Document Co-Organization in an Online Knowledge Community

Harris Wu
University of Michigan
Ann Arbor, MI 48109, USA
harriswu@umich.edu

Michael D. Gordon
University of Michigan
Ann Arbor, MI 48109, USA
mdgordon@umich.edu

Kurt DeMaagd
University of Michigan
Ann Arbor, MI 48109, USA
demaagdk@umich.edu

Abstract
We introduce the concept of "document co-organization" and describe such a system. By document co-organization we mean that individuals are allowed to hierarchically organize documents personally and share their hierarchies with others, while the system generates a "consensus" hierarchy from these personal hierarchies, which provides a full, common, and emergent view of all documents. By allowing users to retrieve documents from their own organization (hierarchy), another user's, the consensus hierarchy, or a time-based hierarchy, we provide access corresponding to different characteristics of knowledge tasks: they are personal, collective, social, and time-sensitive. In a class website experiment, we show that for a complex knowledge task, hierarchies are used more frequently than search. One surprising finding is how often students use others' personal hierarchies.

Categories & Subject Descriptors: H.5.m. [**Information interfaces and presentation (e.g., HCI**]: Miscellaneous.

General Terms: Design, Human Factors, Theory.

Keywords: Document co-organization, knowledge community.

INTRODUCTION
A common challenge facing many online knowledge communities is how to organize large numbers of shared documents in a repository. When the number of documents becomes large, the way these documents are organized becomes critical for locating and sharing them with others in the community. We present an online document co-organization system that helps users build and utilize structures to support knowledge tasks.

We begin this paper with background research on online knowledge communities and document organization. Next we present our document co-organization system and discuss several research questions. We then describe our experiment using the system in a class and present initial evaluation results. We conclude our paper with a discussion and next steps of our research.

CHI 2004, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

BACKGROUND
An online knowledge community is defined as a group of knowledge workers jointly involved with a knowledge domain who meet and share their expertise electronically [2]. A knowledge worker can be described as someone who routinely uses information in his or her task performance. In essence, an online knowledge community is a community in which knowledge is shared and created. The documents used by this community become part of its shared repertoire [11]. Therefore, the knowledge workers must be familiar with some existing information in the document repository before they can participate in contributing new knowledge. Given that the structure of information is important for information sharing, proper organization of these shared documents is critical.

The organization of documents can take various forms, such as a directed graph (e.g. hypertext) or an ordered list (e.g. blog). The most popular way of organizing a large number of documents is to place documents in a hierarchy. For instance, books in a library, files on a computer and entries in yellow pages are all stored in hierarchies. A hierarchy is efficient to cope with a large number of documents, as n documents can be placed in a hierarchy with depth of mere $log(n)$. Also the complexity of a domain is often hierarchical in its nature [7].

Some online knowledge communities organize their shared documents in a common hierarchy. A common hierarchy provides a full, uniform view to all shared documents. Reference to a document in the hierarchy is convenient and the same for all users. For example, a class website may contain folders, subfolders and files within these folders. Usually a central authority of the community maintains such a hierarchy. Maintenance of such a hierarchy can be very expensive.

However, a common hierarchy cannot accommodate conflicting individual perspectives. Most knowledge tasks including learning and knowledge creation are creative and personal. A common document structure adversely affects these tasks, for example by making documents more difficult to find and adding cognitive load to individuals who need to map the way they would organize information onto the way the system does. This discourages community members from contributing to the information repository. Who wants to contribute to an online information

repository if it is difficult to find what you have posted? An unappealing alternative is to file the same document multiple times – once in the common hierarchy, and separately for personal use.

Further, organizing documents involves reflection upon these documents and the collection, which is important to individuals' knowledge acquisition [10]. Personal organization of documents also elicits tacit individual knowledge [1]. So sharing individual organizations of documents among the community promotes knowledge building as a social process. Note that although browser bookmarks allow a user to organize URL's, browser bookmarks cannot be easily shared and not all documents can be accessed by a browser.

Most online knowledge communities lack the support for individuals to organize shared documents into personal hierarchies. One reason is that personal organization presents system design challenges. Storing, presenting and allowing users to interact with multiple hierarchies can be difficult, especially if the number of community members or documents is large. An alternative to personal organization is personal tagging. To accommodate individual perspectives, a system allows individuals to assign properties to documents in the repository. Some researchers [3,5] have developed "placeless" document repositories where retrieval is based on those user-assigned properties. In those systems documents are not accessed through hierarchies or other explicit structures. Rather, documents are accessed by property-based search. It is not clear whether property-based search can replace hierarchies in helping knowledge workers.

