Analyzing Personal Document Spaces

Daniel Gonçalves, Joaquim A. Jorge

Computer Science Department - Instituto Superior Técnico
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
djvg@gia.ist.utl.pt, jorgej@acm.org

Abstract

In recent years, Personal Document Spaces (PDSs) have become more complex, spanning several machines. In order to develop new approaches to help users manage their PDSs, it is important to know their contents and structure. We undertook an empirical study where we made a thorough analysis of eleven PDSs, from which we extracted several research and design guidelines.

1 Introduction

It is increasingly common for users to store their documents in several machines or locations (*locii*), the set of which constitutes a user's Personal Document Space (PDS). Furthermore, the number and diversity of documents in store are increasing. Traditional ways of document handling are quickly becoming ineffective. New tools to find specific documents or manage PDSs as a whole are becoming an imperative necessity. Those concerns led to the development of the Semantic File System [3], where document properties can be used to retrieve them. A more recent work based on this approach is the Placeless Documents [1]. The Lifestreams system [2] allows users to chronologically navigate through their PDSs. Finally, works like Satchel [4] try to manage documents across several *locii*. None of these works has, however, based itself on a thorough characterization of PDSs. The usefulness of those and future works in the area depends on their adequacy to the PDSs they must handle. Knowledge of both their structure and contents of PDSs is of capital importance. In order to acquire such a characterization, we conducted a study where the PDSs of eleven users were extensively analyzed. That analysis provided insights on the nature of PDSs, with direct implications for the development of PDS-handling applications.

2 The Experiment

We developed a computer program to analyze PDSs gathering information on all files and directories therein, across several *locii*. It records statistics about PDS sizes, the distributions of their contents, directory tree topological measures, the numbers, sizes and distributions of files by type, the dates of creation, access and modification of files, file sizes, and the elements that make up file names. The final report is produced in a human-readable format, to alleviate privacy concerns. We analyzed only directories containing documents, and not those containing applications or operating system data. In a two-week period we collected eleven reports. The participants ranged from users of computers as a work tool on office settings to a senior manager for a small software development company and computer science faculty or graduate students. In what follows, all values are averages (other values were omitted for clarity's sake). The number of *locii* in each PDS is 1.45, confirming that users are starting to handle multiple *locii*. As to file

numbers, we identified **file-rich, file-average** and **file-poor** users (10,000-25,000, 1,000–10,000 and less that 1,000 files, respectively), apparently according to user occupation: all file-rich users were teachers, and all file-poor users worked outside academia. The directory trees are narrow and medium-depth, with a branching factor of 1.84, and an average depth of 8.45. They are fairly well balanced: only 10% have significantly larger numbers of sub-directories (over nine). There are thirteen files on each directory, showing users tend to separate and classify documents whenever possible. We identified the type of over 85% of files from their extension. Text, image and archive files were the most common. Notably absent were symbolic links, PDA-related and PIM files. Image, text, PDA, presentation, source-code and spreadsheet files tend to be the main occupants of their directories, accounting for 50% of all files therein. As to PDS activity, we found that about 80% of PDSs are inactive (not used for over a month) at any given time. Finally, we found that file names are, on average, 12.56 characters long, and that nearly 60% contain numbers. Users resort to either underscores or both hyphen and spaces. Dates are seldom used (under 1%).

3 Discussion

Taking into account the narrow and not too deep tree structure, applications need not worry with handling more extreme aspect ratios in the general case. Given that file numbers in each directory are not too large, on most cases they can probably all be displayed at once, and only a handful of directories will require a special approach. Total file numbers, however, can reach the tens of thousands. This makes managing them very difficult, if not impossible. New approaches should be developed to automatically store and retrieve documents. Furthermore, they need not worry with all documents in the same way, given that 80% are not active at any given time. Archiving and retrieving techniques should concentrate on the vast majority of less-used but often necessary 'inactive' documents, where memory problems are sure to arise. Content-based automatic classification and retrieval seems a likely way to solve the problem, and should provide special care for the most common documents: text and image. The discovery of access and reading patterns for automated retrieval of documents from inferred user needs will also become necessary as PDSs grow in diversity and complexity. The low usage of symbolic links shows that multiple classifications are rare, even if necessary, confirming the difficulty of the classification task and suggesting the need for document handling approaches where no such classification is needed.

4 Conclusions and Future Work

We plan to gather more data concerning the increasingly common mobile devices, striving for a wider range of audience, both in number and diversity. We also plan to directly interview the users and record the interaction with their PDS. This will provide a more complete description of PDSs.

References

1. Dourish, P. et al. (2000). Extending Document Management Systems with User-Specific Active Properties. *ACM Transactions on Information Systems*, 18(2), 140-170, ACM Press.
2. Freeman, E. and Gelernter, D. (1996). Lifestreams: A Storage Model for Personal Data, *ACM SIGMOD Record*, 25(1), 80-86, ACM Press.
3. Gifford, D., Jouvelot, P., Sheldon, M. and O'Toole, J. (1991). Semantic File Systems. *13th ACM Symposium on Principles of Programming Languages*, ACM Press.
4. Lamming, M. et al. (2000). Satchel: providing access to any document, any time, anywhere. *ACM Transactions on Computer-Human Interaction*, 7(3), 322-352, ACM Press.