Collaborative Filing in a Document Repository

Harris Wu
University of Michigan
701 Tappan Street
Ann Arbor, MI 48109
734-647-7667
harriswu@umich.edu

Michael D. Gordon
University of Michigan
701 Tappan Street
Ann Arbor, MI 48109
734-763-1387
mdgordon@umich.edu

ABSTRACT

We introduce an emergent, collaborative filing system. In such a system, an individual is allowed to organize a subset of documents in a repository into a personal hierarchy and share the hierarchy with others. The system generates a "consensus" hierarchy from all users' personal hierarchies, which provides a full, common, and emergent view of all documents. We believe that collaborative filing helps translate personal, tacit knowledge into sharable structures, which help the user as well a community of which he or she is a part. Our filing system is suitable for any documents from text to multimedia files. Initial results on an experimental website show promise. For a knowledge task involving extensive document retrieval, hierarchies are not only used frequently but are also effective in identifying high quality documents. One surprising finding is how often subjects use others' personal hierarchies, and upon close examination, social networks play a key role as well.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, clustering.*

General Terms

Design, Experimentation, Human Factors, Theory.

Keywords

Collaborative filing, document organization, co-organization.

1. INTRODUCTION AND MOTIVATION

For a knowledge community to utilize a shared document repository, users need to be able to find documents. When the number of documents in a repository becomes large, the organization of documents becomes critical. We present a filing system that helps users collaboratively build and utilize structures to support filing, retrieval, sharing, and other knowledge tasks. Below, we first describe the problem motivating this research. Then we describe a collaborative filing system in the context of a class website. Finally, we present initial evaluation results and conclude with next steps of our research.

The organization of documents in a repository can take various forms such as a directed graph (e.g. hypertext) or an ordered list (e.g. blog). A hierarchy is an efficient way of organizing documents, as *n* documents may be placed in a hierarchy with depth of mere *log(n)*. Also the complexity of a domain is often hierarchical in nature. Based on the widely confirmed cluster hypothesis [3], documents close to each other in a hierarchy tend to be relevant to the same requests. Books in a library, files on a computer and entries in yellow pages are all stored in hierarchies. The Yahoo Web Site Directory contains categories, subcategories, and finally lists of Websites.

Many shared document repositories are organized in a common hierarchy. A common hierarchy provides a full, uniform view to all documents in the repository. Reference to a document is convenient and the same for all users. However, a common hierarchy cannot accommodate conflicting individual perspectives. Also, a common hierarchy is often out-of-date for document repositories with many new incoming documents on emerging topics. It is far more natural to file documents in a way that matches one's personal perspective on the relationships among documents and topics than to file them using a common structure that does not correspond to one's particular knowledge tasks. Thus, with most shared repositories, one is faced with filing information in a way that may make subsequent retrieval more cognitively demanding, or filing everything twice – once for personal use, and again for the community. Further, personally organizing documents codifies tacit knowledge. Unlike an approach where a system algorithmically classifies documents into an existing taxonomy, or one where a hierarchy is derived from the content of documents using clustering techniques, personal organization of documents adds to the repository genuine structural information beyond the original document content. Further, personal organization is much more reliable than other algorithmic organizations, especially for multimedia documents such as video clips for which content cannot be effectively parsed. In Information Foraging theory [2], document organization is an enrichment activity that not only reduces foraging (search) costs but also returns more valuable information. Unlike individual enrichment activities such as refining a keyword query, document organization results in reusable structures for a community.

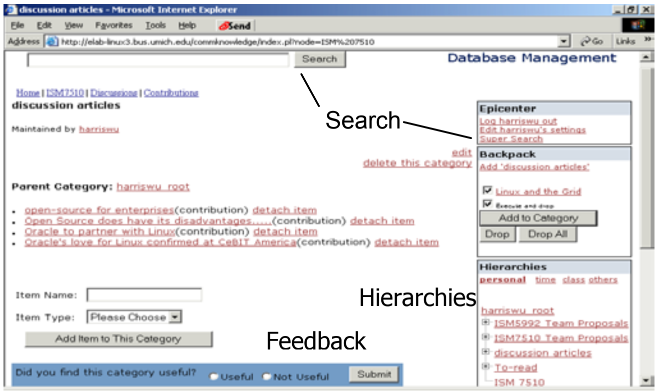
If a shared repository supports personal organizations, it still is desirable to have a common structure for the entire collection. Any individual can only organize certain portions of a large repository, so individual organizations provide incomplete and possibly idiosyncratic structures of the collection. For repository users the collection of individual structures falls short of the ideal. We are interested in preserving these partial, individual structures and building from them a consensus structure of entire collection.

2. SYSTEM AND EXPERIMENT

We developed a collaborative filing system on top of Everything (everydevel.org), a popular open-source content management system. Our system provides structuring capability in addition to

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

Everything's content management capability, which includes sophisticated search functions. We added features to capture all user navigation activities on the website, as well as feedback on whether a retrieved document was judged useful or not. We evaluated our system on a class website (Figure 1) used by 45 students who contributed 1400 document in a variety of formats (text, HTML, Word document, JPG, etc.) in the Fall 2003 term.



