Chapter 7.7
Privacy Concerns for Web Logging Data

Kirstie Hawkey
University of British Columbia, Canada

ABSTRACT

This chapter examines two aspects of privacy concerns that must be considered when conducting studies that include the collection of Web logging data. After providing background about privacy concerns, we first address the standard privacy issues when dealing with participant data. These include privacy implications of releasing data, methods of safeguarding data, and issues encountered with re-use of data. Second, the impact of data collection techniques on a researcher's ability to capture natural user behaviors is discussed. Key recommendations are offered about how to enhance participant privacy when collecting Web logging data so as to encourage these natural behaviors. The author hopes that understanding the privacy issues associated with the logging of user actions on the Web will assist researchers as they evaluate the tradeoffs inherent between the type of logging conducted, the richness of the data gathered, and the naturalness of captured user behavior.

INTRODUCTION

Privacy is an important consideration when conducting research that utilizes Web logs for the capture and analysis of user behaviors. Two aspects of privacy will be discussed in this chapter. First, it is important that governmental regulations, such as the Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada, or organizational regulations, such as a university's local research ethics board (REB) policies, are met. These regulations will dictate requirements for the storage and safeguarding of participant data as well as the use, re-use, and transfer of that data. Secondly, researchers may also find that providing privacy enhancing mechanisms for participants can impact the success of a study. Privacy assurances can ease study recruitment and encourage natural Web browsing behaviors. This is particularly important when capturing rich behavioral data beyond that which is ordinarily recorded in server transaction logs, as is generally the case for client-side logging. It is this second aspect of privacy that will be the primary focus of this chapter.

DOI: 10.4018/978-1-59904-974-8.ch005

There are privacy concerns associated with viewing and releasing Web browsing data. Web browsers are typically used for a wide variety of tasks, both personal and work related (Hawkey & Inkpen, 2006a). The potentially sensitive information that may be visible within Web browsers and in data logs is tightly integrated with a person's actions within the Web browser (Lederer, Hong, Dey, & Landay, 2004). Increasingly the Internet has become a mechanism by which people can engage in activities to support their emotional needs such as surfing the Web, visiting personal support forums, blogging, and investigating health concerns (Westin, 2003). Content captured within Web browsers or on server logs may therefore include such sensitive items as socially inappropriate activities, confidential business items, and personal activities conducted on company time, as well as more neutral items such as situation-appropriate content (e.g., weather information). Visual privacy issues have been investigated with respect to traces of prior Web browsing activity visible within Web browsers during co-located collaboration (Hawkey, 2007; Hawkey & Inkpen, 2006b). Dispositional variables, such as age, computer experience, and inherent privacy concerns, combine with situational variables, such as device and location, to create contextual privacy concerns. Within each location, the social norms and Web usage policies, role of the person, and potential viewers of the display and users of the device impact both the Web browsing behaviors and privacy comfort levels in a given situation. The impacted Web browsing behaviors include both the Web sites visited, as well as convenience feature usage such as history settings and auto completes. Furthermore, most participants reported taking actions to further limit which traces are potentially visible if given advanced warning of collaboration.

Recently the sensitivity of search terms has been a topic in the mainstream news. In August 2006, AOL released the search terms used by 658,000 anonymous users over a three month period (McCullagh, 2006). These search terms revealed a great deal about the interests of AOL's users, and their release was considered to be a privacy violation. Even though only a few of the users were able to be identified by combining information found within the search terms they used, AOL soon removed the data from public access. This data highlighted the breadth of search terms with respect to content sensitivity as well as how much the terms could reveal about the users in terms of their concerns and personal activities.

In addition to taking actions to guard visual privacy within Web browsers, users may also take steps to guard the transmission of their personal information online. When concerned about privacy as they interact on the Web, users may opt to mask their identities by using a proxy server or other anonymizing (Cranor, 1999). The Platform for Privacy Preferences Project (www.w3.org/P3P/) has developed standards that facilitate user awareness of the privacy policies that govern the use of their personal information at participating websites. Research into online privacy generally examines issues concerning the transfer of personal data to business or governmental entities; the relationships are between consumers and corporations. This may be quite different from the privacy concerns associated with others viewing traces of previous Web browsing activity, as in the case of logged Web browsing data in a research context. Although in both cases personal information may be viewed, there are differences in the nature of the relationship to the viewer of the information. When the viewers of the captured information are not anonymous but are known to the user, privacy concerns may be heightened (Lederer, Mankoff, & Dey, 2003).

Field research theoretically allows the study of actual behaviors in a realistic environment. However, the act of observing or recording participants' personal interactions may cause them to alter those behaviors (McGrath, 1995). This is often referred to as the Hawthorne Effect. For example, behaviors deemed to be socially inappropriate (Fisher,