An Empirical Analysis of Personal Digital Document Structures

Sarah Henderson and Ananth Srinivasan

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

Abstract. Hierarchies have long been used as useful structuring mechanisms for organizing and managing documents. This study looks at the problem of personal digital document management in the context of knowledge workers. We study and document strategies that users employ to manage the complexity imposed by the volume and variety of personal digital documents. Exploratory research was conducted, analyzing the file systems of 73 knowledge workers using Microsoft Windows in a university setting. The empirical results of this are presented, and compared to a previous study that examined the file systems of 11 users.

Keywords: Personal document management, personal information management, document organization, file system structure.

1 Introduction

Knowledge workers spend much of their time creating and using digital information. As well as being overloaded with a flood of information and data coming at them from all directions, they are also standing in a rising tide of information of their own making: the morass of reports, memos, articles, notes, presentations, graphics, contacts, web URLs, emails, tasks and appointments that they have been slowly but surely creating and accumulating on their computer. We refer to this collection as personal digital documents. While finding information in databases and on the web is becoming easier, finding information located on a local repository such as their own hard drive is becoming increasingly difficult as a consequence of user driven activity.

Many users will spend a great deal of time using software tools to locate, acquire, manage, communicate, process and otherwise interact with this growing plethora of digital information. Because these tasks occupy such a large amount of their time, it is important that these software tools are usable, that is, they are properly designed to effectively support information management activities. Given the ubiquitous nature of these activities, even small improvements in the usability of the tools could result in a large productivity gain for knowledge workers.

There are numerous digital information types that knowledge workers typically engage with: web pages, email, documents, images, sound, video, memos, contacts, appointments and tasks. Each of these different types of digital information has its

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part I, HCII 2009, LNCS 5617, pp. 394–403, 2009.
© Springer-Verlag Berlin Heidelberg 2009

own particular features and requirements. Due to the relative newness of web, email and multimedia technologies, management of these has been the focus of many research efforts. However, the older and more basic task of managing "ordinary" documents has gone relatively unstudied.

A hierarchical structure as a mechanism for storing and managing documents is a well entrenched systems paradigm. Most people store and manage documents through a user interface that exploits the structure of hierarchies [1]. These tools allow people to recursively create folders and place documents within folders by attaching meaning to the hierarchy thus created. Using this simple containment mechanism, people can build up a large hierarchical structure of folders. This basic paradigm has not changed in the decades since its introduction, although the user interface to it significantly improved with the widespread introduction of graphical user interfaces.

A basic principle of user interface design is that the design of a tool should be thoroughly grounded in an understanding of how the user works, what tasks they perform and how those tasks are carried out. However, with personal digital document management, very little research has been done with regard to investigating how people actually manage their documents and what the requirements are for document management tools. This study attempts to address this knowledge gap by empirically examining document structures that knowledge workers create for themselves.

2 Background

Previous research on personal document management (and personal information management in general) can be divided into two main approaches. The first strand of research examines how people manage various forms of personal information, and the second strand develops and tests new user interfaces and systems for the management of personal information.

In considering how people manage their documents, Lansdale [2] identified the trade-off that exists between the effort spent filing a document when it is first stored and the effort required to find it again later. Many subsequent studies of both email and paper filing systems have found there are two general filing strategies which people adopt in response to this trade-off: filing and non-filing [3-6]. A person adopting a filing strategy generally tries to create a folder structure, and makes an ongoing effort to try and file new information into this structure on a regular basis. They rely on the structure to help them locate documents again, typically using browsing (location-based search) in preference to using a search tool. People adopting a non-filing strategy tend not to maintain much of a formal organization structure. Instead they rely primarily on browsing fairly unstructured lists, or using search tools to locate information when needed.

These two strategies suggest two approaches to improving tool support for information management: improving the efficiency and effectiveness of using an organization structure or improving search tools.

