

Leveraging Metadata to Improve Information Retrieval in Directory Interfaces

Alexander J. Faaborg
Cornell University Human Computer Interaction Group
Kennedy Hall Ithaca, NY 14853
ajf15@cornell.edu

ABSTRACT

This study describes how metadata can be used to organize documents into hierarchical structures that filter against each other. It then discusses several experiments that were conducted to test the underlying usability concerns of this type of organizational system.

Keywords: Information Retrieval, Metadata, Dublin Core, Hypertext, Hierarchical Structures, Semantic Indexing.

INTRODUCTION

Along with the dramatic increase in information availability over the last few years have come advancements in the way information can be both organized and retrieved. Recently XML and RDF have emerged to bring a semantic quality to online information and members of the Dublin Core Metadata Initiative have developed an element set to thoroughly describe any piece of information. While many people have written on how metadata will increase searching accuracy, metadata's impact on how we browse through an online collection of documents may be equally significant. This study focuses on how a document's metadata can be used to improve information retrieval in terms of a directory interface.

Organizing Information into Interlinked Hierarchies

If the metadata attributes assigned to a document come from controlled vocabularies, it is possible to automatically organize a group of documents into a hierarchical structure. This technique can be applied using each of the document's metadata elements, creating a navigational interface that allows users to filter attributes against each other and locate information with fewer navigational steps. Consider a document where several of its metadata attributes come from controlled vocabularies. These attributes could be the document's subject, the document's target audience, and the document's geographic coverage. When browsing through the collection of documents, selecting a particular audience reduces the number of options available under subject and geographic coverage. If these metadata attributes each come from controlled hierarchies that contain multiple levels, then the filtering effect is even more drastic.

An extensive amount of research has been done on breadth vs. depth in information structures, particularly by Larson and Czerwinski [2] at Microsoft Research. They concluded that hierarchies two levels deep were more usable than hierarchies of greater depth. By filtering metadata attributes against each other, the overall depth of the navigational system can be reduced significantly.



One hierarchy interface
second level exposed



Three hierarchy interface
second level hidden

Figure 1: The current one hierarchy interface and the proposed multiple hierarchy interface

Information Clusters and Personal Cognitive Maps

As people search for information, they will simultaneously group documents together. Pirolli and Card [1] from Xerox PARC advocate that as people forage for information they are likely to cluster what they find into unique groups based on the relevance of their particular query. Robertson and Czerwinski [5] from Microsoft Research note that in terms of locating information on the Web, these clusters of information often come in the form of a folder in their favorites menu, or a document that contains a list of related hyperlinks.

A clear way of improving information retrieval in a directory interface is to have a system robust enough to already include these clusters. However, to achieve this we must understand the attributes that users consider when personally grouping documents together. By creating a navigation system that allows a user to filter on each of these attributes, nodes in the interlinked hierarchies should begin to resemble their personal information clusters. This is because the system allows them to filter on the exact same attributes they use to personally group documents together. In the optimal case, the system would have enough nodes to be well customized to each particular user. Their cognitive map of how the information space should be structured would then closely match the actual information space.

The Advantage of Using Multiple Metadata Attributes

The advantage to using metadata to organize online repositories is twofold. By increasing the number of metadata attributes applied to each document users will be able to filter against these attributes decreasing the number of navigational steps and reducing the overall depth of the navigational system. Also, by increasing the number of ways data is catalogued, it is more likely that users will be able to navigate to information based on the attributes they personally feel are the most relevant.

Even if we cannot determine every feasible attribute to describe a document, the 15 Dublin Core metadata elements constitute a good start. In this study we simplify the number of permutations by analyzing how information retrieval can be improved with just three metadata attributes.

Testing the Underlying Usability Concerns

[I will include a more general introduction here when the study is expanded to include the multiple-correct-hyperlinks experiment (and possibly a test of a working directory system).]

Determining the Advantage of Three Attributes

While an extensive amount of research has been done on the correct balance of breadth vs. depth in navigational structures [5], the process of selecting which metadata attributes will best match a user's cognitive map of the information space has not been as thoroughly explored. To organize information available on the Cornell web site, we selected three metadata attributes to test: subject, audience and geographic location. We then conducted a preliminary study to see how well subjects responded to these types of metadata as a classification tool. The experiment consisted of a series of timed reaction trials, where the subjects had to determine if various keywords related to web sites they had just seen.

METHOD

Subjects

Nine subjects ran in the experiment. The group consisted of two female subjects, and seven males. Six of the subjects were undergraduate students at Cornell University; the remaining three subjects were from Cornell's Human Computer Interaction Group. The subjects were aware that the keywords they were evaluating fell into three metadata categories: subject, audience, and geographic location.

Materials

The experiment was conducted online, with subjects using their own personal computers. Because of this, the size and resolution of their monitors varied. All subjects used Windows, browsing the web with Internet Explorer 5.5 or 6.0. Also, all subjects were using an Internet connection faster than 300kbps, although the experiment was designed to insure that connection speed did not affect the timed portions of the experiment. The subject's responses were recorded using the 'J' and 'F' keys on their keyboard.

