, as browser use decreases for observation of such resources.

This demonstrates that some Web applications might benefit from a more adaptable browser user interface—without the urging need for dedicated applications. Browsers should therefore support a simpler and better way to tailor the interface to the habits of the user and the type of Web application used.

Browser Support for Long-Term Revisits

Long-term revisits are usually motivated by the intention to ***rediscover*** content accessed earlier, meaning users are concerned with finding information or a tool they already had accessed before. Due to the extensive nature of the World Wide Web, this *rediscovery* is often a severe problem [22]. We found our participants to apply different strategies for such activities and to face several specific problems.

Interestingly, hyperlinks (>58%) initiated by far the most long-term revisitations (Figure 9). History and bookmarks—provided to support medium *and* long-term revisitation—were only used rarely (16%).

Possible explanations for the low use of bookmarks are, first, that people may have used a bookmark for visiting a first starting page but then created several long-term revisits by following well-known hyperlinks on this and subsequent pages. Further, pages have to be actively bookmarked in advance before being able to rediscover them using this means. The vast majority of our users stated to prefer small, manageable bookmark archives over large, complex ones; the problems with organizing bookmarks are well known [1] and alternative approaches to hierarchical organization, such as the promising *del.icio.us*, have been found to be difficult to manage over extended periods of time [15].

In addition, URL-entry with auto-completion seems to be of limited help. Usually, only commonly used URIs can be memorized by users and directly typed into the address bar. Unfortunately, the auto-completion feature is available solely for addresses that were entered recently or are stored in the browser history. As the latter is also limited in time, pages accessed a few weeks ago are not auto-completed.

Particularly, the browser history remained almost unused and merely 0.2% of all page requests were initiated from it. Only two of our twenty-five participants stated to use it from time to time, but they also reported to only use it, if they *knew* they would find a page there and other alternatives failed. Ten participants were not aware of the browser history at all.

While all above listed approaches can be subsumed as direct '*re-access*' to Web resources, we found our users to apply two additional strategies for long-term revisitation: they '*re-search*' and '*re-trace*' the Web for information.

Re-searching was reported by several participants as a common strategy to rediscover documents. It involves reproducing search engine queries, or using search engines to look for remembered content. Even if it was considered a promising strategy, two drawbacks were reported repeatedly. First, users often had problems to remember the original query. The drop-down box under the search field did not really help to reproduce a query, as they could not search it in temporal order and did not see what queries yielded successful results. A time and task based search history could help to redo searches. The second problem was caused by the rapid change of search result pages of global search engines[4]: even if they remembered the right search term, the result list presented different hits. Therefore, a search history should as well be able to give access to previous search results.

Re-tracing, finally, denotes the following of known paths, e.g. from a search result page or a company's home page. This was with 58% the most frequent strategy of our participants for long-term revisits (Figure 9). Much as in the original conception of hypertext by Vannevar Bush [5], users seem to follow trails to relocate information after a longer period. Unfortunately, the Web does *not* support Bush's concept of trails: a user gets no support by the browser to reproduce previous navigation paths, and even the only clues the browser provides—purple colored link anchors for references to recently visited pages—vanish after a few days. If trails would be preserved in the browser history and be made visible, this could support the retracing of previous paths [34].

A final improvement necessary for long-term revisits seems to be proper support for intended same-content revisits. Due to the dynamic nature of the Web, a local storage of interesting items should be considered. With increasing amounts of permanent local disk storage, there is no reason not to record a searchable history of Web pages allowing for full text search. It would also help to retrieve earlier versions of updated pages or content of volatile pages, for instance booking confirmations or invoices created from POST form data that usually cannot directly be revisited. Not all pages would have to be stored and some pages cannot be stored in a useful way, e.g. pages providing input forms are often useless outside the application context. Pages created using AJAX techniques can often not even be stored locally or printed. Further research is necessary to provide ways to deal with the resulting usability problems.

Although long-term revisits had in average only a share of 7.6% of all page revisits (mean per user: 1.2%–11.4%) the majority of our participants stated that some of these rare revisits were very important to them and that they encountered severe problems as mentioned above. This emphasizes the importance of developing and integrating new improved long-term revisitation tools into common Web browsers.

