Personal digital document management

Sarah Henderson

Department of Information Systems and Operations Management, University of Auckland
Auckland, New Zealand
s.henderson@auckland.ac.nz

Abstract. Knowledge workers today have a lot of digital documents to manage, and most employ some sort of organizational system or scheme to help them. Most commonly used software provides the ability to create a hierarchical organization, but the appropriateness of this structure for personal digital document management has not been established. This research aims to understand how people currently organize their documents, identify the strengths and weaknesses of current systems and explore the usefulness of other information structures. This will provide insight into how personal digital document management systems can be made more usable.

1 Introduction

Personal digital document management is the process of acquiring, storing, managing, retrieving and using digital documents. It is personal in the sense that the documents are owned by the user and is under their direct control, not that they necessarily contain information about the user [6]. Information overload is making document management increasingly difficult. Farhoomand and Drury found that the two most common definitions of information overload were "an excessive volume of information" (reported by 79% of respondents) and "difficulty or impossibility of managing it" (reported by 62%) [4].

One large part of managing documents involves organizing them so that they can later be easily retrieved. Most current software provides a facility to organize documents in a hierarchical set of folders. This organization scheme was adopted over 40 years ago to provide efficient access to files on disk. Although hierarchies are a very powerful and natural organizing scheme, there is no clear reason why these systems must use hierarchies, nor is there evidence that they are necessarily the best option for document management.

Understanding how the current hierarchical model supports users in organizing documents, and more crucially, where it doesn't, is important to being able to develop more usable systems that better support personal document management.

2 Previous Research

Previous work has included studies of how people manage and use paper documents [8, 12], email [3, 9, 13] and files [1]. Findings included identifying two main types of structuring approaches: 'neat' and 'messy' [7, 8], as well as the use of information for reminding people of tasks or events. The two studies of files revealed that many people did not create any kind of digital organizational structure at all [1], and that people used their knowledge of the locations of files to retrieve them again in preference to searching for files.

Technology has changed significantly since some of these findings were published. For example, in the two studies of files that were published in 1995, some of the participants were limited to file and folder names of 8 characters in length (plus a 3 character extension), and many did not have access to a hard drive to store information. Also, the command line interfaces used by some participants did not allow visualization or direct manipulation of information structures. The features offered by current document management software are significantly different from software 8 years ago; hence user interaction with this software is likely to have changed. How current software supports personal document management still needs to be investigated.

Other researchers have created experimental prototypes to explore alternative systems of organizing personal information such as documents. These include primarily logical/topical [2], temporal [5] and spatial metaphors [10, 11]. Many of these researchers appear to operate from the premise that the current predominantly hierarchical system of organization is inadequate for document management, and propose a (sometimes radically) different alternative organizational scheme. Unfortunately, there is not enough information about how people currently use the hierarchical model, and where and how it is inadequate. Additionally, little attention has been given to the fact that current systems do provide some (albeit limited) ability to organize spatially (on the desktop and within folders), temporally (sorting by date last modified/accessed) and logically/topically (through folder and file names). How people actually use these features is not currently known.

3 Research Aims

The aim of this research is to understand how to build more usable software for personal digital document management. The specific objectives of this research are:

- Identify where current document management software is adequate and where it is inadequate.
- Understand how people organize their personal digital documents with current software, particularly how spatial, temporal and logical/topical facilities are used.

4 Methodology

This research uses a number of different methodological techniques in order to provide rich data about the phenomenon of document management. These include semi-structured interviews, observation, and automated data gathering using a software tool that takes a snapshot of the file system. The participants are staff at the University of Auckland Business School, which uses the Microsoft Windows operating system. Twenty participants in total will be included in the study, ten academic and ten administrative staff.

Interviews. The semi-structured interviews ask the participants basic demographic information and then the participants are asked to give a tour of their file systems and email. File System Snapshot software is run during the interviews. These interviews will be fully transcribed and analyzed. This will be used to understand how people structure their file systems, and how these structures have evolved over time. These techniques should provide a thorough understanding of the subjective aspects and rationale for people's current organizations.

File System Snapshot. This software collects information about the folder structures and file names in the file system, and the folder structures used in the email system. It also stores the structure of Internet Bookmarks, My Favorites and captures a screenshot of the Desktop. Software to analyze this data is being written as part of this research. The information gathered will provide an objective empirical description of how people currently organize information, which can be compared and contrasted with the subjective description gained from the interviews.

Document Use Monitoring. Software will be installed on the participants' computers that will track their document management activities over an extended period of time (1-5 days). This will record all document open and close events, document creation, deletion, renaming, copying and moving. The information gathered will provide objective data about how people use their documents over time. It is anticipated that this monitoring will occur with 4 or 5 participants only.

5 Pilot Study Results

A pilot study has been conducted with 4 administrative participants, involving an interview and file system snapshot. The interview data has been transcribed and coded (with the assistance of QSR NVivo qualitative analysis software).