Procedure

A trial would begin with a web page informing the subjects to press the space bar when they were ready to continue. Once the subjects pressed the space bar a web page from the Cornell web site loaded and was displayed for 10 seconds. These web pages were selected using the criteria that they must:

- 1) Contain actual content and not simply be a list of other web sites
- 2) Contain all three metadata attributes, written directly on the page
- 3) Together, the pages must adequately cover the full range of content available on the Cornell web site.

Beyond these three requirements their selection was essentially random. Once the subjects had viewed the page for 10 seconds they were taken to another screen informing them to press the space bar when they were ready for the keyword, and reminding them that the 'F' key mapped to false, and the 'J' key mapped to true. The subjects' task was to determine if the keyword related to the web page they had just seen. As soon as the subjects viewed the keyword, the computer began counting the number of milliseconds it took them to respond. These reaction times were then sent to a server to be analyzed.

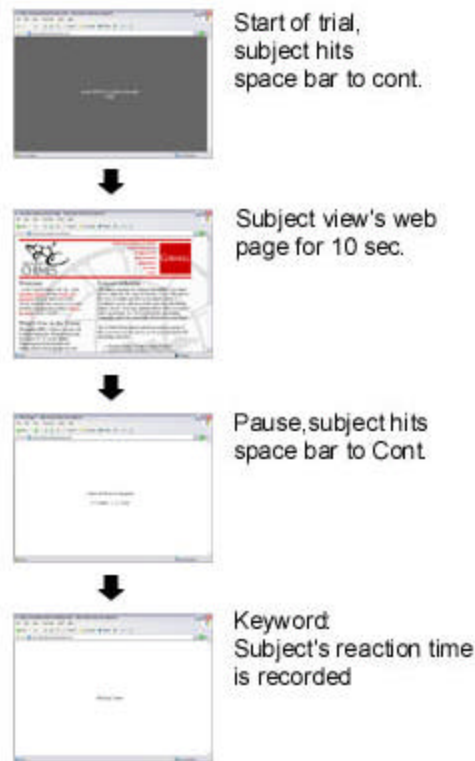


Figure 2: The steps in one trial

In the set of 30 trials, 15 of the pages matched to keywords that were not related, and 15 pages matched to keywords that were subject, audience, or geographic based descriptions of the document. To collect information on each of the three types of metadata, and assure that each subject only saw a document once, there were three different versions of the experiment. Each version of the experiment contained an equal amount of the different metadata types. The order of related and unrelated trials was randomized, as well as the order of metadata types for the related pages.

RESULTS

Accuracy

Initially we did not plan to use accuracy as a benchmark since the trials were set up to be inherently easy. However, overall accuracy was surprisingly bad. On average 22% of the responses were incorrect, this broke down as follows:

Type of Error	Percent of Total Errors	Average Reaction Time
Answering True to a False subject	1%	2002 ms
Answering False to a True subject	5%	1391 ms
Answering True to a False audience	20%	1814 ms
Answering False to a True audience	21%	2575 ms
Answering True to a False geographic coverage	16%	2134 ms
Answering False to a True geographic coverage	27%	2002 ms

The incorrect responses were not included in the average response times for different types of metadata.

Average Response Time for Different Types of Metadata

Across the trials, the average response time for subject metadata was the fastest at 1671 ms, followed by geographic coverage metadata at 1965 ms, and audience metadata at 2053 ms. This broke down as follows.

