Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 228-235

User Interfaces for Supporting Multiple Categorization

Dennis Quan, Karun Bakshi, David Huynh & David R. Karger
MIT AI Laboratory/LCS, 200 Technology Square, Cambridge, MA 02139 USA
{dquan,karunb,dfhuynh,karger}@ai.mit.edu

Abstract: As the amount of information stored on and accessed by computer has increased over the past twenty years, the tools available for organizing and retrieving such information have become outdated. The folder paradigm has dominated existing user interfaces as the primary mechanism for organizing information for day-to-day use. This paradigm encourages many-to-one placement of documents into strictly hierarchical containers. In this paper we examine an alternative organization and navigation mechanism that promotes membership in multiple overlapping categories (as opposed to storage containment). In particular, we explore the user interface consequences of multiple categorization support being made conveniently available from within Web browsers. We have carried out user studies providing evidence that compared to the folder paradigm, multiple categorization not only improves organization and retrieval times but also matches more closely with the way users naturally think about organizing their information.

Keywords: information retrieval, user interface, organization, user models, browsing, classification, categorization, folders

1 Introduction

Advances in personal computing technology over the last two decades have produced powerful tools, such as word processors, e-mail clients, and web browsers, for creating, retrieving, and sharing information of many forms. In contrast, the tools users are given to organize this information have not progressed nearly as much. The hierarchical file system, implemented in UNIX-based systems in the 1970s, remains the dominant paradigm for filing and categorizing documents. Initially, hierarchical directory structures were sufficient for the needs of the user who did not have many files to manage. Users also had to contend with technological problems, such as limited disk capacity, which also served to artificially cap the number of documents a user worked with at once. As network computing was introduced and powerful computers with expansive storage capacity have become ubiquitous, users now have to contend with thousands of e-mails, documents, and Web pages, and the limitations of the hierarchical folder paradigm have become more noticeable.

1.1 Filing Documents

The system of hierarchical folders used on most operating systems today was designed in analogy to that used in filing cabinets for centuries, and as a result, it has inherited many of its physical counterpart's problems. One such problem is the inability to conveniently file documents in more than one category. Fundamentally, a filing cabinet serves a dual role as both a place to store paperwork and also a way to organize it. Moving towards the computer version, we find that this duality makes little sense.

Although hierarchical computer folders are an efficient means for storing documents, the hierarchical folder system presents challenges to users who attempt to use it to categorize documents. Does a document named "PlayStation 2 enters online arena" belong in the "video game" folder, the "Sony" folder, or the "Internet" folder? On a related note, take the example of browsing a collection of recipes stored on a user's hard drive. Figure 1 shows what an example folder organization scheme might look like. Note that it is difficult to place information into only one place within the hierarchy, when it could equally belong in multiple locations. For example, one can ob-

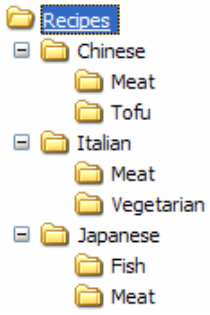

Figure 1: Example folder hierarchy

serve that a Chinese dish with both meat and tofu would have to go into two folders at once. Even worse is the situation that results if a dish's ethnicity cannot be identified, whereby a recipe may actually have no place in the hierarchy, even though we may know its ingredients!

Indeed, there is significant psychological evidence that single classification is the wrong approach. Lansdale reports that the act of classification into a single category is cognitively difficult (Lansdale, 1988). He cites research on office space organization done by Malone, who found that people who indiscriminately pile paper documents do so in order to skirt the problem of having to choose between several potentially overlapping categories (Malone, 1983). Whittaker et al. similarly found in their studies of users and their e-mail corpora that any message of nontrivial length has several axes along which the message may be filed (Whittaker et al, 1996).

Simply supporting files being in more than one folder at once is not sufficient. Commonly used modern operating systems such as Windows, MacOS, and Linux already provide mechanisms (called "shortcuts", "aliases", and "symbolic links" respectively) for placing objects in more than one folder. However, people make relatively little use of these features for simultaneously classifying documents into multiple categories (Dourish et al, 1999).

We postulate that this is because the user interface does not encourage simultaneous classification. How many programs can be found whose file save feature prompts the user for all the possible directories into which to place a file? Many users place their files into a single directory because they are not willing to expend the effort to classify files. Of the fraction that are willing, there is yet a smaller fraction who would be willing to save their files in one place, and then separately create shortcuts, aliases or symbolic links in the other directories.