To support the people inclined towards a filing strategy, another strand of research has worked to create new systems that are different from the currently predominant hierarchical containment approach. Most of these are based around a particular dimension of the information that is assumed to be primary. For instance, Lifestreams [7] is

based on the premise that the most important dimension on which to organize things is time. TimeScape [8] also includes time as a primary dimension, but includes a spatial layout as well, while the Taskmaster system [9] is based on studies of email users that found that task or project is a common organizing principle. Along a similar vein, the Placeless Documents project [10] doesn't impose any type of structured organizing scheme at all, but allow the user to give attributes and attribute values to documents, which can then be used to search and group documents for viewing.

Future work on this strand of research could be helped by having more information about exactly how users currently structure their documents in the relatively unguided context of a hierarchical file system. Very little research has looked specifically at the document structures that people actually create to manage and organize their documents. Studies of paper filing systems in 1982 showed that people tend to create simple classification schemes, rarely more than 2 levels deep [3], however given the physical constraints of folders and filing cabinets, it is unlikely that a more layered system could be developed. The first study to examine computer file use and organization in 1995 [11] found that the study participants did not generally use directory structures at all, although some archived their files by placing them onto separate floppy disks. The only recent study to look specifically at file structures was conducted on 11 users of the Unix file system by Gonçalves and Jorge [12]. They looked at the total number of files and folders, the width and depth of the structure, as well as balance, and the distribution of file types.

To further this research, as part of a larger study into the personal digital document management practices of knowledge workers, we took snapshots of their file systems in order to analyse the document structures they created. The following section will describe how the study was conducted, followed by an analysis of the results. We then present a discussion of the relevance for the design of user interfaces for document management and give our conclusions and suggestions for further research on this subject.

3 Method

As part of a larger study into personal digital document management (including interviews and a survey), a snapshot of the file system of knowledge workers was taken (using custom-written software). The participants were all employees of a large university, drawn from all academic and supporting business units, and at all levels of the hierarchy. All were users of the Windows XP[1] operating system. We thought that the university setting was particularly helpful in understanding the dynamics of the problem given the proliferation in quantity and variety of digital documents that are typically found in such an environment.

The snapshot software instructed the participants to select all the locations where they store documents. The default locations were the My Documents and Desktop subdirectories, although participants could easily remove these and add other locations where they kept their document files. The snapshot was taken on their primary

[1] Microsoft Windows® is a registered trademark of Microsoft Corporation. Henceforth it will be referred to as Windows.

work computer, and could include network locations and flash memory devices, but not other desktop or home computers.

The information captured by the file system snapshot software includes the name, extension, date created, date last accessed and date last modified of every file and folder, as well as the structure of the folders and files. The data was checked and cleaned of any system folders or multi-user shared folders.

A total of 78 participants completed the file system snapshot. However, five of those only included the default locations of My Documents and Desktop, despite indicating in the other part of the study that their primary storage location was a network drive or removable drive. As a consequence, these participants only had a handful of files in the snapshot. These participant's snapshots were removed from this analysis, leaving a total of 73 snapshots for analysis. 34 participants were male and 39 were female. 48 had primarily academic responsibilities while 25 had general administrative responsibilities.

4 Results

4.1 Overall Size

The size of the document collection has an impact on the appropriate software support, since software to support the task of managing a few hundred files is going to be different from managing thousands or tens of thousands.

The mean number of files observed in the document folders was 5,850. However, the number of files ranged from a minimum of only 100 files, to a maximum of 33,902 (standard deviation of 7,605). As Fig. 1 shows, the distribution is significantly right skewed, with a median of 2,754 and a skewness statistic of 2.26.

The average number of folders was 628, with a standard deviation of 860. The distribution was also right-skewed, with the median number of folders being only 350. The smallest number of folders observed was 11, and the largest was 4,694. The participant with the highest number of folders was not the same person who has the highest number of files. As shown in Fig. 1 there are five participants with over 2,000 folders.

