Mission Impossible? Capturing rich yet natural user behaviour on the Web across locations and devices

Kirstie Hawkey
Faculty of Computer Science, Dalhousie University
6050 University Avenue
Halifax, NS, Canada B3H 1W5
hawkey@cs.dal.ca

ABSTRACT
It is often desirous to obtain a full picture of users' web behaviours and activities across a variety of browsing environments. This is challenging due to the wide variety of "normal" environments possible; differences in location, device, web browser, and settings fluctuate both across participants and for each participant. Web data collection tools are required that are able to capture a broad range of data elements and work within multiple environments. This paper reflects on the data collection process undertaken as we investigated how people perceive the privacy of visited pages if others are able to later see traces of that activity. We wanted to capture the majority of participants' web browsing and we did not want our logging of this information to affect their normal browsing environment or patterns. Our data collection choice, a browser helper object, resulted in a minimal impact on the participants' web environment, but limited the types of contextual data we could collect and restricted the population we could study. This paper also includes a discussion of ongoing data analysis challenges and a visualization tool, VOLTS.

Keywords
Web browsing behaviour, field study, client-side logging

1. INTRODUCTION
Studying users' behaviour on the Web is complex because their behaviours can be influenced by a number of factors, such as task [10], motivation [12], and individual differences [7] such as domain expertise [9]. It is important that focused laboratory studies and attitudinal surveys are augmented with field research. Web behavioural studies in a field setting can often provide a more realistic picture of behaviours than can be evoked in a controlled laboratory setting as the tasks are more apt to be motivated by the users themselves and not by the researchers. Furthermore, in the field, participants have access to their usual web tools, browsers, physical environments, and the tasks.

Observing and recording natural behaviours in a dynamic environment such as the Web is challenging. One common method of studying user behaviour in a field environment is through the collection of logged data. This method is unobtrusive to the user and provides researchers with an overview of the user's behaviour. However, logged data by itself does not provide a full understanding of users' activities, goals, attitudes, and processes. Contextual information plays an important role in how we understand and interpret people's everyday behaviour. Information that provides additional details about people, such as their location or task, can help us better understand and interpret their actions. In a web environment, contextual information can be used to determine the activity in which a user is engaging, motivations for engaging in that activity, as well as perceptions about the current tool or the information being viewed.

It is difficult to capture natural web browsing behaviour without altering the browser environment which includes many factors such as the user's physical location and their usual browser application, including all normal settings and convenience features such as user-installed toolbars. Furthermore, normal contexts of web usage can occur across different locations (e.g. home, work), devices (laptop, desktop), and with different web browsers and settings and purposes in these environments [4]. It is important that we not only capture the actual behaviours across these various contexts, but we record the specific aspects of the context that may be influencing behaviour at the time.

There is a tradeoff between the ability to capture rich data about browsing activities across all contexts of use and the ability to maintain the participants' normal web browsing environment as well as implementation costs inherent to each data collection methodology (see [6] for a discussion of the costs and benefits of various logging methods). This paper describes our experiences with client-side logging through a Browser Helper Object (BHO) that works with Internet Explorer (IE) during two field studies. After presenting the study methodologies and challenges, we then reflect on how effective this approach was at capturing rich information and at maintaining a natural browsing environment. We also reflect on the ease of implementation and our ongoing challenges with visualizing the captured data.

2. OUR EXPERIENCES
We needed to collect web browsing activity in the field in order to support our study of incidental information privacy. Incidental information is defined as the information visible on computer displays during co-located collaboration that is incidental to the task at hand. Web browsers were selected as the representative application for this research as they are often used during co-located collaboration to find information or share previously found web sites. In addition, web browsers are used for a wide variety of tasks, both personal and work related which may have different privacy sensitivities. Web browsers have many convenience features, such as History, Auto Complete, and Favorites, that assist users when navigating to previously visited pages, but also display traces of prior activity that users may prefer to remain private. The nature of these traces often leads to their unintentional viewing. For example, Auto Complete will reveal search terms previously entered; during a search for "privacy research" a previous search for "personal bankruptcy laws" may be revealed.

We next discuss the two field studies we have conducted to date. For each we will describe our research goals and the challenges inherent in meeting those goals. We then present the data collection methodologies we employed, both for the logging of browsing activity and participant annotation of their log data.

2.1 Field Study 1

2.1.1 Research Goals

During the first field study (see [3] for details), our goals were to investigate participants' privacy concerns if traces of their web browsing activity were later viewed during collaboration around their personal display. Our hypothesis was that people would be willing to organize their information across a small number of privacy levels or gradients. A 4-tier privacy scheme (public, semi-public, private, don't save) was proposed to see if that level of granularity was appropriate to reflect the privacy needs between types of web sites and potential viewing audience. We also wanted to explore the existence of privacy patterns on a per-browser window basis to evaluate the feasibility of different privacy management approaches.