Even if personal organization of documents were supported for an online community, it still would be desirable to have a common hierarchical organization for the entire collection. Any individual will only organize certain portions of a large repository, so individual organizations provide an incomplete and possibly idiosyncratic view of the collection. For repository users the collection of individual views falls short of the ideal. Our interest is in preserving these partial, individual hierarchical organizations and building from them a consensus view of the entire collection.

THE DOCUMENT CO-ORGANIZATION SYSTEM

We develop our document co-organization system based on an open source web content management system, the Everything engine (http://everydevel.org). The Everything engine supports various forms of content, including anything that can be uploaded as a file. Besides the core Everything engine, many open source add-on modules have been developed containing features such as threaded discussions, instant messaging, etc. The core engine with add-on modules provides a comprehensive set of features for an online document repository and online community.

Documents in an Everything repository form a graph structure from hyperlinks between them. Document authors can include hyperlinks to other documents within and outside the repository. Everything also allows for a central authority to create a hierarchy of documents, through the use of categories. A category is a special type of document that consists of a list of links to other documents. Administrative users can create categories and add documents or subcategories to these categories. A typical user can view but not modify these categories.

The existing Everything software lacks support for users to personally organize documents. Below we describe the additional features that we add to Everything to allow for document co-organization. Figure 1 shows a screenshot of our extended Everything system.

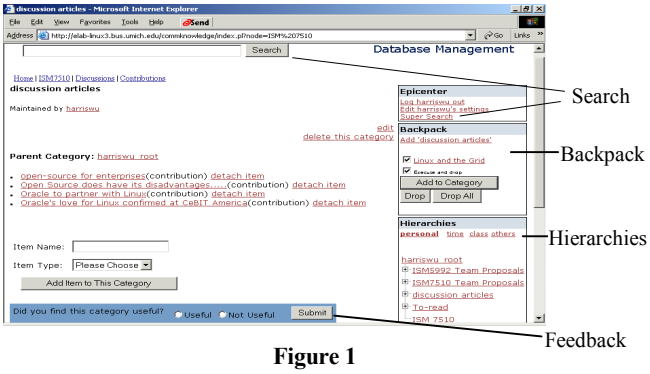

Figure 1

New Features for Document Co-organization

Document organization and many other knowledge tasks are personal. So we allow a user to build a personal hierarchy of documents. Besides granting users permissions to create and modify his/her own categories, we add two components to Everything: *Backpack* and *Hierarchies*. Backpack is a staging area where users keep a list of documents to be further organized. Hierarchies allow a user to switch between several different hierarchies; but at any given time only one hierarchy is displayed. The Hierarchies feature is implemented using Javascript and allows for user interactions such as adding or deleting a category. Both the Backpack and Hierarchies components appear in the right frame of the browser throughout a user's navigation in our system. To build a personal hierarchy, a user first adds interesting documents to the Backpack when these documents are encountered. The user goes to a category to be appended, chooses documents from Backpack by checking corresponding checkboxes, and then clicks the "Add to Category" button to add these documents to the given category (Figure 1).

Document organization and other knowledge tasks are also social. To support this perspective, we allow users to utilize other people's personal hierarchies by switching to *others* in Hierarchies. By default, categories in a personal hierarchy have read permission granted to the public. The *others* hierarchy (shown in Figure 2) leads to individuals' personal hierarchies. For privacy reasons, a user can grant read permission on his/her categories to selected users, or

nobody at all. To facilitate collaboration, a user can also grant write permission to his/her categories to other users. The Everything engine has native support for sophisticated access control mechanisms.