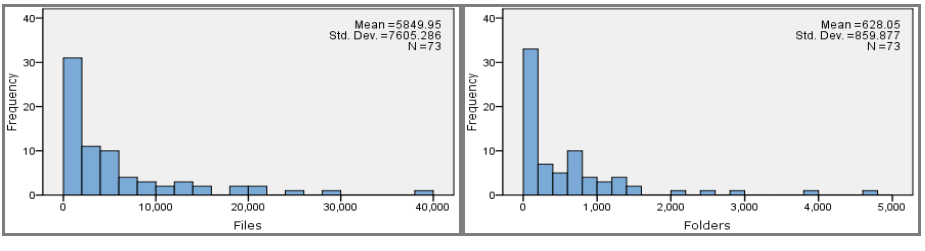

Fig. 1. Distribution of total number of files (left) and folders (right) in snapshot

As would be expected, there is a significant correlation between the number of files and the number of folders a person has in their file system (correlation coefficient

0.88) [2]. There is no correlation between the number of files or folders a person manages and any of the demographic data collected (age, gender, academic or general staff status, department, length of time they have been working in the same field, or duration of employment).

4.2 Tree Characteristics

Trees vary in several dimensions. Trees can be shallow or deep, broad or narrow, and can contain varying numbers of files in each folder. The maximum depth for each participant is the depth of the deepest folder in their document collection.

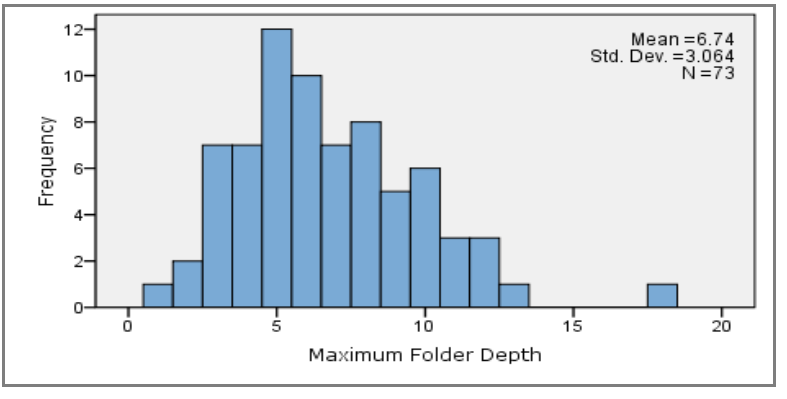

Fig. 2. Frequency distribution of the maximum depth of each document structure

There is significant variation in the (maximum) depths of folder structures. The shallowest structure was only 1 level deep, while the deepest was 18 levels deep. The average of the maximum depth across all participants was 6.8, with a standard deviation of 3.1. There is a significant correlation between the number of files a person has and the average depth of their file system ($r = 0.78$).

The width of a tree is determined by the average number of subfolders in each folder. On average 74% of folders did not contain any subfolders at all, only (possibly) files. These are considered leaf folders, and are not included in the average subfolders metric. The interior (non-leaf) folders by definition must contain at least one subfolder. The mean number of subfolders per folder was 4.1, with a standard deviation of 1.3. The highest average observed was 9.5, and the lowest was 1.8 subfolders per folder. The distribution of this metric is shown in Fig. 3.

There is no correlation between the average number of subfolders per folder and the total number of files and folders in the file system, so both small and large systems do not differ in their average breadth. There is also no significant correlation between the average depth of the tree and the average number of subfolders. Since depth does vary with the total size, this would imply that the bushiness of the tree varies independently of these factors.

[2] All correlations reported are statistically significant at the 0.01 level unless otherwise stated.