Address: http://elab-linux3.bus.umich.edu/commknowledge/index.pl?node=ISM%207510

Database Management

Home | ISM7510 | Discussions | Contributions

discussion articles

Maintained by harriswu

edit
delete this category

Parent Category: harriswu_root

- open-source for enterprises(contribution) detach item
- Open Source does have its disadvantages.....(contribution) detach item
- Oracle to partner with Linux(contribution) detach item
- Oracle's love for Linux confirmed at CeBIT America(contribution) detach item

Item Name:

Item Type: Please Choose

Did you find this category useful? ○ Useful ○ Not Useful

Epicenter
Log harriswu out
Edit harriswu's settings
Super Search

Backpack
Add 'discussion articles'

☑ Linux and the Grid
☑ Execute and drop

Hierarchies
personal time class others

harriswu_root
⊞ ISM5992 Team Proposals
⊞ ISM7510 Team Proposals
⊞ discussion articles
⊞ To-read
ISM 7510

Figure 1. A screenshot of the class website

Users can find documents in the repository by search, following hyperlinks, or navigating several hierarchical structures. A *Personal* hierarchy contains a user's personal categorization of documents. Each user creates his or her own categories and category labels. For reasons of space, we omit other details explaining how a user builds a personal hierarchy. One's personal hierarchy can be shared with other users, who typically only have read access. An *Others* hierarchy contains links to Personal hierarchies other than one's own. The *Class* hierarchy is the consensus hierarchy built by an agglomerative hierarchical clustering algorithm. The input to the algorithm is a categorization matrix A containing one row for each document. The columns in the matrix represent categories in individuals' personal hierarchies. $A_{ij} = 1$ if document i is in category j, and 0 otherwise. The similarity between documents is defined using the Jaccard measure between the row vectors, which is the number of 1's in the "intersection" of the vectors divided by that in the "union" of the vectors. The algorithm constructs the consensus (class) hierarchy by merging the most similar documents into groups, and then merging the most similar groups using an average linkage method [1]. The categories in the consensus hierarchy are labeled using the keywords from titles of the documents within it as well as the labels of subcategories, if applicable. The consensus hierarchy is dynamically re-generated periodically, which takes only a few seconds. For large document databases, there are several approximation techniques available for agglomerative hierarchical clustering [4], so our method is scalable. Note that there is no central administrative body involved in creating the consensus hierarchy. From a complex system perspective, the consensus hierarchy emerges from self-organization of individual hierarchies. Finally, a *Time* hierarchy categorizes documents based on the day (week, month) a document is contributed.

The "critical" task in our experiment was an end-of-term course research paper for which most documents in the repository would be potentially useful. Students focused on this assignment for the last month in the term. So far we have just had time for an initial analysis of the data collected for this critical task period. Figure 2 shows the individual usage of search and hierarchies in retrieving documents, where each dot indicates an individual's usage of a certain mechanism. The *personal* hierarchy and the *others* hierarchy are used more often than search. Some students extensively used the class hierarchy -- the self-organized, consensus structure emerging from individual hierarchies. Hierarchies also seem to identify useful documents more often than other methods of retrieval, as measured by user feedback. Overall the ratio of retrieved documents being rated as useful to not useful is about 5 to 1. These would be accessed by following hyperlinks, search, or via hierarchies. However for documents accessed through hierarchies, the ratio of being rated as useful to not useful is 25 to 1. It is interesting that the *others* hierarchies are used so often. Upon closer examination, there are a few user "cliques" based on common research interests that frequently use each others' hierarchy. This result suggests that collaborative filing is more useful for a community with strong social ties, or a group focusing on the same task.

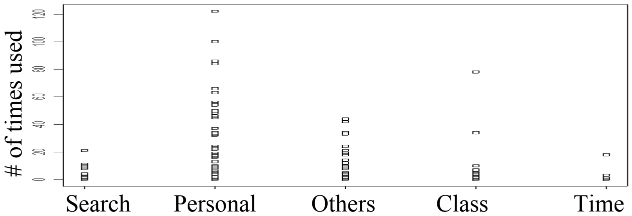


Figure 2. Frequency using hierarchies and search

3. NEXT STEPS

The data we have captured is very rich, including users' step-by-step actions on the website. We plan to further analyze users' action patterns, and perform qualitative studies using interviews and questionnaires. We hope to identify if and how collaborative document organization helps users in information retrieval. In the experiment we described, some technical problems with updating caused some downtime for the consensus hierarchy. In experiments under way, we are exploring if its usage will be greater without this defect. We are also working on different algorithms for building the consensus hierarchy, and plan to evaluate them against each other. We would like to extend the experiment to different retrieval tasks, and other environments such as corporate intranets and Internet knowledge communities. Note that while some research has tried to utilize browser bookmarks, bookmarks can organize only web documents with an underlying hyperlink structure. Our research does not assume an initial hyperlink structure among the documents.

4. REFERENCES

[1] Jobson, J.D. (1992). Applied Multivariate Data Analysis. Springer-Verlag, New York.

[2] Pirolli, P. and Card, S.K. Information Foraging. Psychological Review, 106 (4). 1999, 643-675.

[3] Van Rijsbergen, C. J. (1979). Information Retrieval. Butterworths, London.

[4] Willett, P. Recent trends in hierarchical document clustering. Inf. Process. Manage. 24,5 (1988), 577-597.

[5] Marais, H. and Bharat, K. Supporting cooperative and personal surfing with a desktop assistant. ACM UIST'97.