The Utility of Controlled Vocabularies within Bookmark Management Tasks

Siu-Tsen Shen
Department of Multimedia Design
National Formosa University
Hu-Wei 63208, Taiwan, R.O.C.
stshen@nfu.edu.tw

Stephen D. Prior
School of Engineering and Information Sciences
Middlesex University
London N14 4YZ, UK
s.prior@mdx.ac.uk

***Abstract*—This research investigates the utility of adopting a controlled vocabulary approach to bookmark management. An initial user survey conducted for this research has shown that just over half the population use bookmarks to save important websites and that 75% of these people use up to three sub-levels only. The bookmark facility within all current web browsers is therefore underutilized and the argument that users need and want greater freedom and flexibility to create their own unique file structure is disputed. We conclude that users need a simple, logical and contextual system of bookmark management which complements their daily lives.**

Keywords - controlled vocabulary; bookmark management; web browser; information search; information retrieval.

I. INTRODUCTION

Personal file management has become significantly more important as the daily amount of digital data we view and store on our computers, smart mobile devices and web browsers increases. The familiar concept of the hierarchical file system allows us to group our important information in an organized tree structure, i.e., the use of folders, referred to as directories within folders. Folders (directories) normally include other possible sets of files or sub-folders (subdirectories). This is useful for desktop organization and bookmark management in relation to relevant topics and information.

Nevertheless, the usefulness and efficiency of the hierarchical file system has been debated over the last few decades [1-3]. Bloehdorn and Volkel (2006) stated that current file systems are problematic (with their single location ascribe) to browse to maximum specificity (retrieval needs the exact directory), miss-orthogonality (all orthogonal dimensions being forced into one access path), path order dependence (contrast to the directories seen as independent attributes), have no query refinement (no list of relevant directories to search), and have no navigational aid (no indication of the content of subfolders) [4].

In Section II, we will discuss the related work with particular references to controlled vocabularies. Following this section will be data on a preliminary user study. A discussion of the initial findings, analysis, and conclusions follows on from this.

II. RELATED WORK

Tags, also known as metadata, have been widely used within Web 2.0, and serve as labels to be easily-identified in information retrieval tasks. Successful examples include del.icio.us [5] , the social bookmarking website, Flickr [6] for image collection and YouTube for video collection. Compared to the inflexible, one-way system of the hierarchical file system, a tagging system gives the user a great deal of freedom to mark their wanted items, which could be multiply-tagged. Especially popular, is the social and collaborative sharing of information, also known as Folksonomy [7]. Websites with a large collection of text listed tags in alphabetical order (normally) shown as tag clouds are helpful for browsing and searching by their various fonts, size and colour. A successful social tagging website like '43 Things' is an example. Nevertheless, users may create thousands of tagged items and end up spending more effort on sorting and finding the tags that are needed.

A. Controlled Vocabularies Applied to IT

In contrast to the concept of Folksonomy, which is an informal and liberal collaborative tagging system, a controlled vocabulary is a restricted system of textual tags normally used for large datasets. Examples include the Library of Congress Subject Headings (LCSH) [8], the European Patent database [9] and the Yellow Page phonebook. For example, the International Patent Classification (IPC), established by the Strasbourg Agreement 1971 [10], provides for a hierarchical system of language independent symbols for the classification of patents and utility models, according to the different areas of technology to which they pertain. In total, this consists of eight main subject headings under which every patent application has to be categorized [11]. These are listed below:

Section A — Human Necessities
Section B — Performing Operations, etc
Section C — Chemistry; Metallurgy
Section D — Textiles; Paper
Section E — Fixed Constructions
Section F — Mechanical Engineering, etc

 ISBN: 978-1-61208-177-9

Section G — Physics
Section H — Electricity

The purpose of controlled vocabularies is to classify the terms such as words or phrases defined by experts or authorities in order to make retrieval performance more efficient. However, it requires a certain level of preciseness in the interpretation of the terms. It is common that users experience a familiar situation, whereby they conduct an online search by typing in keywords which may have a more general and broad meaning, and might therefore come up with a long list of irrelevant or unwanted information. A successful performance when using a search engine normally relies on the individual user's capability as to whether he or she could select the appropriate keywords or not. Controlled vocabularies are thus generally applied to thesauri, taxonomies and ontology [12].

Several researchers concluded that people's capability of categorizing information is cognitively difficult [13-16]. It has been stated that human's ability to categorize is hard to identify and is definitely not in a strict hierarchical structure, but shall be assumed to be more fluid and flexible [17, 18]. File systems, rigorous hierarchical mechanisms, such as 'My Documents' and 'My Favorites/Bookmarks' have been shown to have the usability problems of usefulness and appropriateness, including filing management, document organization, and document retrieval [2, 19-22].