1.2 Retrieving Documents

There are a number of usability problems to be overcome on the retrieval side that occur as a result of the inability to place documents into more than one category. Using a hierarchical system, the user is bound to a static organization scheme. Hence, he or she cannot view the information using a different scheme during retrieval than what was used when the information was initially organized.

For example, consider again Figure 1. If the user is interested in recipes that must have meat since he or she has meat that must be used before it goes bad, he or she will have to search multiple folders. Although the hierarchy of recipes based first on cuisine and then on ingredients made sense at the time of organization, it now serves to restrict the ways in which the information can be retrieved. In particular, the user must remember the *ordered* sequence of topics and subtopics that were used to organize the information when attempting to retrieve it, even though the topics of interest during retrieval might be different from those during organization.

Ultimately, the purpose of filing a document is to make it easier to retrieve later (Landsdale, 1988). However, if the user is not given proper tools with which he or she can place the document into the categories he or she will likely look for it in later, retrieval performance will ultimately suffer.

This is not to say that hierarchies are not useful for retrieval. Barreau et al. discuss in their paper that hierarchies were the primary means by which people retrieve documents on their machines (Barreau et al, 1995). Below we present a retrieval system that uses a hierarchical display but also addresses the issues that arise with single category classification and needing to remember the sequence of topics used to retrieve a given document.

1.3 Contribution

Lansdale alluded to the idea of "multiple categorization" to solve many of the issues described earlier (Lansdale, 1988). Indeed, our approach adopts a category-based organization and navigation scheme that allows information to be placed in multiple thematic "bins", or categories, simultaneously. Allowing multiple categories lets the user organize documents in a more intuitive, richer information space and supports our belief that information inherently has multiple, relevant categories that the user can readily (albeit subjectively) identify.

In this paper we propose that multiple categorization is a useful technique for organizing many commonly-used forms of information, such as documents, web pages, and e-mails. However, to take full advantage of this approach, we postulate that multiple categorization user interfaces must become *pervasive* throughout the system. In other words, multiple categorization must pervade the user experience at least to the extent to which the folder paradigm does today; it is not sufficient to have multiple categorization functionality being provided by an "add-on" or as an afterthought.

This notion of pervasive multiple categorization is consistent with previous research. Whittaker's study revealed that users did not regard creating categories in current mail clients (most of which are geared towards folder-based organization) as a

"lightweight activity" and were hence discouraged from creating them. In a related domain, Abrams et al. found similar problems with the usability of a folder-based system for organizing bookmarks (Abrams et al, 1998). They found that users need more scalable tools because even though their bookmark collections grow quickly, the effort required to create and maintain bookmarks remains a constant hindrance. One finding was that a user "cannot place it [a bookmark] in the prescribed folder easily at the mouse click." These problems were important motivations in our attempt to prevalently expose convenient means for creating categories and organizing documents into categories.

To examine the usefulness of multiple categorization, we performed a user study that allowed users to categorize documents using multiple categorization. This user study tested the effectiveness of both categorization and retrieval. Not only do our prototype interfaces foster true multiple categorization, but they are also integrated into a commonly used information client: Microsoft Internet Explorer. We feel that the breadth of the different kinds of information accessible from Internet Explorer, from Web pages to Word documents on the local file system, make Internet Explorer an ideal test environment.

This research is being conducted in association with the Haystack project (Huynh et al, 2002). The goal of the Haystack project is to develop a tool that allows users to easily manage their documents, e-mail messages, appointments, tasks, and other information. Haystack uses a semistructured data model to describe the connections between different documents in a user's corpus as well as the metadata concerning each document. Users are then able to browse and retrieve documents based on this metadata, such as importance, sender, author, or category, rather than just by name and location in a hierarchy. The user interfaces presented in this paper are stripped-down versions of the ones used by Haystack, created specifically to explore the specific aspects of organization presented here.

1.4 Related Work

A few commercial products implement multiple categorization functionality. A somewhat hidden feature of Microsoft Outlook, the Categories dialog box, which resembles the list of category checkboxes used in our prototype, can be shown for any object by selecting the Categories option from a context menu. However, folder-based organization plays a predominant role in Outlook's paradigm, and burying multiple categorization functionality in a context menu makes it inconvenient for users to access it frequently.

