How do Students Organize Personal Information Spaces?

Sharon Hardof-Jaffe[1], Arnon Hershkovitz[1], Hama Abu-Kishk[2], Ofer Bergman[3], Rafi Nachmias[1]
{sharonh2, arnonher, nachmias}@post.tau.ac.il, hama@bgu.ac.il, o.bergman@sheffield.ac.uk
[1] Knowledge Technology Lab, School of Education, Tel Aviv University, Israel
[2] Department of Communication Studies, Ben-Gurion University, Israel
[3] Information Studies Department, Sheffield University, UK

Abstract. The purpose of this study is to empirically reveal strategies of students' organization of learning-related digital materials within an online personal information archive. Research population included 518 students who utilized the personal Web space allocated to them on the university servers for archiving information items, and data describing their directory hierarchies. Several variables for measuring folders size and depth were defined, and four of them were chosen as best representing different aspects of the user's archive structure. Then, as a result of cluster analysis of the students, four organization strategies emerged, refining the classical piling/filing classification: piling, one-folder filing, small-folders filing, and big-folder filing. Also, associations were found between the organization strategies and archive size, students' studies degree. A discussion of this study and further research is provided.

1 Introduction

Personal information management (PIM) is an emerging research field focusing on the activities by which a person keeps, saves and organizes information items in order for her or him to later retrieve them [4]. In the current knowledge age, PIM has a central role in learning processes, as students create and collect many information items, and organize them into personal information archives. During PIM activities, students construct knowledge regarding the subject matter as they collect, evaluate, choose, tag, sort, classify and name information items. The purpose of this study is to investigate students' organization strategies of personal archives using data mining techniques.

Previous research have identified two main organizational strategies for PIM: Piling and Filing [14]. The pilers are those who tend to gather many items in the main documents directory (e.g., "My Documents" for files, "Inbox" for e-mails). The filers, by contrary, tend to sort the items into labeled folders, according to some categorization. The resulted structure of the personal information space reflects the user's organization strategy, hence examining students' archives might shed light on how they deal with PIM activities [2, 8].

Over the years, PIM studies have heavily relied on traditional data-collection methodologies which usually allow only a small number of participants, thus their external validity is limited. Recently, data mining methods have been suggested as enabling identification and measurement of PIM activities and personal information space structures for large populations [9, 11]. During this study, we have investigated online storage space used by university students using data mining techniques, in order to identify students' personal information space organization strategies. Applying data mining techniques on data drawn from online storage spaces presents PIM-related research with new and fascinating opportunities, and is the core of this research.

2 Background

2.1 PIM and Learning

The nature of information has dramatically changed in the digital era, as information is easily accessible, mostly distributed, presented in multiple formats, and hypertext-oriented. While learning, students create personal information spaces, negotiating between the huge amount of available information - from various resources and environments - and their limited processing abilities at any given time. Students therefore need to acquire Personal Information Management (PIM) literacy in order to efficiently manage their own learning environment, which is normally associated with the nature of the subject matter and the assignment requirements [16].

PIM literacy [16] is not just a set of practical actions of saving and retrieving information items; it is an integral and a centric part of the learning process, as through it, and by constructing an information archive, students construct knowledge. The constructive approach to learning emphasizes the fact that knowledge is constructed through a process in which learners actively integrate new knowledge with previous knowledge [7]. During the process of information seeking, students organize collected items into an information construction, by using cognitive skills, such as naming, sorting and categorizing [13].