Several attempts to use different approaches to replace the standardized hierarchical system have provided solid results in previous research. Barreau and Nardi (1995) found that people prefer to use location-based search and visual grouping, rather than complex data structures [1]. Gifford, Jouvelot, Sheldon, and O'Toole (1991) employed an associative attribute-based approach to access files and directories by semantic indexing, and proved to be more effective than the hierarchical structure [23]. Dourish et al. (2000) implemented a property-based approach to amend the traditional file system's problem with a uniform framework [2].

III. PRELIMINARY USER STUDY

In order to gain an understanding of user behavior, with regards to the use of current file systems and browsing patterns, a preliminary study was conducted online via an academic-based social blog in August 2011. The recruitment of participants aimed for experienced computer users, in that they could provide more insightful views according to their intensive usage on task performances such as searching, browsing and organizing information. There were a total of 60 participants, consisting of 37 females (62%) and 23 males (38%). Their age was between 18 to 25 years old, with 50% of them being aged 20 years old (see Figure 1).

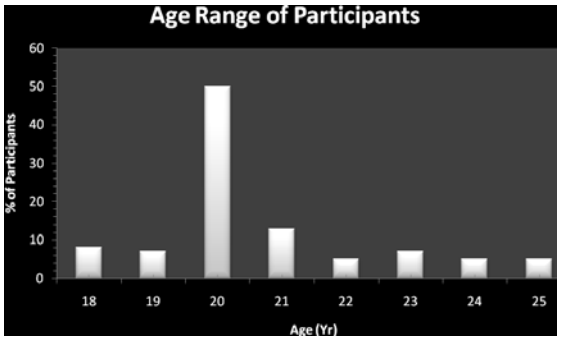

Figure 1. Age range of participants.

The group consisted of 88% of the participants who were University students, 10% from graduate schools, and 2% with a senior high school degree. In terms of computer experience, 33% of the participants had 10-12 years of experience, 22% with 8-10 years, 17% with 14-16 years of experience, and 12% with 12-14 years of experience (see Figure2).

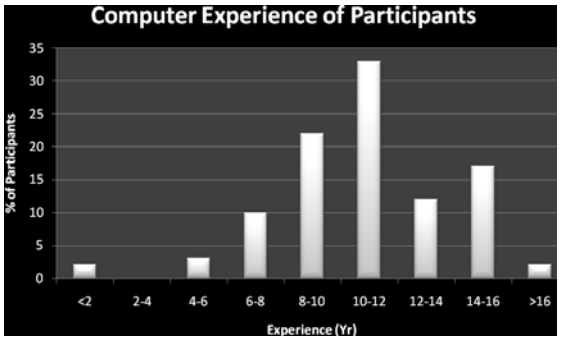

Figure 2. Computer experience of participants.

In terms of their Internet usage, 28% of the participants spent 6-8 hours, a further 23% of the participants spent 4-6 hours and 8-10 hours, and 12% participants who spent more than 12 hours a day online (see Figure 3).

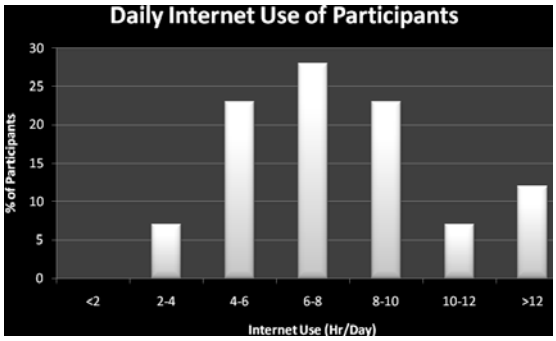

Figure 3. Daily internet use of participants.

The majority of the participants (93%) did search for information from their Bookmark folders, whilst only 7% of the participants did not. There were 83% of Bookmark

folders who had less than 50, 13% who had 50-100 folders, and 3% who had 100-150 folders (Figure 4).

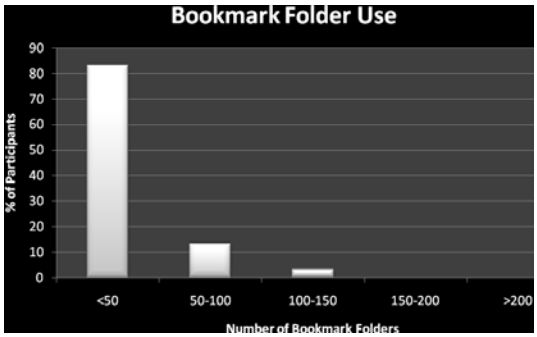

Figure 4. Bookmark folder use.

With regards to the type of computers used, there were 83% PC users, 7% Mac users, and 10% users who used both PC and MAC. 73% of participants use Google Chrome as their default browser, 15% with IE, and 12% with Firefox. Most of the participants had installed the latest versions of their browsers: 60% with Google Chrome 13, 12% with Firefox 5, 10% with IE 8 and 9. Nevertheless, there were 18% of the participants who did not know the version of their browser.