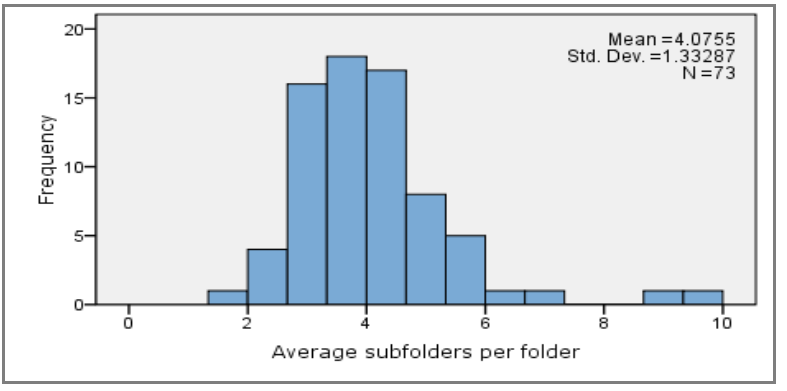

Fig. 3. Measure of bushiness - frequency distribution of number of subfolders per folder

Rather than comparing the average values for each user, we can also do the comparison at the individual folder level. There is a high average number of subfolders at the root of the tree (approximately 9), which sharply drops off (to less than 2) by two or three folders down.

In order to compare the result of this study directly to the study conducted by Gonçalves and Jorge [12], we also calculated branching factor as a metric of bushiness. The average branching factor in this study was 1.93. This ranged from 1.27 to 2.97 and had a standard deviation of 0.34.

As would be expected for two measures of bushiness, the branching factor and the average number of subfolders are correlated (r = 0.5). In common with the average subfolders metric of bushiness, there is no correlation between the branching factor and the total number of files and folders in the file system.

There is a significant negative correlation between the branching factor and the average depth of the tree (r = -0.36), indicating that wider trees tend to be shallower. There is also a positive correlation between the branching factor and the number of top level locations (r = 0.41). This is expected, since the locations essentially represent the top level of tree branching.

One of the key differences between the branching factor and the average subfolders is that branching factor assumes a perfectly even tree, whereas the average number of subfolders is affected by the tree's unevenness.

The more files people store in each folder, the more 'leafy' their folder tree becomes. Leafiness is the average number of files per folder. Higher leafiness indicates a denser tree. The average number of files per folder across all file systems was 11.1 (standard deviation 7.8). The highest number of files observed in a single folder was 1168.

The least leafy file system had an average of 4.5 files per folder, and the leafiest averaged 64.3 files per folder. However, this was a significant outlier, with the second leafiest file system averaging under 30 files per folder.

There is no significant correlation between the average number of files per folder and the overall number of files, nor with the average depth or bushiness of the document structure. However, similar to bushiness, there is a significant correlation between the *maximum* number of files per folder and total number of files (r = 0.56).

As shown in Fig. 4, the average number of files per folder is highest at the top levels of the tree, and then drops off sharply. It is fairly constant at levels 1 to 5 of the tree and then tapers off. Note that the average file system has a maximum depth of 6.8 levels.

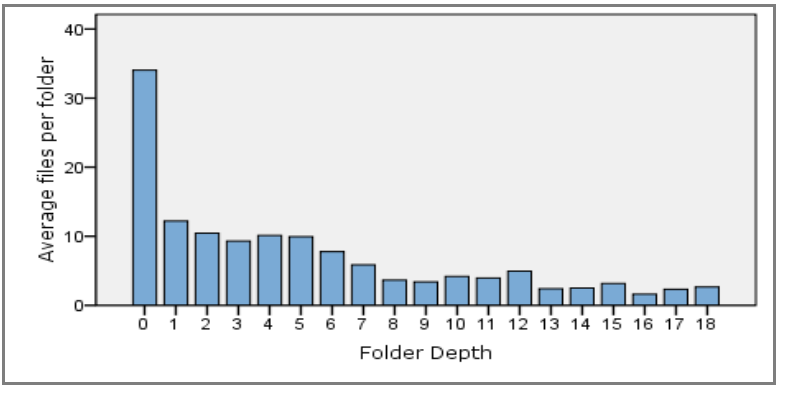

Fig. 4. Leafiness vs Depth - how the average number of files in a folder varies with the depth of the folder

To assess how even the distribution is across the tree, the standard deviation of the number of subfolders was used as a measure of balance. The lower the standard deviation, the more evenly balanced the tree. The average balance was 5.6, with a standard deviation of 3.5. While this indicates that most trees are fairly balanced, there is one significant outlier having a standard deviation of 27.9. This person has a relatively small file system of 628 folders in total. While most of their folders only have one or two subfolders, they also have folders that contain 105, 142 and 151 subfolders, giving them an extremely unbalanced tree.