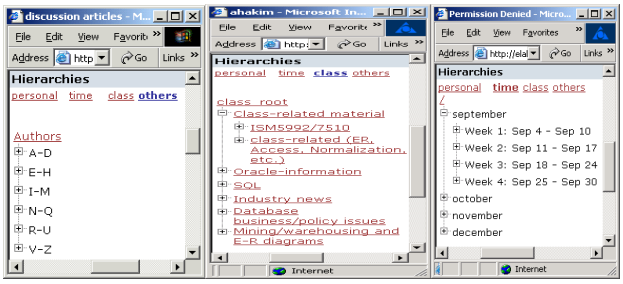

Figure 2. The *others*, *common*, and *time* hierarchies

In addition to being personal and social, document organization is a collective task. An individual hierarchy only provides an incomplete view of a collection. The set of different individuals' hierarchies may present an inconsistent view since individuals organize somewhat idiosyncratically. Our system generates a common "consensus" hierarchy from personal hierarchies using a hierarchical clustering algorithm adapted from [12]. The common, co-organized hierarchy emerging from individual hierarchies not only accommodates individual perspectives but also provides a full, unified view to the repository. Figure 2 shows the common hierarchy labeled as "class" for the screenshot taken from a class website.

Finally, knowledge tasks are also time-sensitive. Users may be only interested in documents created, modified or accessed within a certain time period. The *time* hierarchy structures all documents in the repository based on a document's creation time, which can reconfigured to be based on the time of other actions on a document [4].

Research Questions and Data Collection Mechanisms

The goal of our research program is to provide improved access to information through co-organization. In this effort, we are interested in exploring the value of document co-organization relative to other means of locating information. For information retrieval, an alterative to using hierarchies is search engines. In fact, in many information repositories, users rely on keyword or attribute based search to access documents. Search can be very fast – often faster than clicking to navigate a hierarchy. Search can also return high quality documents ordered by relevance or in other ways.

However using a hierarchy often introduces a lower cognitive load, as navigating the hierarchy involves recognition, whereas specifying a query involves recall. Using a hierarchy is also more interactive, as a user can go up and down a hierarchy to refine or adjust his/her focus. For a human-built hierarchy, the structure itself has extra descriptive ingredients not included in the content of documents. Hierarchy building can be seen as an environmental enrichment activity in Information Foraging theory [6]. However, the benefit of enrichment need not be limited to a single forager. Rather, in our system users can share personal hierarchies with each other, thus using these hierarchies in a social process.

Much research in information retrieval has confirmed the Cluster Hypothesis, which states that closely associated documents tend to be relevant to the same requests [9]. For many knowledge tasks, users benefit by looking at document clusters. Both documents close to each other in a hierarchy and documents in a search result may be considered closely associated. However, it is difficult to design a query so that returned documents are associated in a particular way. For example, searching for "Jaguar" may return documents about cars, football or animals. In contrast, a user can use a hierarchy that associates documents on a particular dimension suitable to a given task. Sometimes the association between documents is tacit knowledge that cannot be easily codified, however can be elicited by hierarchies [1].

Thus the questions we want to answer using our system are: First, for a given task, do individuals use hierarchies more often than search? Second, which hierarchy is used most often? Obviously choosing which tool to use depends on the knowledge task. However, our research questions serve as important first steps that inform us of the user preferences for a given knowledge task.

Our system is capable of collecting data to answer the above questions. Everything comes with comprehensive search functionality, including both keyword based full-text search and advanced attribute search. We extended Everything so that all user navigations are captured in detail. We also added features to collect user feedback on whether a document/category is useful (shown in Fig 1).

CLASS WEBSITE EXPERIMENT & INITIAL RESULTS

We deployed our system in a class website for two class sessions on data management in Fall 2003 with a total of 45 students. A student typically made 2-3 contributions related to data management to the website every week. A majority of contributions were short write-ups that discuss news articles or other Web resources, or other students' contributions. These write-ups contain hyperlinks to references. There were also contributions in Microsoft Word, PDF, or JPEG formats. Over the semester students made about 1,400 contributions. Students organized documents interesting to them into their personal hierarchies. They were told that the final paper would utilize the documents in the class website, thus being able to find relevant documents was critical. The final paper assignment was not revealed until three weeks before the end of semester. The paper was a complex, hypothetical case analysis that would potentially utilize most of the class' contributions. It required that all references be from contributions to the website. We analyzed user navigation

data collected over these three weeks, during which students focused on the single task of writing the paper.