2.1.2 Research Challenges

The choice of the data collection tools presented several challenges. First, we needed to explore normal web browsing activities to see if privacy patterns existed. Therefore, it was important that the experimental software not interrupt the flow [1] of participants' web browsing. Second, we also wanted to maintain the participant's normal web browsing environment (i.e. their usual web browser with all the convenience features and settings intact). Finally, we were also concerned about participants' privacy; we did not want the recording of the sites visited to impact their normal web browsing activity.

2.1.3 Methodology

This week-long field study was conducted in August 2004. To qualify for inclusion, participants needed to be Internet Explorer (IE) users who performed the majority of their web browsing on a laptop computer. This allowed us to capture most of participants' personal and work/school related web browsing as they moved between these locations with their laptop.

Data collection during this study consisted of two primary data elements. The first element consisted of a record of the web page visits, including the date/time stamp, page title, URL, and the browser window in which the page visit occurred. The second element consisted of participants' perceived privacy of their web usage. Standard logging tools did not support our data collection requirements. Although there are several research and commercial logging tools that record visited page data, none include the browser window ID. We therefore had to develop two client-side data collection tools: one to log users' web activities and the other to allow participants to annotate their web activity.

2.1.3.1 Logging of Browsing Activity

The ability to maintain participant privacy (recording data locally) and to gather rich information about user activity on a per-window basis led us to a client-side solution. To record the browsing activity of participants, a browser helper object (BHO) was implemented to work with IE. A BHO is a .dll file that loads every time IE loads. As each IE window opens, the BHO loads and logs all web sites visited until the window closes. For this study, the visited web page (URL and page title), time stamp, and ID number of the browser window were recorded. All pages viewed in the browsing process were logged, even if navigation continued before the document fully loaded. Individual frames or images loaded within a web document were not logged, just the complete document. An advantage of the BHO was that the users' browsing environment did not change; they continued using IE with their normal settings intact.

2.1.3.2 Participant Annotation of Log Data

An ongoing challenge within the research community, is qualitative annotation of log data [11]. An electronic diary was developed to allow participants to assign privacy gradients to their web browsing on a daily basis (similar to that shown in Figure 1). The diary displayed all the logged data and required participants to indicate how they would classify the privacy level of each web page they visited if others were able to view the history of this activity later. Participants could annotate individual entries with a privacy level or select multiple entries for annotation. The entries could be sorted by any field (time, URL, page title), allowing participants to easily classify groups of page visits (e.g., repeated visits to the same site). After classification, participants clicked a button to generate a report to email to the researchers. In this report, the viewing history was sanitized so that the URL and page title were eliminated (to protect participant privacy). Participants could view the data about to be sent, but could not modify it. While this approach to maintaining privacy was designed to maximize the participants' willingness to engage in their usual browsing activities, the lack of URL information means that the number of unique web sites visited or the extent of site re- visitation is unknown

2.2 Field Study 2

2.2.1 Research Goals

A second field study (see [5] for details) was needed to extend the information learned in the first study. The goal of this study was to gather additional contextual information about regular web browsing activity such as the page title, URL, and location of the browsing. We required this information to examine the relationship between the context of the browsing activity (page content, location) and the privacy comfort levels that participants applied to their web browsing. Also, we wanted to confirm that the findings from the first study were replicable with a broader population. The first field study consisted solely of laptop users with a primarily technical background. This second study

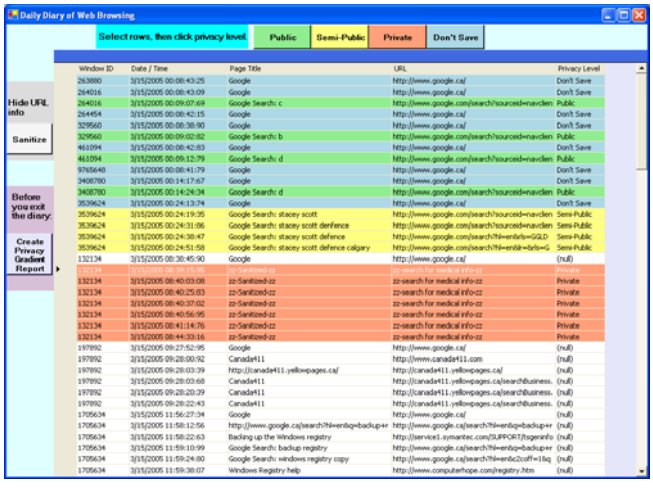

Figure 1. Screenshot of electronic diary used by participants in the second field study to annotate their web browsing with a privacy level.

included participants with varying technical experience and computers in use.

2.2.2 Research Challenges

As we wanted to not only collect the URL and page title, but to also send that information to the researchers, we needed to re-address privacy issues as we did not want to impact participants' willingness to visit sensitive sites. Participants needed to be able to selectively blind the data that was sent to researchers.