There is no correlation between the balance of the tree and the overall size or the number of top level locations. Nor was there any relationship with the depth of the tree.

There is a statistically significant correlation between the balance of the tree and the bushiness, using both the average subfolder metric ($r = 0.73$) and the branching factor metric ($r = 0.42$). This would indicate that trees that are wider on average also tend to be less evenly balanced than narrower trees.

In addition to assessing how balanced the folder structure is, we can also examine how evenly distributed files are throughout the tree. The standard deviation of the number of files in a folder was used as a measure of balance. The average file balance was 23.4, with a standard deviation of 29.5. As with the folders, this was also considerably right skewed due to one outlier. This participant has 966 out of their total 1028 files in the My Documents folder itself. They have not created any folders to structure these documents, and do not appear to have made use of any of the system created folders to organise these documents.

There is no correlation between how evenly the files are distributed and the balance of the tree structure itself. There is also no correlation between the file spread and the overall size or depth of the document structure.

There is a statistically significant correlation between the file balance of the tree and the leafiness (r = 0.93). This would indicate that trees that are wider on average also tend to be less evenly balanced than narrower trees.

Empty folders are a potential sign of inefficiencies in the file structure, since the participant has expended effort created those folders but then has not made use of them. Only 3 out of the 73 file systems did not contain a single empty folder. The highest number of empty folders was 610. The mean number of empty folder was 37.8 (s.d 83.0). The distribution of the number of empty folder was extremely right-skewed (skewness 5.4), with a median number of empty folders being 13.

Since the number of empty folders can be expected to increase with the size of the file system, the proportion of empty folders is perhaps more important. The emptiest file system had 47.6% of the folders being empty.

Most file systems had only a small proportion of empty folders, with the mean proportion of empty folders being 7.9%.

4.3 Duplication

A very efficient system is likely to contain a low level of duplicated folder and file names. A measure of the proportion of duplication can be calculated from the number of non-unique files divided by the total number of files. Duplication can be calculated separately for files and folders. This measure of duplication will only reflect the fact that multiple files or folders are named identically. These files or folders may be exact copies of each other, or they may have entirely different content.

The mean level of file duplication was 21.8%. This means that on average, 21.8% of the documents in the file system have the same name as another file. The amount of file name duplication ranged from 0.4% to 60.4%. The level of folder name duplication was slightly higher, with a mean of 23.5%, and ranging from 0 to 73.4%.

There is a significant correlation between the level of duplication and the overall size of the file system. The correlation between total number of files and file duplication had r = 0.61, and the number of folders and folder duplication were correlated with r = 0.65.

There is also a significant correlation between the level of folder and file name duplication (r = 0.79). One explanation for this might be that entire folders and their contents are being duplicated together.

5 Discussion

It is surprising how similar the results of this study are to the study performed by Gonçalves and Jorge [12], despite the fact that many of their participants used different operating systems. The mean number of files that people are managing was slightly lower than their study, but we found similar levels of individual variation in file system size.

We also had fairly similar results in terms of the document structures. We found trees that were on average slightly bushier and leafier, a little deeper and a little less balanced. The larger the tree, the deeper and more unbalanced it tends to be. We

found directory trees to be slightly deeper on average than Gonçalves and Jorge found - 9.65 compared to their 8.45. However, we also noted that people tend to average a lower depth value - only 3.4 folders deep on average. In fact, most people's maximum depth is about 3 subfolders deeper than their average.