In terms of the use of Bookmark sub-folders, it was found that 53% of the participants did use them, and 47% who did not use them at all.

With regards to the maintenance of folder levels, 20% of the participants used 3 levels, 13% used 2 levels, 10% used 4 levels, 7% used only one level, and 3% used 5 levels (see Figure 5).

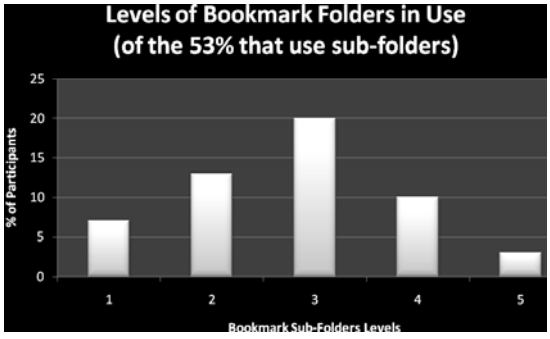

Figure 5. Levels of bookmark sub-folders in use.

IV. DISCUSSION OF INITIAL FINDINGS

From the outset, it was a desirable goal to have even numbers of male and female participants. However, the study was conducted online and therefore the researchers could not foresee or manipulate the numbers from the participants, since it was freely available 24-7. The vast majority of the participants were university students. Although 10% of the participants were from graduate schools, they were studying related digital media subjects. It is important to note that they are dependent on computers, in that computers are the essential tools for their future career prospects. It is surprising to note that even in a group of young people such as these; one third of the participants had 10-12 years of computer experience which corresponds to the prevalence of Internet boom back in late nineties. It was found that about 75% of the participants spent between 4-10 hours/day using the Internet, which indicates the strong need for information and communication tools. It is not surprising that the PC is still dominating the market. However, with the latest popularity of multi touch devices like the iPad, it is still hard to tell when these will filter down to common use. With the recent trend for using smart phones, it was not surprising that 30% of our survey used one; there were 20% of the participants who used the HTC android platform, compared to 7% for iPhone users and 3% for Blackberry users.

A surprisingly high number of participants (70%), used Google Chrome as their default browser (even the Mac users); this indicates the advantage for web browser which are integrated within powerful search engines. Most of the participants kept up-to-date with the latest versions of their browsers, which also reflect their fast adaptation to the newest technology available. The frequency of using file systems under Bookmarks to search for information and files proved that the vast majority of the participants were indeed making efforts to organize their data. However, a majority of the participants managed their files into less than 50 folders.

It is interesting to note that 47% of the participants did not use sub-folders to organize their Bookmark files into more refined levels even though they created main folders to store information. Furthermore, 75% of the participants who used sub-folders managed to organize their personal information using 3 levels or less. This suggests that the current file systems that give almost unlimited creation of folders and sub-folders are perhaps unnecessary, because they are not well utilized by users who may either not want to make much effort on sorting their database, or may be aware that they might not be able to retrieve their desired information efficiently. Even high tech users find it difficult to manage the overwhelming data tsunami which hits us all on a daily basis; resorting to either not bookmarking or bookmarking without using folders and sub-folders.

A. Limitations of the study

This study can be criticized for only having a small number of homogenous participants and an uneven balance of genders represented. However, it is important to note that all the participants were studying multimedia design relevant subjects which require professional skills within several advanced software, as the aim of this study was focused on the experienced computer user.

Furthermore, this study was conducted via an online process and offered no cash in return for participation; therefore it was hard to manipulate the exact even numbers of both males and females compared to a lab-controlled setting environment. It is interesting to note that the majority

 ISBN: 978-1-61208-177-9

of participants were aware of the traditional hierarchical file system and did use them. However, nearly half of the participants did not create any sub-folders. Either they did not have a habit of organizing information or they thought that the traditional file system structure might not be helpful.

Traditional file system structures (developed by IT specialists) have existed for several decades and provide unlimited freedom for creating files and folders for users to organize their personal information. Based on our findings, three quarter of the participants used files and directories only up to 3 levels. This suggests that the current bookmarking system with 255 levels may be overly complex levels. This could be made simpler and more intuitive in terms of categorization via controlled vocabularies.

The purpose of getting online is mainly for searching, socializing and communication. Users may not appreciate the hierarchical filing system as others do. Moreover, it requires a lot of time and effort in organization information and does not guarantee users could successfully retrieve their required data when needed. Therefore, if we take this notion further, it would be a better idea to find an adequate approach to make the existing file systems into a simplified and deductive knowledge repository.