A related example is Lotus Agenda, a text-based personal information manager developed before the Web (Kaplan et al, 1990). The dominant paradigm in Agenda was multiple categorization; however, due to the technology available at the time, the categorization process was separated from the text mode applications with which it could interoperate.

Another product that incorporates multiple categorization functionality is Bibliographix (available at http://www.bibliographix.com/), a software package that facilitates the management of bibliographic ref-

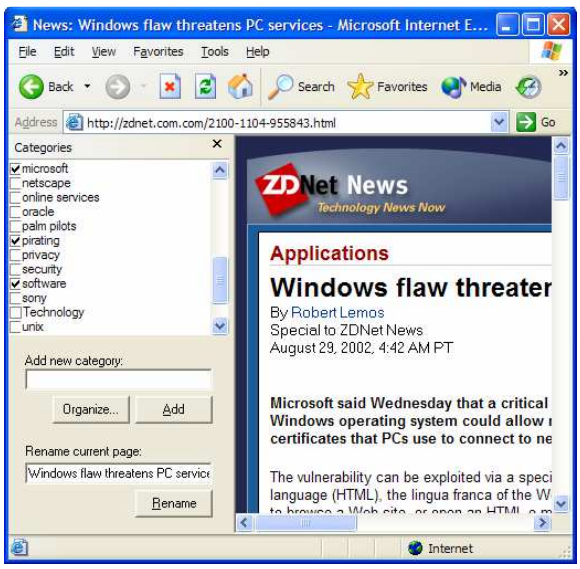

Figure 2: UI for organizing documents

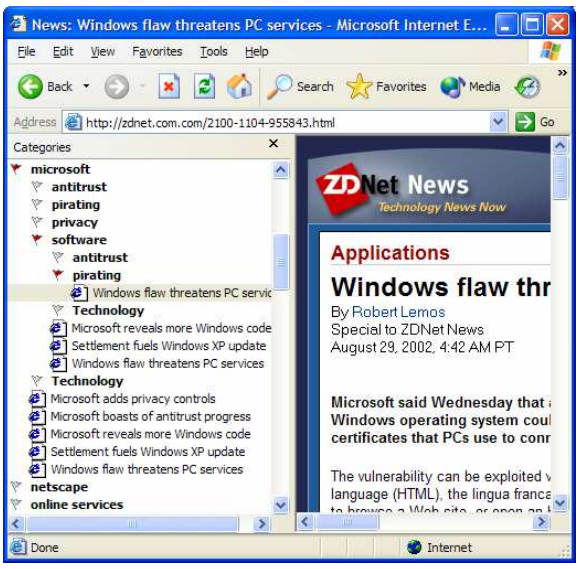

Figure 3: UI for retrieving documents

erences. The program's user interface greatly resembles the ones presented in this paper; however, we have tried to broaden the application of this technique to include other forms of information. The study presented in this paper can be seen as a validation of this style of multiple categorization interface for general use.

Our approach to retrieval can be thought of as a specific instance of the more general idea of metadata-aided retrieval, a technique Hearst notes has been employed successfully by sites such as Epicurious.com (Hearst, 2000), which gives its users the ability to browse their collection of recipes by iterative refinement. Haystack also gives the user tools with which to perform metadata-aided retrieval in its user interface. However, this paper focuses specifically on the aspect of categorization-aided retrieval in order to narrow the scope of our study.

Another possibility for improving the usability of categorization interfaces is to allow the computer to automate the process of categorization. Sophisticated clustering algorithms are available, and many have had success in applying them to this problem (Agrawal et al, 2000; Cutting et al, 1992). In this paper we will restrict our focus to the problem of usability in human-assigned categorization interfaces, noting that the incorporation of automated processes is an interesting and already somewhat explored area worthy of future research.

One issue heretofore not discussed is the idea of full text search playing a role in retrieval. The success of Internet search engines such as Google may suggest to some that search alone may solve most retrieval problems. We argue that this is not the case. Searching is of little use when the precise details of the target documents are not easily recalled. Lansdale suggests that retrieval consists of two sessions: recall-directed search, followed by recognition-based scanning (Lansdale, 1988). Folders and categorizations serve as guides for the user when visually scanning for the document being sought. In this paper, we restrict our attention to this latter phase, noting that full text search complements both folder hierarchies and categorization schemes and is a necessary part of any complete system.