Bloom's Taxonomy of Educational Objectives [5] presents six levels of cognitive skills: knowledge, comprehension, application, analysis, synthesis, and evaluation. The three main filing skills (i.e., naming, sorting, categorizing) might be related to different levels in Bloom's taxonomy: a) *Knowledge* "includes those behaviors […] which emphasize the remembering, either by recognition or recall of ideas, material or phenomena" (p. 62). By *naming* a folder, the student has to recall some basic knowledge about the files within it, or to recognize their main theme, in order to define and label it; b) *Analysis* is the separation of materials or concepts into component parts, during which the student "is required to determine their connections and interactions" (p. 145) and to recognize their organizational principles. When *sorting* materials, students select the related folder(s) for each of the new information items, hence explicitly identify the relationships among the items using the hierarchy; c) *Synthesis* is defined as "putting together elements and parts so as to form a whole" (p. 162). In order to construct a personal information space, many items are being combined together to form a hierarchical structure – a process which requires *categorization* skills.

These PIM activities are part of a process of integrating new knowledge into previous constructed knowledge as any information item the student adds to her or his personal information space, is being connected to the other items by its location in the hierarchy. While information items are being connected, knowledge, analysis and synthesis skills are constantly being applied in a spiral process during which the personal information space is being formed and is continually evolving. Therefore, we believe that PIM activities have an inherent learning component.

2.2 *PIM Organization Strategies*

Malone [14] was the first to classify Personal Information Management (in the context of office organization) into two types of strategies: Piling and Filing. The Piling style is characterized by papers being heaped on top of each other (latest papers are on the top of the heap), with the pile carries no label. Filing is characterized by papers being distributed into physical files, labeled according to a certain categorization (determined by the filer). Malone found that piles were useful for small collections, where the users could still remember the location of each paper within the pile, however as piles grew users could not keep track of their papers.

The folder hierarchy is the standard mechanism for organizing personal information in digital environments. This mechanism allows users to create a personal classification scheme, based on categories and dimensions they see as relevant (e.g., role, project, time). In today's offices, papers are replaced by digital information items (e.g., files, e-mails), filing is done into directories (folders) with labels referring to their category, and piling is typically done by heaping the information items in a root directory, such as "My Documents" for files and "Inbox" for e-mails. Previous research has shown that most of the users tend to employ a mixture of Piling and Filing [19].

The binary classification of Piling/Filing was refined by many other PIM classifications, and was extended mainly to describe different filing activities over time (i.e., *when* do users file their files?) [1, 6, 20]. In the context of learning, strategies were defined regarding the creation time of new folders: a) *Pre builders* - students who create new folders before they produce any items to put in them; b) *Post builders*, who prefer to create new folders after a set of new items is collected [8]. Our study is aiming on refining the Piling/Filing classification, based on empirical data describing personal online archives.

2.3 *Data Mining Methods in PIM Research*

Data describing how users organize their personal information space had been usually collected by means of traditional research methodologies, e.g., in-depth interviews, semi-structured interviews, screen captures, and questionnaires [3, 6]. Over the last few years, data mining has been suggested as a promising methodology for PIM research, and several PIM studies have already demonstrated the strength of this approach [11, 18]. For example, Clustering algorithms were used for identifying groups of files (on desktop) having the same context, and for grouping together email messages according to their content [10, 15], demonstrating the collection and analysis of large datasets, which would not have been possible using traditional methods.

The main purpose of this study is to empirically examine personal information space organization strategies in the context of learning processes on a large population of students, in order to refine the traditional piling/filing classification.

3 Methodology

3.1 *Research Field*

Tel Aviv University enables each of its students to keep and manage personal information items on the Web, within the university's Learning Content Management System (*HighLearn* by Britannica Knowledge Systems Inc.), which serves about 26,000 students and comprises of over 4,300 courses [17]. Users of this environment can upload files, create folders, and retrieve files by navigating or searching.

3.2 *Research Population and Data File*

The study was conducted on data describing online archives of 2,081 undergraduate students, graduate students and staff who kept information items in their virtual personal directory. The data included the list of files and folders (full paths) for all the users, where each personal information space had a unique random identification. The raw data included more than 70,000 rows, each of which refers to one file or folder. Data were collected on August 2008. After excluding students with less than 10 files in their archive, a new data file for analysis was created, holding 48,744 rows of 518 students.

