| | |
|---|---|
| Title | A quantitative comparison on file folder structures of two groups of information workers |
| Author(s) | Zhang, H; Hu, X |
| Citation | The 14th IEEE/ACM Joint Conference on Digital Libraries (JCDL 2014), London, UK., 8-12 September 2014. In ACM / IEEE Joint Conference on Digital Libraries Proceedings, 2014, p. 1-2 |
| Issued Date | 2014 |
| URL | http://hdl.handle.net/10722/213514 |
| Rights | ACM / IEEE Joint Conference on Digital Libraries Proceedings. Copyright © IEEE. |

# A Quantitative Comparison on File Folder Structures of Two Groups of Information Workers

Hong Zhang
School of Library and Information Science,
University of Kentucky
Lexington, Kentucky, U.S.A.
zhanghong920@gmail.com

Xiao Hu
Faculty of Education
University of Hong Kong
Pokfulam Road, Hong Kong S.A.R.
xiaoxhu@hku.hk

## ABSTRACT
This study compares file folder structures on personal computers of two groups of information workers, administrative staff and PhD students. A set of quantitative measures are calculated which disclose the differences and similarities between folder structures of the two user groups. The results shows that the group conducting more administrative activities has broader and shallower folders than the PhD group who performs more research activities, and the folders of the PhD group are more populated over deeper levels of the trees than those of the administrative group. The study improves our understanding of the various quantitative measures in investigating personal computer folder structures, and furthermore contributes to our knowledge of the information organization structure in personal information systems.

## Categories and Subject Descriptors
H.3.2 [**Information Storage and Retrieval**]: Information Storage – *file organization;* H.1.2 [**Models and Principles**]: User/Machine Systems - *human factors*

## General Terms
Measurement, Management, Human Factors

## Keywords
File structures, personal information organization, information workers, user groups, quantitative measures

## 1. INTRODUCTION
File folders on computers are places where information workers spend much effort and time creating, organizing, and accessing information for daily work and study. However, our knowledge of this familiar phenomenon is still limited [2]. Existing studies on file organization in personal computer and information systems have drawn inclusive conclusions: some of them observed broad and shallow tree structures while others observed deep tree structures among files in computer of information workers [2].

This study compares the file folder structures of two groups of information workers using a set of quantitative measures, aiming to find out if the different findings in previous studies can be at least partially attributed to the different user groups who conduct different information tasks. Results of the study will help deepen our understanding of how information workers organize information in file folders on personal computers, based on which implications for designing future systems could be proposed to facilitate information workers conducting their tasks.

## 2. DATA COLLECTION
The file folder structures of personal computers of 12 participants were examined in this study. The participants include six Ph.D. students and six administrative staff in an academic environment. The home folder (the topmost directory of a directory tree where a user puts most of his/her documents and folders) of each participant's computer was scanned as well as two to four selected top-level folders (subdirectories of the home folder). The top-level folders were purposefully selected such that they included directories for a current working project, a completed or archived project, and miscellaneous files. All the top-level folders of system and application software were excluded because they largely were not managed by the participants.

All of the six administrative participants (Adm) and two of the Ph.D students (PhD) were using Windows XP. Three PhDs were using Mac Operating System, and the other was using the Unix operating system. The length of time that the Adms had been in the institution ranged from three months to 29 years, while the PhDs had been in their programs for one to six years.

## 3. PRELIMINARY RESULTS
We examined the depth, breadth, and shape of folder structures, as well as file distributions in folders.

### 3.1 Tree Depth
The depth of each leaf folder (a folder that has no subfolder) is calculated for each participant's scanned folders. The mean and standard deviation values of each group's depths are listed in Table 1.

**Table 1: Two Groups' Average and Maximum Depths**

| | Ave. Depth | | Max. Depth | |
|---|---|---|---|---|
| | Mean | Stdev | Mean | Stdev |
| Adm | 2.50 | 0.44 | 4.00 | 1.10 |
| PhD | 5.12 | 2.18 | 9.17 | 4.71 |

The PhD group generally had deeper folder structures than the Adm group as reflected in the means of average and maximum depth. An unequal variance *t*-test on the two groups' maximum depth confirms that they are significantly different ($p = 0.043$). Similar result exists when comparing the average depths of the two groups ($p = 0.032$). Although the data set is small, the Q-Q plots of the maximum and average depths of the two groups show that they are approximately normally distributed. Overall, the result shows that the Adm participants generally have shallower folders than the PhD participants.

978-1-4799-5569-5/14/$31.00 ©2014 IEEE

## 3.2 Tree Breadth

A folder structure's breadth can be measured by the average number of subfolders per folder in a hierarchical folder structure. Overall, the PhD group had many more (sub)folders than the Adm group: PhDs had a total of 3,127 subfolders, while the Adm group only had 691 subfolders in total. The average number of subfolders per folder was 8.13 (stdev: 5.73) for the six Adm participants, 3.40 (stdev: 0.39) for the PhDs, which means the Adm group generally had broader tree structures than the PhD group. The large value of standard deviation for the Adm group also indicates that the Adm group has larger inter-personal variation than the PhD group. Although the two groups had very different average number of subfolders per folder, all the participants in the two groups had similar median values: mostly one or two except for one Adm participant having a median of four. Similar situation apples to the mode values: most of them were one and some were two, meaning that it is more common for the folders to have only one or two subfolders.