2 Approach

Our categories pane, as depicted in Figure 2, consists of a list of checkboxes corresponding to categories with a series of widgets below for adding and removing categories and renaming the currently displayed page. To indicate membership in a category, users can check the box next to the category's name in the list box. Categories appear sorted in the list box alphabetically. Adding a category is accomplished by typing the name of the category into the text field below the "Add new category" label and clicking Add. Deletion and renaming of categories can be accomplished by using the Organize button.

The categories pane was designed to expose similar functionality to that of the Internet Explorer favorites pane to help users become familiar with it quickly. Internet Explorer's favorites pane allows users to create, delete, move and rename folders. Analogously, the categories pane supports creation, deletion and renaming of categories. The categories pane also exposes context menus with options for renaming and deleting categories for those used to this modality.

The user may then traverse his or her collection of information using the theme that is relevant to the task at hand. The browsing paradigm we present (Figure 3) gives the user a listing of his or her categories using a dynamically-generated hierarchy. When a category is clicked upon, it expands to display a list of the articles assigned to that category, as well as a list of the other categories to which those articles have been assigned. These other categories are in effect "subcategories" in that they represent subsets of the articles in the parent node. A category node in the tree corresponds to the conjunction of the categories associated with the node and its ancestors. The leaf nodes correspond to Web pages, which when clicked upon, trigger the Web browser to display the corresponding Web pages.

In other words, the user can continuously refine his or her search by recursively clicking on categories and seeing fewer and fewer documents corresponding to the intersection of the categories selected. Furthermore, the order in which these category tree nodes are expanded is not important. Hence, no matter which relevant category a user begins a search in, he or she can continue refining the relevant collection based on other categories to which items in the collection belong, and is likely to find the answer without encountering a "miss" or having to check in multiple locations. This approach can be seen as the presentation of a hierarchy based on a set of overlapping categories.

For example, clicking on the "microsoft" category reveals a list of five articles, and five other categories. The five articles are further assigned to these additional categories, namely "antitrust", "pirating", "privacy", "software", and "Technology". Expanding the "software" subnode shows which three of the original five articles fall under the "software" category. Descending down the tree further,

we find that the article "Windows flaw threatens PC services" falls under the "microsoft", "software", and "pirating" categories.

3 Experimental Method

We conducted a user study to compare users' preferences and performance between the two approaches: multiple categorization and hierarchical folder organization. In the first session of the study, users organized two separate corpora of news articles using the two different approaches. They then navigated those organizational schemes in the second session of the study after one week, in order to answer questions about several topics brought forth in the corpus. The study collected two types of metrics: quantitative performance measurements and qualitative feedback from the users describing their attitudes towards the two paradigms.

3.1 Participants

The users in this study were MIT computer science graduate students recruited by a general e-mail to the departmental mailing list that advertised the opportunity to participate in this study, ensuring an equal opportunity to participate for all people on the mailing list. The 21 participants (15 male, 6 female) were entered into a drawing for three gift certificates to an electronics store, one valued at US$50 and two valued at US$25.

Admittedly, computer science graduate students do not represent the general population, but we believe that their participation provides a number of benefits, in addition to convenience. First, computer science students are sufficiently proficient with computers as to not require extensive training to use our interface, avoiding the bias of any associated learning curve. Second, they deal with a lot of information on a constant basis, to the point that they would likely have experienced any problems with current information filing systems. Third, they represent a diverse populace in terms of cultural backgrounds and opinions of different computer operating environments. Finally, while computer science students are generally adept at organizing information, the style in which they organize their information is typically hierarchical, perhaps emphasizing any performance improvements shown for multiple categorization over hierarchical folders.

3.2 Test Environment

The test application consisted of a modified Internet Explorer window with three panes: an organization or navigation pane (depending on the session), a content pane for viewing a web page, and an instructions pane prompting the user to either categorize a page or to answer a question. Internet Explorer's favorites pane was chosen as a representative of the folder paradigm because it is widely used and condenses the key operations and aspects of a folder browser into a single area of the screen, making the experiment more controllable.

Our prototype does not represent our notion of the "perfect" user interface for categorization and retrieval. Instead, we designed the test environment to help us study specific aspects of users' preferences in classifying and browsing documents. A number of possible improvements, such as hierarchical categorization, were avoided in favor of keeping the number of variables being analyzed to a minimum. In a later section we detail some of these improvements, many of which actually came up as suggestions during the user surveys.

3.3 Organization Session

The main purpose of the organization session was for the user to organize two corpora of information using two different techniques: the folder hierarchy and multiple categories. Users began by viewing the directions for working with the test system and a demonstration of both techniques on a sample corpus in order to ensure a nominal level of familiarity.