Gonçalves and Jorge found an average number of top level folders per locus of 2.75. This study differed from theirs in only considering one locus (work computer), so therefore the number of locations in this study is comparable to the number of top level folders in a locus. Our value of 3.4 was probably inflated by the fact that the snapshot software automatically included the Desktop. Many users would probably not have added it themselves if it wasn't suggested to them, and thus the figure might be lower if the users had freedom to choose their top level folders themselves.

Our average branching factor was slightly higher than that found by Gonçalves and Jorge, although well within one standard deviation of their value. However, the branching factor metric eliminates all the variability in the tree and assumes the tree is completely uniform. The average number of subfolders in a folder is a better metric, since it eliminates the leaf (empty) folders, and better reflects the actual internal structure of the tree.

In terms of visualizing small sections of the tree, no extreme techniques are required, since the tree structures are not particularly bushy or leafy.

While there were some interesting correlations to emerge, what is perhaps more interesting is correlations that were not present that might have been expected. For instance, it might be expected that participants would have a tendency to create either wide tree or deep trees. Thus there would be an expected negative correlation between depth and either bushiness or leafiness. However, no such correlation was found.

6 Conclusion

The use of hierarchies as structuring mechanisms is an inherent part of how we approach document management. With digital documents, the volume and variety that knowledge workers typically encounter present a set of issues relating to how they ought to be managed effectively. We know that a significant number of computer users rely on a structuring strategy to enable effective retrieval. This study is an attempt to observe and document the strategies that individuals employ to assist them with this task. This study examines the behavior of knowledge workers who exploit a hierarchical structure to manage their documents. The results show that users vary considerably in terms of how the hierarchy is employed to manage the complexity of the problem. The results of such studies will help us improve tools that are integral to computer systems to help users be more productive. While this study is descriptive in that it examines various dimensions of usage, the real value of such work lies in our ability to construct predictive models of usage. Such models will form the foundation of improved usability of tools that support personal digital document management.

References

1. Faichney, J., Gonzalez, R.: Goldleaf Hierarchical Document Browser. Australian Computer Science Communications 23(5), 13–20 (2001)
2. Lansdale, M.: The psychology of personal information management. Applied Ergonomics 19(1), 55–66 (1988)
3. Cole, I.: Human aspects of office filing: Implications for the electronic office. In: Human Factors Society Annual Meeting, Seattle, Washington, USA (1982)
4. Ducheneaut, N., Bellotti, V.: E-mail as Habitat: An Exploration of Embedded Personal Information Management. Interactions 8(5), 30–38 (2001)
5. Malone, T.W.: How do people organize their desks? Implications for the design of office information systems. ACM Transactions on Office Information Systems 1(1), 99–112 (1983)
6. Whittaker, S., Sidner, C.: Email Overload: exploring personal information management of email. In: CHI 1996 Conference on Human Factors in Computing Systems, Vancouver, Canada (1996)
7. Freeman, E., Gelernter, D.: Lifestreams: A Storage Model for Personal Data. SIGMOD Bulletin 25(1), 80–86 (1996)
8. Rekimoto, J.: Time Machine Computing: A time-centric approach for the information environment. In: UIST 1999 Symposium on User Interface Software and Technology, Asheville, North Carolina, USA (1999)
9. Bellotti, V., et al.: What a To-Do: Studies of Task Management Towards the Design of a Personal Task List Manager. In: CHI 2004 Conference on Human Factors in Computing Systems, Vienna, Austria (2004)
10. Dourish, P., et al.: Presto: An Experimental Architecture for Fluid Interactive Document Spaces. ACM Transactions on Computer-Human Interaction 6(2), 133–161 (1999)
11. Barreau, D.K., Nardi, B.A.: Finding and Reminding: File Organization from the Desktop. SIGCHI Bulletin 27(3), 39–43 (1995)
12. Gonçalves, D., Jorge, J.A.: An Empirical Study of Personal Document Spaces. In: Jorge, J.A., Jardim Nunes, N., Falcão e Cunha, J. (eds.) DSV-IS 2003. LNCS, vol. 2844, pp. 46–60. Springer, Heidelberg (2003)