3.3 *Procedure*

In order to examine different strategies for personal information space organization, four variables describing the organization were chosen and computed for each student: 1) *Files per folder* – average folder size; 2) *Largest folder* – number of files in the largest folder, including root directory; 3) *Pile rate* – ratio between pile size (root directory) and archive size (total number of files); and 4) *Inner-pile rate* - ratio between the largest folder size (not including root directory) and archive size (total number of files). *Files per folder* and *largest folder* were transformed for having a maximum value of 30, and 100 accordingly, in order to normalize their distribution. Then, Two-step Cluster Analysis of the students into k disjoint groups was applied (using SPSS), in order to classify students according to their personal information space organization strategy by the four variables. After several iterations, k=4 was chosen as resulting in the best fitting clustering.

4 Results

A short descriptive statistics of the data file is given in Table 1. On average, each student has 80.52 (SD=170.17) files and 13.58 (SD=45.33) directories.

Table 1. Descriptive statistics for the four describing variables

Variable	Minimum	Median	Maximum	Mean (SD)
Files per folder	0.34	10.55	235	16.16 (23.06)
Largest folder	1	16	339	27.15 (35.24)
Pile rate	0	0.17	1	0.38 (0.40)
Inner-pile rate	0	0.20	1	0.28 (0.28)

After clustering the students according to the four variables, we have calculated means and SD for each variable within each cluster; results are given in Table 2, where maximum and minimum values for each variable are **bolded** and *italicized*, accordingly.

Table 2. Means (SD) of the four variables by which the clusters were formed

Cluster	N	Files per folder	Largest folder [# files]	Pile rate	Inner-pile rate
1	141	17.78 (7.33)	22.71 (16.60)	**.97** (.08)	*.02 (.06)*
2	49	14.70 (7.49)	18.77 (8.53)	*.09 (.11)*	**.86** (.13)
3	262	*6.10 (4.49)*	*14.52 (11.55)*	.18 (.20)	.26 (.16)
4	66	**23.10** (7.67)	**71.62** (28.00)	.13 (.19)	.48 (.27)
All	518	12.26 (8.97)	24.42 (24.19)	.38 (.40)	.28 (.28)

As might be seen from the table, Cluster 1 (n=141) is characterized by extreme values of two variables' means among clusters: *Pile rate* gets a maximum (0.97), and *inner-pile rate* gets a minimum (0.02). These results imply that in this cluster, most of the students' files are stored in the root directory (hence it is not surprising that the second largest folder is extremely small). These two extreme values of variables are typical for Piling organization strategy.

In Cluster 2 (n=49), again the means of the same two variables as in Cluster 1 get to their extreme values, however in different direction. In this cluster, the mean of *pile rate* is minimal (0.09), and we may think that this is a non-piling strategy. However, the mean of *inner-pile rate* is relatively high (0.86), which indicates on the existence of a folder holding a large share of the archive. That means that the files were saved in one main folder out of the root directory – a strategy that we may call One-folder Filing.

Cluster 3 (n=262) has minimum mean values for two variables: *Files per folder* and *Largest folder*, i.e., students in it have small folders on average (6.1), and their largest folder is also relatively small (14.52). This suggests that the cluster represents a Small-folders Filing organization strategy.

In Cluster 4 (n=66), the means of the same two variables as in Cluster 3 take their extreme values: Both *files per folder* (23.1) and *largest folder* (71.62) are maximal. By examining the mean value of *pile rate* (0.13), it might be concluded that about 87% of their files are filed, with one folder containing about half of their files (0.48). Therefore, this cluster, which we call Big-folder Filing, describes a mixture of filing and piling.

According to this analysis of the clusters, we present the following classification of personal information space organization strategies: Piling, One-folder Filing, Small-folders Filing, and Big-folder Filing. Table 3 shows the distribution of the four types in the research population.