Each of the two corpora consisted of a collection of 60 articles taken from ZDNet.com. The choice of ZDNet.com as a source was made in order to ensure that users would remain interested during the study session and would have sufficient understanding of the topics to make informed organization schemes. We chose this number of articles both to motivate users to organize the articles (a lesser number may have been manageable in a flat list) and to prevent users from becoming overly bored or frustrated with a larger number of articles.

The organization session was divided into two phases; the second phase followed immediately after the first phase. In each phase, the user was asked to organize one of the two corpora with a specific technique (either categories or folders). The order in which the users organized the two corpora and the assignment of which corpus was organized with folders or categories were varied per user during the organization session to avoid a systematic bias.

Articles were presented one at a time to the user, in sequence. The user was then asked to create an organization scheme from scratch using either hierarchical folders (via the Internet Explorer favorites pane) or multiple categories (via our prototype categories pane depicted in Figure 2). Users were being

timed but were advised to spend as much time as needed in organizing the documents. We allowed users to reshuffle folders as they saw potentially more relevant groupings with each new article. Similarly, users were permitted to go back and further classify past articles using multiple categories. Users were encouraged to mark an article with as many categories as they felt necessary.

Users were required to place each article in exactly one place in the hierarchy. This was mandated in order to simulate the common condition that documents fall into only one folder at once. The system monitored these collections to ensure compliance with these rules. However, users were permitted to not check any checkboxes associated with an article, as every article was implicitly part of the "all articles" category, which appeared at the root of the navigation pane in the navigation session.

3.4 Navigation Session

The navigation session required users to answer two sets of questions by using two different navigation techniques corresponding to the organization structures that they created during the first session. Users waited a week after their organization session before performing the navigation session in order to reduce the effects of memory on retrieval. The session commenced with the user reading the directions for working with the testing application and viewing a demonstration of the testing application's GUI.

Each navigation task involved the user responding to a set of about two dozen questions based on the corpus used in the corresponding organization task. Questions were presented one at a time to the user, in sequence, and identified one or more themes that were discussed in an article in the corpus. Special care was taken to ensure that keywords in the articles were not used in the wording of the questions to avoid allowing the users to easily create and use categories based on keywords. In order to answer a question, the user had to navigate the corpus using his or her navigation scheme. For the corpus organized with the favorites pane, the user was again given the favorites pane for navigation. The other corpus was displayed in the navigation pane depicted in Figure 3.

Articles themselves served as answers to the questions. When a user navigated to what he or she felt was the correct page, he or she was required to click the "I found it" button. If the answer was incorrect, i.e., the user navigated to the wrong article, the user was required to continue with the navigation and answer submission process until he or she identified the correct article or clicked the "I give up" button. (This latter button was provided in order to minimize frustration on the part of the user in the event that their organization scheme made it impossible for them to locate the article. The uses of this button were recorded.) The user was told that the session was being timed and to attempt to answer the questions in each set as quickly as possible.

4 Results

We present the results for the user study below taking into consideration both timing data (how long it took to categorize an article and how long it took to retrieve an article) and subjective survey responses. First, we discuss users' relative performance with the two paradigms for the organization session. We then similarly elaborate on the results of the navigation session. Herein we detail the evidence we found in favor of the specific advantages of multiple categorization.

4.1 Organization Session

The mean time to complete the organization session was approximately 1 hour and 29 minutes. We compared users' performance on the first phase with the second phase and found that on average, the second phase took only 70% as long. We attribute this to users hurrying through the second phase in order to minimize the total amount of time they spent on the study, and this reasoning was reflected in their comments. Furthermore, because the participants were inherently more familiar with folders than multiple categorization, we felt that their haste might have created an unfair bias against multiple categorization. As a result, we consider the results from the first phase to be more indicative of the relative performance of the two organization techniques.

On a one-tailed t-test basis, $t(19) = 1.1$, $p = 0.14$, users took considerably less time (19% reduction) organizing their corpus using multiple categorization (mean 2778.2, σ 833.7 seconds) compared to folders (mean 3441.2, σ 1693.9 seconds). (When considering both phases, users took on average 2754 seconds, σ 1434 for multiple categorization versus 2586 seconds, σ 1083 for folders, a less statistically significant result due to the factors discussed above.) We feel this result is substantial despite many factors working against multiple categorization, such as users' unfamiliarity with multiple categorization and the simplicity of the prototype user interface, which lacked the ability to show existing members of a category during organization. Further, many users told us that they found the corpora interesting and hence spent more time reading each article than was

necessary to organize it. This may have increased the variance in our results.

