Int. J. Human-Computer Studies 74 (2015) 1–13

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

An empirical characterisation of file retrieval ☆

Stephen Fitchett, Andy Cockburn *

University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 4 February 2014
Received in revised form
20 June 2014
Accepted 3 October 2014
Communicated by K. Hornbaek
Available online 14 October 2014

Keywords:
Personal information management
Retrieval
Files
Navigation
Search

ABSTRACT

Retrieving files on personal computers is a fundamental component of interaction, yet there is surprisingly little empirical data characterising how it is carried out in realistic settings. We developed software, called FileMonitor, to dynamically record users' file retrieval activities, including data describing the files retrieved and the tools used to retrieve them. We then deployed the system in a four week log study of 26 participants' actual file retrievals on their personal computers. Follow-up interviews contextualised the findings. Results are presented in two sections focusing on the files (the number of files, patterns of revisitation, file types, etc.) and on the interface mechanisms used to retrieve them (file browsers, search tools, 'recent files' lists, etc.). We conclude by discussing implications for the design of next-generation file retrieval interfaces.

1. Introduction

Computer-based work normally begins with an explicit user action to open or retrieve a file, such as a word processing document, spreadsheet, or PDF document. Many alternative tools can be used to retrieve files, including hierarchical file browsers (such as Windows' File Explorer and the Mac OS X Finder), tools to view recently accessed files, and search utilities such as Mac OS X's Spotlight. Understanding how these tools are used is important in assisting the development of next-generation file retrieval user interfaces.

The capture and description of how people retrieve files is best supported by observations that are: (1) *longitudinal*, to reveal patterns of behaviour over time; (2) *in context* of everyday activities, rather than the result of some artificial stimuli; (3) *involve a large and heterogeneous sample* to capture a broad range of behaviours and to facilitate statistically rigorous analysis. Different experimental methods vary in the degree to which they support these objectives, with Kelly and Teevan (2007) reviewing the strengths and limitations of methods for studying personal information management.

Log-based analysis is an attractive research method for characterising many forms of user activities. By equipping software to automatically capture users' actions, it is possible to unobtrusively monitor long-term patterns of behaviour by large user groups. Log analysis has been used in several studies of personal information management such as web navigation (Tauscher and Greenberg, 1997; Tyler and Teevan, 2010), email management (Whittaker et al., 2011; Elsweiler et al., 2011), various forms of search (Teevan et al., 2007; Dumais et al., 2003), window switching (Tak, 2011), and the flow of file resources between applications (Jensen et al., 2010). Of course log analysis is not a research panacea, and its limitations include the difficulty in knowing why the user carried out an event, whether the event was successful or erroneous, and determining the starting point for an action sequence.

Given its importance in everyday computer use, there are surprisingly few (if any) log-based studies of file retrieval. Reasons for this are suggested by Bergman et al. (2011): first, there are privacy concerns in monitoring file use that can act as a disincentive for participation as well as a barrier for human-ethics approval processes; and second, developing robust logging software is a non-trivial software engineering exercise that, when done poorly, results in unacceptable instability in the users' computing environment. The studies that come closest to a log-analysis of file retrieval are Agrawal et al. (2007), who examined annual snapshots of the file systems of ~60,000 Microsoft employees but did not examine retrieval methods, and Li et al. (2010), who used file retrieval logs to evaluate task identification algorithms.

This paper describes the results of a four week study characterising file retrieval by 26 participants. The study analyses explicit retrieval of named files (such as word processing documents, spreadsheets, etc.), rather than implicit retrieval of files through applications such as email clients. The participants ran our logging tool, called FileMonitor, on their personal compu-

☆This paper has been recommended for acceptance by Hornbaek.

* Corresponding author.
E-mail address: saf75@cosc.canterbury.ac.nz (A. Cockburn).

http://dx.doi.org/10.1016/j.ijhcs.2014.10.002
1071-5819/

ters for four weeks, after which we interviewed them to gain contextual insight and to clarify unexpected or anomalous observations.

Results provide a characterisation of actual file retrieval, presented in two sections. The first focuses on *retrieved files*, describing the types of files used, their locations, frequency of access, folder management, etc. The second focuses on *retrieval methods*, characterising the use of tools for retrieval, such as search, file browsers and recent items tools. We finish by discussing design implications for next generation file retrieval tools.

2. Related work

Personal information management is an active and broad area of research, and a complete review is beyond the scope of this paper (see Jones, 2010 for a general introduction). This section briefly reviews key literature describing how users organise and retrieve information, beginning with a comparison of the methods used across different domains, such as physical files, email and the web. Section 2.2 then summarises key findings on electronic file management and retrieval, based on the file management subtasks identified by Barreau (1995): acquisition of items, organisation of these items, maintenance of information, and information retrieval.