[4] For our participants, about 97% of all result pages had changed in content after a single week.

CONCLUSION

We presented results of an extensive long-term click-stream study that captured the Web usage behavior of 25 participants with diverse backgrounds and tasks. Seen in contrast to earlier studies, our results indicate that many aspects of interaction with the Web have changed. This has different effects on revisitation requirements. For short-term revisits multiple windows and tabs allow for new navigation strategies, but create new problems with locating a document, as *backtracking* by the back button is often not possible. The strong increase of the proportion of submit events stands for a growing number of dynamic Web pages and 'Web applications'. However, these often do not support the back button either and call for an *undo* function in browsers.

We identified opportunities for the development of new browser tools that target not the bulk of revisits, but specialize on certain user requirements for revisitation. Support for *observational* behavior is already given by RSS feeds, but little is known on their usability and presentation. *Re-utilization* is partly provided by special browser extensions and appliances like 'widgets', yet the integration of Web services and office applications is still rarely possible.

Finally, a neglected field of research seems to be browser support for *rediscovering* resources that have been accessed a longer time ago. These revisits were quite rare, but often important. Neither browser history nor bookmarks seem to be reliable tools for long-term rediscovery. Instead, users re-searched and re-traced the Web for the desired information. As missing original pages often caused problems, a searchable copy of search terms, visited pages and user trails could severely enhance long-term revisitation support.

ACKNOWLEDGEMENTS

We thank Horst Oberquelle, Saul Greenberg, Winfried Lamersdorf, all reviewers, our colleagues at Hamburg and Twente, and our participants for their spiritual support, practical help, and for their comments. We are indebted to *C1WPS*, *conftool.net* and *mmsc.de* for financial support.

REFERENCES

1. Abrams, D., Baecker, R., Chignell, M. Information Archiving With Bookmarks: Personal Web Space Construction and Organization. In *Proc. CHI 98*, ACM Press (1998), 41-48.
2. Barrett, R., Maglio, P.P., Kellem, D.C. How to Personalize the Web. In *Proc. CHI '97*, ACM Press (1997), 75-82.
3. Berners-Lee, T.: Information management: a proposal. CERN, March 1989.
4. Brown, B., Sellen, A. Exploring Users' Experiences of the Web, *First Monday* 6(9), 2001.
5. Bush, V. As We May Think. *The Atlantic Monthly*. July 1945; Reprinted in *Interactions, (III)2*, 1996, 35-46.