Subjective measures of user preference for multiple categorization over folders indicated a similar approval or preference. Of the 21 users who participated in the study, 10 users felt that conceiving and maintaining a folder hierarchy required a greater amount of cognitive effort as opposed to 6 users who felt that multiple categorization required more effort. The remaining 5 users felt that both techniques required an equal amount of effort. This result is encouraging, considering that users have been "practicing" folder-based organization for years.

Also, 8 of the 21 users felt that multiple categorization by itself more closely matched the way they think about information, and an additional 11 felt that some combination of the ideas embodied in the folder and multiple categorization paradigms captured how they modeled information. Only 1 user each felt that folders alone or neither technique matched how they modeled information. Finally, 9 participants felt that they would always prefer one style of categorization over the other, with 6 preferring multiple categorization to 3 preferring folders. Many of the users who found that a combination of the two techniques was closest indicated that they found folders useful for "file and forget" archiving of documents as well as for organization schemes with hard and fixed structures.

Earlier we mentioned that previous research pointed to a cognitive barrier to creating folders. One of the motivations for developing multiple categorization was to lower this barrier. We postulate that our progress can be quantified by measuring the total number of folders created versus the number of categories. On average across both corpora, we found during our study that users created 22 folders; in contrast, users created twice as many categories (45).

4.2 Navigation Session

In keeping with our reasoning from the organization session, we elected to use only data from the first phase of the navigation session. Also, because users were allowed to "give up" on questions they were not able to answer in a reasonable amount of time, we excluded these questions from our averages; we feel this is justified because users had a similar amount of success in answering questions using both techniques (12.4% of questions given up for multiple categorization versus 14% for folders). During this session, users took on average 36.9 seconds (σ 8.6) using categories compared to 44.7 (σ 12.3) seconds with folders, a 17% improvement, $t(19) = 1.66$, $p = 0.056$, one-tailed test. (For both phases, the results are 38.1 seconds, σ 11.0 for categories versus 44.3 seconds, σ 13.9 for folders.) The distribution of users' performances for folders and multiple categorization is plotted in Figure 4. As can be seen from the figure, the distribution of users' retrieval times using folders is shifted towards higher times than when they were using multiple categorization, thereby yielding the different means.

This result is bolstered by the subjective responses we collected during the survey. A significant number of users (15) felt that the ability to begin searching for an article in one of several possible categories was useful. Overall, users overwhelmingly preferred our categories-based navigation scheme (15) to either folders (3) or neither (3).

Other experimental results supported many of the hypotheses we used in constructing our method of multiple categorization. First, to validate our claim that information access is highly context-dependent, we inserted 3 pairs of questions into each of our question sets, where both questions in each of the 3 pairs shared the same article as their answer. Not surprisingly, about 2 times out of 3, users ended up taking different paths to the same answer when using multiple categorization.

In addition, users took advantage of the fact that by using multiple categorization, the user could narrow down through as many categories as needed before deciding to scan the resulting list for the desired article. This is in contrast to the folder system, where the user must navigate through every parent folder in order to reach the leaf folder with the desired article. We found that on average, when looking for a specific article, the number of categories users used to refine the list of articles was only about half of the total number of categories assigned to the desired article.

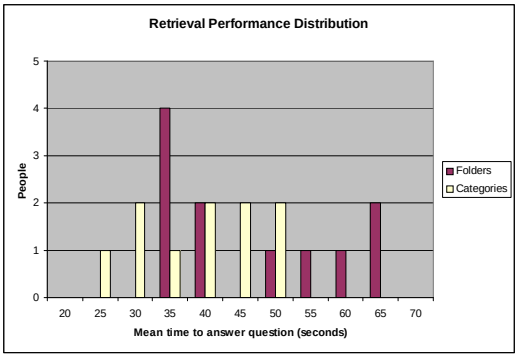

Figure 4: Relative retrieval performance of folders versus multiple categorization

5 Discussion

Our analysis has shown a noteworthy amount of user interest in the use of multiple categorization for organizing documents. However, we have so far presented a comparison of two extremes in order to expose the differences between the two paradigms. Of course the ideal system would bring to bear the best features of each paradigm. This is borne out by users' comments.