V. CONCLUSION

Based on our literature review, together with the results of the preliminary study listed above, several usability problems have been identified. The elements in need of improvement are: categorization, optimum levels of sub-folders, ambiguity of the use of vocabulary and contextual user mental models.

It is anticipated that if we could make the filing system less complicated and less strict, it would encourage users to be more willing to organize their information under such architecture. It is not our intention to replace the existing file system, but rather to offer a fresh perspective in visualization and user interface design.

The use of controlled vocabularies to assist in structured information storage and retrieval tasks looks to be promising, yet due to the natural ambiguity of descriptions of any specific term or object, it may appear not sufficient enough to achieve adequate understanding of precise meaning.

From an analysis of people's daily lives, further work is proposed to use a controlled vocabulary, which is divided into four primary facets, i.e. Work, Home, Travel, and Health. These could each further be broken down into a secondary level of say ten sub-categories. The use of these primary and secondary facets could help users reduce confusion and simplify the procedure when they organize their web information using bookmarks.

REFERENCES

1. Barreau, D. and B. Nardi (1995) *Finding and Remembering File Organization from the Desktop* SIGCHI Bulletin. **27**(3): pp. 39-43.
2. Dourish, P., W. Edward, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. Terry, and J. Thornton, *Extending Document Management Systems with User-Specific Active Properties.* ACM Transactions on Information Systems, 2000. **18**(2): pp. 140-170.
3. Sullivan, K. *The Windows 95 User Interface: A Case Study in Usability Engineering*. *Proceedings SIGCHI*. 1996: ACM: pp. 473-480.
4. Bloehdorn, S. and M. Völkel. *TagFS - Tag Semantics for Hierarchical File Systems*. *WWW2006*. 2006. Edinburgh, UK.
5. AVOS. *Delicious.* 2011 [retrieved: December, 2011]; Available from: www.del.icio.us.
6. Yahoo. *Flickr*. 2011 [retrieved: December, 2011]; Available from: www.flickr.com.
7. Lamere, P., *Social Tagging and Music Information Retrieval.* Journal of New Music Research 2008. **37**(2): pp. 101-114.
8. Library of Congress, *Library of Congress Subject Headings*. 2011 [retrieved: December, 2011]; Available from: http://www.loc.gov/aba/.
9. European Patent Office, *European Patent Database*. 2011.
10. WIPO, *Strasbourg Agreement Concerning the International Patent Classification*, WIPO, Editor. 1971, World Intellectual Property Organization.
11. WIPO, *International Patent Classification (IPC)*, WIPO, Editor. 2011, World Intellectual Property Organization.
12. NISO, *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. 2010, Bethesda, Maryland, U.S.A.: National Information Standards Organization Press.
13. Malone, T. *How do people organize their desks? Implications for the design of office information systems*. *ACM Transactions on Information Systems (TOIS)*. ACM Press, 1983. **1**(1): pp. 99-112.
14. Lansdale, M., *The psychology of personal information management.* Applied Ergonomics, 1988. **19**(1): pp. 55-66.
15. Kidd, A. *The Marks are on the Knowledge Worker*. *Proceedings of CHI on Human Factors in Computing Systems*. 1994. Boston, USA: ACM: pp. 186-191.
16. Whittaker, S. and C. Sidner. *Email overload: exploring personal information management of email*. *Proceedings of CHI'96*. 1996. Vancouver, Canada: ACM: pp. 276-283.

 ISBN: 978-1-61208-177-9

17. Eysenck, M.W. and M.T. Keane, *Cognitive psychology: a student's handbook*. 5th Edition. 2005, Hove: Psychology.
18. Oren, E. *An Overview of Information Management and Knowledge Work Studies*. *Proceedings of the ISWC Workshop on the Semantic Desktop*. 2006.
19. Bowker, G.C. and S.L. Star, *Building Information Infrastructures for Social Worlds - The Role of Classifications and Standards.* Community Computing and Support Systems, 1998: pp. 231-248.
20. Hearst, M.A., *Clustering versus Faceted Categories for Information Exploration.* Communications of the ACM 2006. **49**(4): pp. 59-61.
21. Bondarenko, O. and R. Janssen. *Documents at hand: Learning from paper to improve digital technologies*. *Proceedings of CHI2005*. 2005. Portland, Oregon, USA: ACM: pp. 121-130.
22. Marsden, G. and D.E. Cairns. *Improving the Usability of the Hierarchical File System*. *Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology*. 2003. South African Institute for Computer Scientists and Information Technologists, Republic of South Africa: pp. 122-129.
23. Gifford, D., P. Jouvelot, M. Sheldon, and J. O'Toole. *Semantic File Systems*. *Proceedings of the 13th ACM Symposium on Operating Systems Principles*. 1991: ACM Press: **25**(5): pp. 16-25.

 ISBN: 978-1-61208-177-9