## 3.3 Tree Shape

The shape of a tree can be roughly depicted by its breadth and depth at the same time. If a tree's breadth (as measured by average number of subfolders per folder) is larger than its depth, then its shape can be summarized as relatively broad and shallow. Similarly, if a tree's breadth is smaller than its depth, then its shape can be summarized as relatively narrow and deep. We found that four of the six participants in the Adm group had larger average tree breadth than average tree depth while all participants in the PhD group had smaller average tree breadth than average tree depth. According to the Fisher's Exact Test that was specifically designed to test on small sample size [1], the difference was significant at the 90 percent confidence level ($p$ = 0.06). This result suggests that the folder trees in the Adm group tended to be relatively broad and shallow while the trees in the PhD group tended to be relatively narrow and deep.

## 3.4 File Distribution

Distributions of individual files can give further information on how information is organized by these information workers. Figure 1 illustrates the histograms of the numbers of files per folder of the two groups. As it shows, the PhD group had much more files than the Adm group, with 35,721 for the PhD group and 6,146 for the Adms. In addition, the frequency distribution of the number of files per folder did not fall in normal distribution, but seemed to follow Zipf's law [4] when number of files per folder was larger than zero. A maximum-likelihood estimation (MLE) was conducted to calculate the exponents and test the fitness of file distribution to Zipf's law distribution. The results show that both distributions fit Zips' law. The exponent values are 1.30 and 1.27 for Adm and PhD group respectively, which indicates that the frequencies of the two groups decrease at similar speed.

The two groups had the same median value, four, and the same mode, one, for the number of files per folder. That is, both groups' most popular number of files per folder was one, as can be seen from Figure 1. The two groups had similar mean values of number of files per folder: with 10.87 for the Adm group and 8.92 for the PhD group, but as the distributions are so skewed, median and mode are better measures than mean.

Overall, the results show that most of the folders examined in this study included small number of files, although there were exceptional folders with large numbers of files in both groups (see Figure 1). It is particularly noteworthy that there were a large number of folders with single files, as intuitively it is less efficient to include only one file in a folder.

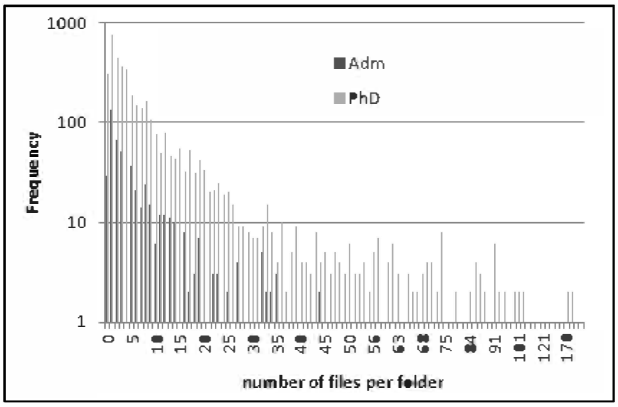


**Figure 1: Frequency distribution of number of files**

# 4. CONCLUSIONS AND FUTURE WORK

This study compares the computer file folder structures of two groups of information workers, administrative staff and PhD students. The folder tree depth, breadth, shape and file distribution were examined and measured quantitatively. There are some common patterns in the two groups, such as both have many folders with one or two subfolders and/or one single file, but what is more evident is the differences between the file folder structures of the two groups. The PhD group had significantly deeper folder structures, more (sub)folders and files than the Adm group; and the Adm group tended to have broader and shallower folder structures while the PhD group tended to have narrower and deeper ones. These different characteristics of folder structures may reflect the scale of administrative activities and the depth of research activities done by the two groups of information workers respectively [3]. Our findings suggest that the natures of information activities routinely conducted by the users should be taken into account in investigating personal digital document organization. The study improves our understanding of the various quantitative measures in investigating computer folder structures, and furthermore contributes to our knowledge of the information organization structure in information workers' information spaces. Due to the limited data size, this study focuses on exploratory analyses and does not intend for generalization of the findings. Larger scale comparative studies are needed to verify the findings and extend the exploration.

# 5. REFERENCES


[1] A. Agresti, "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, vol.7, no. 1, pp. 131–153, 1992. doi:10.1214/ss/1177011454

[2] O. Bergman et al., "The effect of folder structure on personal file navigation," *Journal of the American Society for Information Science and Technology*, vol 61, no. 12. pp. 2426-2441, 2010.

[3] O. Bondarenko and R. Janssen, "Documents at hand: Learning from paper to improve digital technologies," In *Proc, of the SIGCHI Conference on Human Factors in Computing Systems*, Portland, Oregon, 2005, pp. 121-130.

[4] G. K. Zipf, "*Human Behavior and the Principle of Least Effort*," Addison-Wesley Press, 1949.