The most troublesome problem reported by the participants was managing different versions of documents (reported as a significant problem by three participants) Two reported trouble identifying where the most recent version of a document is (whether in the email system or the file system, and in which folder). Three had systems in

place for tracking multiple versions of documents using conventions based on file name, folder name or folder location.

The data collected by the file system snapshot software has been analyzed to reveal basic statistics about the file structures used by each participant, as shown in Table 1. Only folders nominated by the participant as document directories were included in this analysis (for instance, the Windows and Program Files directories were always excluded).

Table 1. File System Snapshot summary data. This shows some basic statistics about the file systems of the pilot participants

Metric	A	B	C	D
Years Experience	3	3	3	10
Files	4,395	44,196	3,793	1,545
Folders	426	7200	854	211
Files per Folder	10.3	6.2	4.5	7.3
Maximum Depth	6	16	11	8
Average Depth	2.6	5.9	6.2	3.8
Duplication (same name)	6.3%	80.1%	14.1%	14.5%

What these statistics show is firstly, the variation in the size of the collections managed by these participants, and also some very different patterns of use. For example, participant A has a relatively high number of files per folder and a shallow hierarchy, indicating a classic 'non-filer' who tends not to spend much time on organizing files, and relies more on search to locate them. In contrast Participant C has a low number of files per folder and tends towards deeper hierarchies, indicating a 'frequent filer' who stays organized and uses the hierarchy to locate documents.

The duplication figure counts the proportion of files that have the exact same name as another file. This is likely to understate the true duplication figure, as a copy of a file with a different version number would not be counted as a duplicate. The relative magnitudes of the duplication figures correlate well with the severity of the version management problem as reported by the participants.

6 Discussion

Much of the version management problem centers on the difference between files and documents. The participants are attempting to manage documents, using an interface that supports the management of files. As far as the user is concerned, a document is a structured set of information, to which changes and events occur over time. A user might talk about a status report that went through five drafts, was edited once by the boss and sent to a client. However this is actually represented as six separate files in the file system plus two in the email system, with no relationship between any of them (except perhaps a similar file name, but that is up to the user). An interface that recognizes and manages documents (rather than files) could help overcome the version management problems reported by these participants.

6 Future Work

Additional interviews and file system snapshots are planned with both academic and non-academic participants. In addition, some participants will have their document use over time monitored. More comprehensive analysis of the file system snapshot data will also be carried out, including age profiles of files.

References

1. D. K. Barreau and B. A. Nardi: Finding and Reminding: File Organization from the Desktop. SIGCHI Bulletin, 27 (3). (1995) 39-43.
2. P. Dourish, W. K. Edwards, A. LaMarca and M. Salisbury: Presto: An Experimental Architecture for Fluid Interactive Document Spaces. ACM Transactions on Computer-Human Interaction, 6 (2). (1999) 133-161.
3. N. Ducheneaut and V. Bellotti: E-mail as Habitat: An Exploration of Embedded Personal Information Management. Interactions, 8 (5). (2001) 30-38.
4. A. F. Farhoomand and D. H. Drury: Managerial Information Overload. Communications of the ACM, 45 (10). (2002) 127-131.
5. E. Freeman and D. Gelernter: Lifestreams: A Storage Model for Personal Data. SIGMOD Bulletin, 25 (1). (1996) 80-86.
6. M. Lansdale: The psychology of personal information management. Applied Ergonomics, 19 (1). (1988) 55-66.
7. W. E. Mackay: More than just a communication system: diversity in the use of electronic mail. In CSCW'88 Conference on Computer-Supported Cooperative Work, Portland, Oregon, USA. (1988) 344-353.
8. T. W. Malone: How do people organize their desks? Implications for the design of office information systems. ACM Transactions on Office Information Systems, 1 (1). (1983) 99-112.
9. J. McKay and P. Marshall: The dual imperatives of action research. Information Technology and People, 14 (1). (2001) 46-59.
10. J. Rekimoto: Time Machine Computing: A time-centric approach for the information environment. In UIST'99 Symposium on User Interface Software and Technology, Asheville, North Carolina, USA. (1999) 45-54.
11. G. G. Robertson, M. Czerwinski, K. Larson, D. C. Robbins, D. Thiel and M. van Dantzich: Data Mountain: Using Spatial Memory for Document Management. In UIST'98 Symposium on User Interface Software and Technology, San Francisco, California, USA. (1998) 153-162.
12. S. Whittaker and J. Hirschberg: The Character, Value, and Management of Personal Paper Archives. ACM Transactions on Computer-Human Interaction, 8 (2). (2001) 150-170.
13. S. Whittaker and C. Sidner: Email Overload: exploring personal information management of email. In CHI'96 Conference on Human Factors in Computing Systems, Vancouver, Canada. (1996) 276-283.