For example, users responded that in some instances multiple categorization gave them too much freedom and their categorization scheme degenerated into a keyword system. Instead of being crystallizations of specific concepts or themes present in the corpus, categories were created to represent the presence of a specific word in a document. Hence, synonymy became rampant in these categorization schemes thereby creating a source of possible confusion and delay during retrieval. This made multiple categorization cumbersome. As mentioned earlier in the paper, the incorporation of a search engine should address this need. Also, users felt that some information was inherently hierarchical in nature, and folders were better suited to these situations.

On the other hand, users found multiple categorization useful when the information being organized was highly interrelated and fell under many overlapping topics. Multiple categorization was reported to be more robust in situations where the topic space was initially unfamiliar or rapidly evolving. Finally, users told us that multiple categorization offered the unique advantage of being able to navigate a corpus from multiple perspectives.

In addition to these observations, users also informed us of desirable features that would improve their experience with an information management system. For example, users wanted to know which articles belonged to each category as they were organizing. Users also found that being able to organize categories into a hierarchy would reduce the amount of scanning needed to locate relevant categories.

We propose that future research investigate means of addressing the weaknesses of each paradigm individually while improving the user interface in ways elucidated by our surveys. Such a hybrid system should incorporate both a hierarchical categorization pane and a navigation pane in order to allow users to conveniently file and retrieve documents simultaneously. Furthermore, the inclusion of a keyword-based search engine should help to reduce the tendency to use multiple categorization as such. We are currently using Haystack as a basis for experimenting with these ideas.

6 Conclusion

In this paper we have pointed out several weaknesses inherent in existing information organization systems and considered the use of pervasive support for multiple categorization as a possible solution for addressing some of these weaknesses. Our user study showed that users experienced improved organization and retrieval performance. In addition, participants of our study appreciated several aspects of the multiple categorization paradigm, even in light of the fact that most had probably been extensively conditioned to use folders. Future work should continue to investigate the full potential of multiple categorization as a pervasive component of next generation information environments.

7 Acknowledgements

We would like to thank all of the participants of our study for generously giving us their time and comments. We would also like to thank Mark Ackerman, Jimmy Lin, and Vineet Sinha for their comments. This work was supported by the MIT-NTT collaboration, the MIT Oxygen project, a Packard Foundation fellowship, and IBM.

References

Abrams, D., Baecker, R., and Chignell, M. (1998), Information Archiving with Bookmarks: Personal Web Space Construction and Organization, *Proceedings of CHI 1998*, 41–48.

Agrawal, R., Bayardo, R., and Srikant, R. (2000), Athena: Mining-based Interactive Management of Text Databases, *Extending Database Technology*, 365–379.

Barreau, D. and Nardi, B. (1995), Finding and Reminding: File Organization from the Desktop, *SIGCHI Bulletin* 27(3), 39–43.

Cutting, D., Karger, D., Pedersen, J., and Tukey, J. (1992), Scatter/gather: A cluster-based approach to browsing large document collections, *Proceedings of the 15th SIGIR*, 318–329.

Dourish, P., Edwards, W., et al. (2000), Extending Document Management Systems with User-Specific Active Properties, *ACM Transactions on Information Systems* 18(2) , 140–170.

Hearst, M. (2000), Next Generation Web Search: Setting Our Sites, *IEEE Data Engineering Bulletin, Special issue on Next Generation Web Search*, Luis Gravano (Ed.), September 2000.

Huynh, D., Karger, D., and Quan, D. (2002), Haystack: A Platform for Creating, Organizing and Visualizing Information Using RDF, *Semantic Web Workshop, The Eleventh World Wide Web Conference 2002 (WWW2002)*. http://haystack.lcs.mit.edu/papers/sww02.pdf.

Kaplan, S., Kapor, M., Belove, E., Landsman, R., and Drake, T. (1990), Agenda: a personal information manager, *Communications of the ACM* 33(7), 105–116.

Lansdale, M. (1988), The Psychology of Personal Information Management, *Applied Ergonomics* 19(1), 55–66.

Malone, T. (1983), How Do People Organize Their Desks? Implications for the Design of Office Information Systems, *ACM Transactions on Office Information Systems* 1(1), 99–112.

Whittaker, S. and Sidner, C. (1996), Email Overload: Exploring Personal Information Management of Email, *Proceedings of CHI 96: Human Factors in Computing Systems*.