2.2.3 Methodology

This second field study was conducted in March 2005. Three different classes of participants were recruited: technical desktop users, non-technical desktop users, and non-technical laptop users. A screening process assessed participants' technical background and identified computers on which they conducted their web browsing. Participants were required to use IE and to have logging software installed on all their computer(s) so that the full picture of their personal and work/school related web browsing could be captured.

As in our previous study, data collection consisted of date/time stamp, page title and URL of visited pages, and the browser window ID. Two new elements of data collection were introduced in this study. First, the physical location of the web browser was logged (e.g., home, work, school). Second, focus events were logged so that we could determine when participants moved between windows, not just when they moved between windows for the purpose of navigating to a new page. We also logged the IE window open and close events.

2.2.3.1 Logging of Browsing Activity

The BHO was modified to record web document loading and focus events. Participants' location was hard coded for desktop computers. Laptop users indicated their current location with a radio button that appeared in a form as the browser window closed. The options listed were home, work, school, and other (a text box was provided for entry of the specific location).

2.2.3.2 Participant Annotation of Log Data

The electronic diary (see Figure 1) was modified to allow participants to sanitize entries in the diary by removing the page title and URL after applying a privacy level. Participants were asked to give a general reason for the sanitized browsing (e.g., "looking for medical information"); the default label was "no reason given". After classification, participants emailed a report to researchers. We expected that the privacy afforded by allowing participants to selectively sanitize their browsing record would contribute to their willingness to engage in normal web activities while providing us with the context for most visited pages.

3. REFLECTIONS

3.1 Effectiveness at Capturing Desired Data

We have found our data collection methodologies to be effective at capturing some types of behavioural data. In particular, we found that contextual information provided in the form of participant annotations, coupled with logs of web usage, afforded valuable insight into our participants' normal patterns of activity. The choice of the BHO did limit us in several respects. Data was limited to visited pages and basic navigation events. In order to study participants across different locations, we needed to install the BHO on each of their computers. Our sample population was also limited to those that use IE on a Windows machine.

During the second field study, we wanted to capture windows focus events; but, due to an inability to hook into the IE browser window itself, our focus events are limited to the web document. In times of rapid browsing, not all events were captured, making analysis difficult (i.e. not all on focus events match a lost focus event). Furthermore, as documents could load in the background, it could be difficult to determine when viewing of one page ended and another began. Due to time limitations, this problem was not resolved to our satisfaction. We would like to resolve this issue in order to study how people move between different browser windows and tabs while conducting browsing activities.

3.2 Effectiveness at Maintaining a Natural Environment

One of the main reasons for selecting field studies as a methodology was to capture natural user behaviour. The focus of our research included not only an investigation of the sites they visited but also of their normal patterns of activity. The BHO was ideal in that it did not impact participants' normal web browsing environment. In both studies participants could continue to use their usual browser (i.e. IE) and had access to all of their usual features, such as Favorites, History, and the Google toolbar. The automatic loading of the BHO with IE meant that participants did not have to remember to use the study instrument. As long as they were using IE on a computer with the BHO, their browsing data was captured.

In the first field study, as we did not receive the page title and url of visited sites, we have no way of knowing if the browsing captured was indicative of normal behaviours. We were able to inspect the visited pages in the second field study. We did not observe a large number of blinded of URLs (only 6/15 participants had occasions of blinding for a total of 433/31160 page visits). We also observed several instances of adult content. The proportion of participants in the field study with instances of adult content was comparable to frequency reports of erotica viewing in a previous survey [4]. This may indicate that we have captured participants' normal web usage, including those activities not considered to be socially desirable [2].

In both field studies, participants provided privacy ratings at the end of each day using the task diary. Privacy ratings may change from one page to the next, so it would not have been feasible to interrupt the flow for each and every page to assign privacy ratings. Overall, participants did not find it problematic to assign their privacy ratings at the end of the day. The electronic diary also allowed them to return to their annotations at a later time if they were unable to complete their daily classification. In the second field study, location information was provided by laptop users in real-time through a browser pop-up window. We did not expect that participants would be able to accurately assign location information at the end of the day for all of their web usage, especially if they accessed the web from several locations. We were therefore willing to accept occasional interruption of flow for the benefit of more accurate location information. In order to minimize the disruption, the pop-up window appeared when a browser window was closing.

3.3 Implementation Challenges

It is difficult to develop experimental software that is sufficiently robust in the field for multiple versions of the operating system and web browser. For example, during the course of the first study, five participants had difficulties with their software, their hardware, or their internet connections. These participants did

complete seven days of the study, but the days were not all consecutive. Maintaining frequent backups on the participants' laptops of the raw data log and the sanitized report allowed recovery of data when problems arose. Asking participants to email the data daily allowed us to question gaps in received data, then investigate and fix problems as they arose.