Table 3. Personal Information Space Organization Strategies distribution

Personal information space organization strategies (cluster number)	N	% of students
Piling (1)	141	27
One Folder Filing (2)	49	9
Small Folders Filing (3)	262	51
Big Folder Filing (4)	66	13

For examining the association between the archive size and its organization strategy, mean values for archive size (total number of files) were compared between the clusters. Using Univariate ANOVA test, it was shown that the means are significantly different. As may be seen from Table 4, two strategies (Piling, One-folder Filing) have a small archive size on average (24.4 and 22.31, respectively), while the largest mean value for archive size (284.73) was found in the Big-folder Filing cluster. This indicates that larger archives are associated with strategies of filing into more than one directory.

Table 4. Archive size in the different clusters

Personal information space organization strategies (cluster number)	N	Archive size statistics			
		Minimum	**Median**	**Maximum**	**Mean *(SD)***
Piling (1)	141	10.00	17	155.00	24.40 (20.30)
One-folder Filing (2)	49	10.00	18	52.00	22.31 (10.85)
Small-folders Filing (3)	262	10.00	38	967.00	70.18 (101.04)
Big-folder Filing (4)	66	42.00	144	2170.00	284.73 (369.04)

5 Discussion

The main purpose of this study was to empirically identify different types of personal information organization strategies, which are part of Personal Information Management (PIM), and to do so for a large population, using data mining methodologies. PIM is not only a coherent and integral part of the learning process in the digital era - it is a process through which students learn. Therefore, researching PIM in the context of learning is very important for having a broader understanding of the learning process. Applying data mining techniques for PIM research brings new and fascinating opportunities to this field, as was demonstrated in this study.

Focusing on users' management of online personal archives, we were able to empirically identify four types of archiving strategies: a) Piling – most of the files are in the root directory; b) One-folder Filing – most of the files are located in one folder, under the root directory; c) Small Folders Filing – items are being divided into many relatively small folders (about 6 files per folder on average); d) Big-folder Filing – items are being divided into folders (about 23 files per folder on average) with about a half of them located in one big folder. These four types refine the classical Filing/Piling binary classification [14]. As the results suggest, students who tend to be Big-folder Filers, manage the largest archives and have relatively many files per folder on average. In order to construct a hierarchy of large coherent folders of different items related to a certain context (represented by each folder's name), students are required to a meaningful integration and generalization processes regarding the subject matter.

Our analysis showed that more than half of the participating students were categorized as Small-folders Filers. As this strategy is characterized by the use of small folders, this might imply that there are relatively many near-empty folders. Empty folders might indicate on a pre-building strategy, as was previously observed in the context of students' PIM [8]. Having many empty folders might increase PIM complexity, as well as having big folders. The strategy of Big-folder Filing was found in this study as associated with large archives, supporting previous findings [11].

In the context of learning, increasing PIM complexity is of special interest as PIM activities require cognitive skills. Bloom's cognitive taxonomy for learning objectives [5] enables us to analyze the three main PIM activities – i.e., naming, sorting, and categorization – in the light of three different levels of the taxonomy's cognitive skills: knowledge, analysis, and synthesis, accordingly. Regarding the four personal organization strategies found in this study, we might suggest different levels of reflected activities. In Piling strategy, the students neither name, sort nor categorize any information items. In One Folder Filing strategy, the students name only few folders and don't sort or categorize at all. In Small Folders Filing, the students name folder and sort information items into them, however they only do little categorization (since they join only few items into each folder). Only in Big-folder Filing strategy, students name, sort and categorize many items into folders. As the results suggest, managing bigger archives requires a wider range of cognitive skills. Replicating the process described in this article over several points in time might enlighten issues regarding changes over time of PIM strategies and their related cognitive activities.

PIM is subjective and idiosyncratic, and because PIM research mostly uses qualitative data collection from relatively small populations, it might seem that there are as many PIM variations as there are researched users [12]. However, using a large research population and data mining techniques, unexpected patterns might arise, suggesting similarities between groups of users, as was shown in this study. To promote the creation of large datasets, Chernov et al. [9] have suggested building a repository of PIM activity log files; this then would serve the PIM research community. Since it is likely that there will be problems obtaining participants' consent to trace their PIM activity over time, it might be easier to collect structural data reflecting accumulating activity.