6. Catledge, L.D., Pitkow, J.E. Characterizing browsing strategies in the World-Wide Web. In *Proc. WWW 1995*, ACM Press (1995), 1065-1073.
7. Cockburn, A., Greenberg, S., Jones, S., McKenzie, B., Moyle, M. Improving Web Page Revisitation: Analysis, Design, and Evaluation. *Information Technology and Society*, 3(1), 2003, 159-183.
8. Cockburn, A., McKenzie, B. What Do Web Users Do? An Empirical Analysis of Web Use. *International Journal of Human-Computer Studies*, 54(6), 2001, 903-922.
9. Cockburn, A., McKenzie, B., Jason-Smith, M. Pushing Back: Evaluating a New Behaviour for the Back and Forward Buttons in Web Browsers. *International Journal of Human-Computer Studies*. 57(5), 2002, 397-414.
10. comScore Media Metrix. Surfing Down Memory Lane to January 1996: comScore Media Metrix Revisits First-Ever Web Site Rankings. February 25, 2004 http://www.comscore.com/press/release.asp?press=434
11. Conklin, J. Hypertext: A survey and introduction, *IEEE Computer*, 20(9), 1987, 17-41.
12. Cooley, R., Mobasher, B. & Srivastava, J. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems* 1, 1999.
13. Abrams, D., Baecker, R., Chignell, M. Information archiving with bookmarks: personal webspace construction and organization. In *Proc. CHI'98*, ACM Press, (1998), 41-48.
14. Danielson, D. Web navigation and the behavioral effects of constantly visible site maps. In *Interacting with Computers,* 14(5), 2002, 601-618.
15. Guy, M., Tonkin, E. Folksonomies: Tidying up Tags? In *D-Lib Magazine*, 12(1), 2006.
16. Halasz, F.G. Reflections on Notecards: Seven Issues for the Next Generation of Hypertext Systems. In *Communications of the ACM*, 31(7), 1988, 836-852.
17. Harris Interactive. More Than Four in Ten Internet Users Now Have Broadband – Doubled in Two Years. Poll #63, 8.9.2004, http://www.harrisinteractive.com/harris_poll/index.asp?PID=492
18. Heer, J., Chi, E.H. Separating the Swarm: Categorization Methods for User Access Sessions on the Web. In *Proc. CHI2002*, ACM Press (2002), 243-250.
19. Hilbert, D.M., Redmiles, D.F. Extracting usability information from user interface events. *ACM Computing Surveys* 32, 4 (2000), 384-421.
20. Internet Systems Consortium. ISC Internet Domain Survey 2005, http://www.isc.org/index.pl?/ops/ds/
21. Jansen, B.J., Pooch, U.W. A Review of Web Searching Studies and a Framework for Future Research." In *Journal of the American Society of Information Science* 52(3), 2000, 235-246.
22. Jones, W., Bruce, H., Dumais, S. Keeping found things found on the Web. In *Proc. CIKM'01*, ACM Press (2001), 119-134.
23. Jones, W., S. Dumais, et al. Once Found, What Then?: A Study of "Keeping" Behaviors in Personal Use of Web Information. In *ASIST 2002, 39*, 1 (2002), 391-402.
24. McKenzie, B., Cockburn, A. An Empirical Analysis of Web Page Revisitation. In *Proc. HICSS'01*, 2001, 501-509.
25. Milic-Frailing, N., Jones, R., Rodder, K., Smyth, G., Blackwell, A., Sommerer, R. Smartback: supporting users in back navigation. *Proc. WWW'04*, ACM Press (2004), 63-71.
26. Mozilla Project. Mozilla Firefox browser, 2006. http://www.mozilla.org/products/firefox/
27. Obendorf, H., Weinreich, H., Hass, T. Automatic Support for Web User Studies with SCONE and TEA. In *Ext. Abstracts CHI'04*, ACM Press (2004), 1135-1138.
28. Pitkow, J. 4th GVU WWW User Survey, 1995. http://www.cc.gatech.edu/gvu/user_surveys/survey-10-1995
29. Pitkow, J.E. Summary of WWW Characterizations. In *World Wide Web*, 2, 1-2 (1999), 3-13.
30. Reeder, R., Pirolli, P., Card, S.K. WebLogger: A data collections tool for Web-use studies. UIR Technical report UIR-R-2000-06, Xerox PARC, 2000.
31. Spiliopoulou, M., Mobasher, B., Berendt, B. and Nakagawa, M. A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS J. on Computing* 15 (2), 2003, 171-190.
32. Tauscher, L. and Greenberg, S. How People Revisit Web Pages: Empirical Findings and Implications for the Design of History Systems. *International Journal of Human Computer Studies*, 47(1), 1997, 97-138.
33. Tauscher, L. *Evaluating History Mechanisms: An Empirical Study of Reuse Patterns in WWW Navigation*, M.Sc. Thesis, University of Calgary, 1996.
34. Weinreich, H. and Lamersdorf, W. Concepts for Improved Visualization of Web Link Attributes. In *Computer Networks*, 33 (2000), 403-416.
35. Weinreich, H., Obendorf, H., Herder, E. Data Cleaning Methods for Client and Proxy Logs. In *WWW 2006 Workshops: Logging Traces of Web Activity*, 2006.
36. Weinreich, H., Obendorf, H., Herder, H., Mayer, M. Off the Beaten Tracks: Exploring Three Aspects of Web Navigation. In *WWW 2006*, ACM Press (2006), 133-142.
37. Weinreich, H., Obendorf, H., Mayer, M., Herder, E. Der Wandel in der Benutzung des World Wide Webs. In *Mensch und Computer 2006*, Oldenbourg (2006), 155-164.