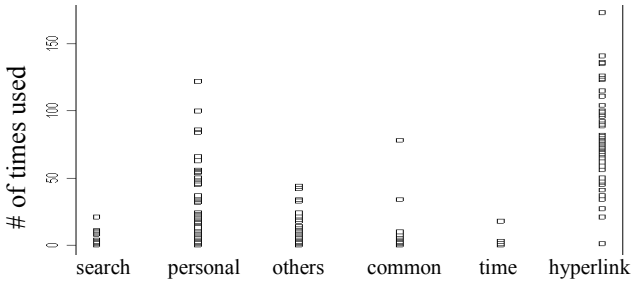

Figure 3. Strip chart on usage of different tools

The strip chart in Figure 3 shows the usage summary comparing search and different hierarchies in retrieving documents. An alternative to using search and hierarchies is browsing documents through hyperlinks contained in them. As a comparison, the numbers of hyperlinks traversed by students in browsing are also shown. Clearly students use their personal hierarchies more often than search. However it is surprising that students use the *others* hierarchy even more frequently than using search. We use a two-paired t-test and the difference is significant at a 0.01 level ($t = 3.0289$, $df = 43$). This result confirms that knowledge tasks are indeed social as well as personal. By examining the data closely, it seems that students tended to look at hierarchies of the top students in the class, although this cannot be confirmed statistically. The difference between usage of the *common* hierarchy and usage of search is not statistically significant. Interestingly as shown in Figure 3, at least two students have used the common hierarchy significantly, which indicates that some students see the automatically generated "consensus" hierarchy as useful. The *time* hierarchy is only occasionally used although one student used it 18 times. That student perhaps was trying to find documents he/she saw around a certain time.

DISCUSSION AND NEXT STEPS

Our goal is to understand how to provide better structuring tools for documents that let users work personally, with each other's structures, and at the emergent "consensus" level. The above class experiment is the beginning of our investigation on how users utilize various structures in a knowledge community for different tasks. The experiment is imperfect. There were some technical problems with updating the consensus hierarchy, which caused some downtime of that hierarchy during the experiment. Otherwise the usage of the consensus hierarchy may have been higher. We are also working to improve the algorithm building the consensus hierarchy, for which we omit the detail due to limited space. The experiment is limited in that the final paper assignment is the only task evaluated. We plan to extend the experiment to a variety of knowledge tasks and other online communities.

For the data collected last month from the class experiment, we have just had time for an initial analysis. Initial results show that students frequently utilize their personal hierarchies as well as each other's hierarchies. In the next a few months, we will regenerate user navigation sessions and identify patterns such as the longest repeated sequences [8]. These patterns will help us to understand more about how users perform knowledge tasks using a combination of tools. Together with user feedback these patterns will show how often different tools lead to useful documents, the circumstances that lead searchers to switch among tools, and the likelihood of switching. We will also perform qualitative user studies with interviews and questionnaires.

It deserves mention that developing on an open source project has many advantages. The Everything engine is robust and scalable, being used by many popular online communities including everything2.org and PerlMonks.org. We plan to contribute our development back to the open source community.

REFERENCES

1. Cooke, N.J., 1994, Varieties of knowledge elicitation techniques, *Intl. Journal of Human-computer Studies*, 41: 801-849.
2. De Vries, S, Bloemen, P. and Roossink, L. Online Knowledge Communities, *Proc. WebNet 2000.*
3. Dourish, P., Edwards, W.K. Extending document management systems with user-specific active properties. *ACM Trans. on Info Systems*,18(2),2000.
4. Gordon, M. and Moore, S. Depicting the use and purpose of documents to improve information retrieval. *Information System Research*, 10:1, 1999.
5. Huang, J. and Michiels, J. Exploring Property-based Document Organization in a Collaborative Note-Sharing System. *Proc. CHI 2000,* Pages 327-328.
6. Pirolli, P. and Card, S.K. Information Foraging. *Psychological Review*, 106 (4). 1999, 643-675.
7. Simon, H. A. *The Sciences of the Artificial*, MIT Press, Cambridge, MA, 1969
8. Tauscher, L. and Greenberg, S. (1997) Revisitation Patterns in World Wide Web Navigation. *SIGCHI* 97.
9. Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London.
10. Welbank, M. An overview of knowledge acquisition methods, *Interacting with Computers* 2(1):83-91,1990.
11. Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge U. Press.
12. Wu, H., Gordon, M., DeMaagd, K. and Bos, N. "Link Analysis for Collaborative Knowledge Building," Proc. *ACM Hypertext* 2003, Pages 190-191.