2.1. Information retrieval across information domains

To help understand how best to support electronic file management, many researchers have examined how office workers manage paper documents (e.g., Kwasnik, 1989; Case, 1986; Mackay, 2003; Whittaker and Hirschberg, 2001; Malone, 1983). One common finding is that spatial properties of document placement can be an important facet for retrieval, with paper documents often organised in ordered piles (Lansdale, 1988). Malone (1983) also found that spatial organisation on the desk often served an important role in reminding the user about current tasks. Whittaker and Hirschberg (2001) conducted a study of the paper document collections of workers who were shifting offices, finding that participants kept physical documents for a variety of reasons, including their ready availability, to act as reminders, for certainty of retrieval in case electronic copies became unavailable, and just in case they later turn out to be useful.

Email information management is increasingly challenging. Many people receive hundreds of messages each day (Fisher et al., 2006), and studies suggest that the size of email archives increased ten-fold between 1996 and 2006 (Whittaker and Sidner, 1996; Fisher et al., 2006). Mackay (1988) identified two stereotypical strategies for managing email: *prioritisers*, who use a set of rules (either manual or automatic) to sort email messages based on priority, and *archivers*, who maintain a large number of folders that are subject-based, rather than priority-based. More recently, Whittaker and Sidner (1996) classified email users as either *no filers* (who do not sort emails into folders, instead relying on opportunistic retrieval methods such as search), *frequent filers* (who minimise the size of their inbox by frequently filing messages into folders) and *spring cleaners* (who periodically sort their inbox into folders). Elsweiler et al. (2011) found users were split roughly evenly between these strategies. Fisher et al. (2006), however, found that although there was substantial variation between users in terms of folder use and inbox size, there were no clearly discriminable groups. Similarly, Boardman and Sasse (2004) found that most users employed multiple management strategies and could not easily be assigned to Whittaker and Sidner's categories.

While many users make substantial use of folders to store email (Whittaker and Sidner, 1996; Fisher et al., 2006), the effort required to maintain this organisation is significant (Bälter, 2000), with more time spent organising email than retrieving it (Bellotti et al., 2005). Furthermore, organising emails into folders may yield little benefit – Whittaker et al. (2011) found that those who frequently file emails are no more likely to revisit emails or successfully find them, and are slower revisiting items, than those who do not file extensively.

Web page retrieval is another extensively studied area of Personal Information Management, with retrievals commonly categorised as either *finding* (retrieval of content not previously accessed) or *refinding* (returning to a previously visited web page) (CapraIII and Pérez-Quinones, 2005). Several studies show that refinding is prevalent on the web, accounting for 44–81% of web page retrievals (Tauscher and Greenberg, 1997; Cockburn and McKenzie, 2001; Obendorf et al., 2007; Mayer, 2009). Web browsers provide a range of features to assist revisitation, including bookmarks, history lists, back/forward buttons, and autocomplete URL fields. While most of these features are automatically populated, bookmark collections require explicit action from the user to mark a page as a likely target for future revisitation. Abrams et al. (1998) analysed the way users organise their bookmarks, finding a strong relationship between organisational tendencies and the size of bookmark collections. Few users with under 35 bookmarks organised their collections, while users with large collections (over 300) typically used multiple hierarchy levels. Other studies suggest that users face problems maintaining bookmark collections (Jones et al., 2001), with contents quickly becoming outdated due to changing needs (Abrams et al., 1998) or website changes (Cockburn and McKenzie, 2001 showed that a quarter of bookmarks in personal collections were invalid, and that 5% were duplicates).

Web search engines provide powerful functionality for both finding and refinding (Aula et al., 2005). CapraIII and Pérez-Quinones (2005) found no significant difference between frequency of use of search engines for finding compared to refinding, and Teevan et al. (2007) found that up to 40% of search engine queries were conducted for the purposes of refinding and nearly 30% of URLs clicked in search results were clicked multiple times by the same user. Furthermore, 24% of queries were *navigational queries*, defined as queries where a single result is selected and where both the search query and result selection are identical to that of an earlier query. Tyler and Teevan (2010) found that many refinding queries are often shorter than their initial finding queries and rank the target item higher, suggesting that people learn information about the pages they visit that helps them when later searching again.

2.2. Studies of personal electronic file organisation and retrieval

Goncalves and Jorge (2003) analysed the structure of 11 participants' document file hierarchies, finding that participants averaged about 8000 files, with each folder containing an average of 13 files. Henderson and Srinivasan (2009) conducted a similar but larger scale analysis of Windows XP users, yielding similar results to Gonçalves and Jorge. They found a mean of 5850 files and an average of 11.1 files per folder. They also found that 74% of folders contained no subfolders, and that non-empty folders contained an average of 4.1 subfolders. 7.9% of folders were completely empty.

In a large-scale study with 289 participants, Bergman et al. (2012) examined how various factors affected file navigation (retrieving a file by traversing through the hierarchy using a file browser). Their method involved statically recording the state of participants' 'recent documents' list, then asking them to navigate