References

1. Abrams, D., Baecker, R., and Chignell, M. (1998). Information archiving with bookmarks: personal Web space construction and organization. *Proceedings of the SIGCHI conference on Human factors in computing systems*. Los Angeles, California, United States: ACM Press/Addison-Wesley Publishing Co.

2. Barreau, D. (2008). From Novice to Expert: Personal Information Management Behaviors in Learning Contexts. *CHI 2008 Workshop*. Florence, Italy.

3. Bergman, O., Beyth-Marom, R., and Nachmias, R. (2008). The user-subjective approach to personal information management systems design: Evidence and implementations. *The American Society for Information Science and Technology*, 59(2), 235-246.

4. Bergman, O., R. Beyth-Marom, and R.Nachmias. (2003). The User Subjective Approach to Personal Information Management Systems. *Journal of the American Society for Information Science*, 54(9), 872-878.

5. Bloom, B.S. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*, McKay, New York.

6. Boardman, R. and Sasse, M.A. (2004). "Stuff goes into the computer and doesn't come out": a cross-tool study of personal information management. *Proceedings of the SIGCHI conference on Human factors in computing systems*. Vienna, Austria: ACM.

7. Brooks, J.G. and Brooks, M.G. (1993). *In Search of Understanding: The Case for Constructivist Classrooms*, Association for Supervision and Curriculum Development, Alexandria, VA.

8. Chang, S.-J. and Ko, M.-H. (2008). Behaviors of PIM in Context of Thesis and Dissertation Research. *CHI 2008 workshop*. Florence Italy.

9. Chernov, S., Demartini, G., Herder, E., Kopycki, M., and Nejdl., W. (2008). Evaluating Personal Information Management Using an Activity Logs Enriched Desktop Dataset *CHI 2008 Workshop*. Florence, Italy

10. Chirita, P.A., Gaugaz, J., S.Costache, and W.Nejdl. (2006). Desktop context detection using implicit feedback. *SIGIR 2006 Workshop on Personal Information Management*. Seattle WA, USA.

11. Fisher, D., Brush, A.J., Gleave, E., and Smith, M.A. (2006). Revisiting Whittaker & Sidner's "email overload" ten years later. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*. Banff, Alberta, Canada: ACM.

12. Kelly, D. (2006). Evaluating personal information management behaviors and tools. *Communications of the ACM*, 49(1), 84-86.

13. Lansdale, M.W. (1988). The psychology of personal information management. *Applied Ergonomics*, 19(1), 55-66.

14. Malone, T.W. (1982). How do people organize their desks? (Extended Abstract): Implications for the design of office information systems. *Proceedings of the SIGOA conference on Office information systems*. Philadelphia, Pennsylvannia, United States: ACM.

15. Manco, G., Masciari, E., and Tagarelli, A. (2008). Mining categories for emails via clustering and pattern discovery. *Journal of Intelligent Information Systems*, 30(2), 153-181.

16. Mioduser, D., Nachmias, R., and Forkosh-Baruch, A. (2009). New Literacies for the Knowledge Society, in *International Handbook of Information Technology in Primary and Secondary Education*, J.M. Voogt and G.A. Knezek, Editors. Springer. p. 23-41.

17. Nachmias, R. and Ram, J. (2009). Insights from a Decade of Campus-wide Implementation of Blended Learning in Tel Aviv University. *The International Review of Research in Open and Distance Learning*, 10(2).

18. Teevan, J., Dumais, S.T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. Salvador, Brazil: ACM.

19. Whittaker, S. and Hirschberg, J. (2001). The character, value, and management of personal paper archives. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(2), 150-170.

20. Whittaker, S. and Sidner, C. (1996). Email overload: exploring personal information management of email. *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*. Vancouver, British Columbia, Canada: ACM.