Over time it can be difficult to continually update and refine tools to work with new versions and features of commercial software. The BHO would likely require modifications to work with the latest IE release which includes tabbed browsing.

3.4 Data Analysis Challenges

One other aspect that has remained challenging for us is the visualization of the data generated by the logging tools (see [8] for a framework of challenges in extracting information from logged data). There can be thousands of page visits per participant and it is not enough to rely on descriptive statistics for the data, patterns of activity also need to be examined. There are often multiple attributes related to each visited page (e.g. content category, privacy level, secure/non-secure page, browser window, etc.). Visualization of the time series data is complicated by frequent rapid bursts of browsing activity and long periods of inactivity. Karen Parker of UBC has been working to provide us with a visualization tool that not only addresses the specific needs of our project, but that can be dynamically customized to reflect the needs of other researchers with similar types of logged data.

The VOLTS (Visualization of Logged Time Series) tool[1] allows users to view logged events over a given time period in a Gantt-chart-like display with a horizontal axis representing time, a vertical axis defined by the user (in our case, browser windows), and logged data events appearing as discrete, colour-coded boxes within a bar representing a channel of information. The system provides for a user-defined primary channel of information (in our case, the privacy levels of visited pages) as well as an un-limited number of user-defined secondary channels (in our case attributes such as secure/non-secure pages, primary content category, location, etc.) This tool will allow researchers to dynamically select which attributes of their data they would like to visualize and which values they would like to focus on through customizable colour coding (e.g. set colours for a few values and the remainder to black to highlight specific values of interest). Once development is complete, we will assess the effectiveness of VOLTS at addressing our analysis requirements.

4. CONCLUSIONS

This paper has reflected upon our research investigating how people perceive the privacy of visited pages if others are able to later see traces of that activity. We wanted to capture the majority of participants' web browsing and we did not want our logging of this information to affect their normal browsing environment or patterns. Our data collection choice, a browser helper object, resulted in a minimal impact on the participants' web environment, but limited the types of contextual data we could collect and restricted the population we could study. We also presented implementation and data analysis challenges. We hope that these reflection are useful to others embarking in research requiring the collection of natural web browsing activity and continue in our quest for methods of capturing rich yet natural user behaviour on the Web across multiple locations and devices.

[1] www.jkparker.ca/volts

5. ACKNOWLEDGMENTS

Thanks to Melanie Kellar for help with the BHO, to Karen Parker for taking on the visualization challenges, and to Kori Inkpen and the members of the EDGE Lab for their support. This research is funded by NSERC and NECTAR.

6. REFERENCES

[1] Chatterjee, P., Hoffman, D. L. and Novak, T. P. (2003). Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science* 22(4): 520-541.

[2] Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research* 20: 303-315.

[3] Hawkey, K. and Inkpen, K. (2005). Privacy gradients: exploring ways to manage incidental information during co-located collaboration. In *Proc. of CHI 2005*, Portland, OR, 1431 - 1434.

[4] Hawkey, K. and Inkpen, K. (2006). Keeping up Appearances: Understanding the Dimensions of Incidental Information Privacy. In *Proc. of CHI 2006*, Montreal, Canada, To appear.

[5] Hawkey, K. and Inkpen, K. M. (2006). Examining the Content and Privacy of Web Browsing Incidental Information In *Proc. of WWW 2006 (to appear)*, Edinburgh, Scotland.

[6] Hawkey, K. and Kellar, M. (2004). Recommendations for reporting context in studies of web browsing behaviour. Dalhousie University, Halifax, NS. Technical Report No. CS-2004-16.

[7] Herder, E. and Juvina, I. (2004). Discovery of Individual User Navigation Styles. In *Proc. of the Workshop on Individual Differences in Adaptive Hypermedia (Adaptive Hypermedia 2004)*, Eindhoven, The Netherlands.

[8] Hilbert, D. and Redmiles, D. (2000). Extracting Usability Information from User Interface Events. *ACM Computing Surveys* 32(4): 384-421.

[9] Holscher, C. and Strube, G. (2000). Web Search Behavior of Internet Experts and Newbies. In *Proc. of WWW 2000*, Amsterdam, The Netherlands, 337-346.

[10] Kellar, M., Watters, C. and Shepherd, M. (2005). A Field Study Characterizing Web-based Information Seeking Tasks. Dalhousie University, Halifax, NS. Technical Report No. CS-2005-20.

[11] Kort, J. and de Poot, H. (2005). Usage analysis: Combining Logging and Qualitative Methods. Workshop presented at CHI 2005.

[12] Loeber, S. C. and Cristea, A. (2003). A WWW Information Seeking Process Model. *Educational Technology & Society* 6(3): 43-52.