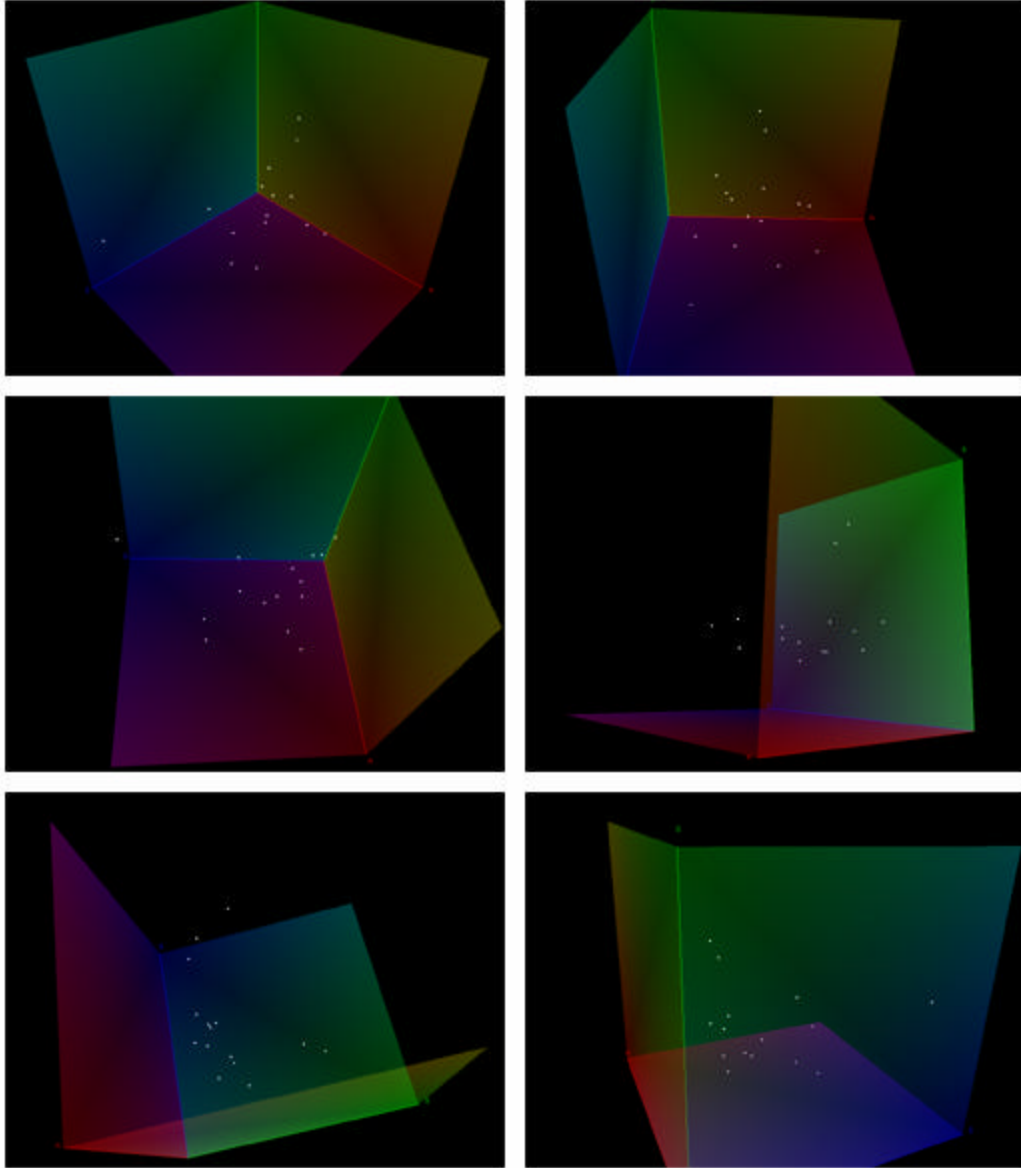
Trial	Audience	Subject	Geographic
2	2125	3270	1008
3	2504	1255	3185
6	2974	1535	1626
8	2243	2945	1187
9	955	1267	2404
11	2248	1658	1272
13	1942	1217	1658
14	1182	2025	4867
16	3230	1315	1355
17	1265	1743	912
21	3165	1602	3184
23	2199	1352	1973
25	1703	1056	2404
27	1488	1577	1347
28	1578	1257	1102
Average	2053.283	1671.5	1965.489

In this figure the best time for each trial is represented in red. You can hover over the reaction to see the keywords tested, or select a trial to see the web page the keywords related to.

DISCUSSION

Improving Information Retrieval

To better visualize this data we plotted the trials in three-dimensional space:



In these images each web page trial is represented as a dot. The reaction times for the three different types of metadata are represented by the distance that dot is from each wall. Note that while some pages are far from one of the walls, they are closer to another wall. This varies between all the pages. In terms of the reaction times it means that while subjects were on average slow with one type of metadata, they were also on average fast with another type of metadata. Because the dots form a three dimensional cloud, we see that no one type of metadata best describes all pages. Numerically, we can see that when we take the fastest type of metadata for each trial, the average reaction time across all of the trials reduces to 1219 ms, as apposed to the 1671 ms reaction time for just subject metadata. This represents a 27% improvement.

Effect on the Design of Directory Structures

According to these findings, by increasing the number of attributes used to describe data, subjects on average reacted faster than if the data had been described using just one attribute. To some extent this is intuitive: Trial 9, a bulletin for graduate students, is best described by audience, while trial 17, a page about the Mathematics Library, is best described by its geographic location.

From these findings it may seem that we should not interlink the hierarchies in a directory structure but instead organize documents under which metadata attribute they are identified with the quickest. The problem with this design is that users will likely vary in respect to which type of metadata they responded to the quickest for different documents depending on their personal cognitive map. This should theoretically be compounded as the audience becomes more diverse. To generate an improvement in information retrieval, the documents in the directory should be available under each type of metadata.

[Note: the following sections will change as more experiments, and possibly a test of an actual filtering directory interface are conducted.]

Weaknesses and Limitations of Experiment

A reasonable concern with this study is that the task of determining if a keyword is related to a web page is significantly different from browsing for a web page in a directory. While the user task is different, their cognitive map of how the directory should be organized will be the same. The assumption is that the user's fast reaction time for a certain type of metadata in this study will map to them clicking on a particular option in the directory. If this assumption is correct, the effects found in this study should carry over to subjects browsing for a document. To test the results of the timed reaction trials, further experiments using the actual directory interface should be conducted in the future.

Although this is just a preliminary study addressing one of the underlying usability concerns of a filtering directory interface, one of the biggest limitations was the small and not very diverse subject group. More subjects will be tested in future experiments.

While the findings showed that increasing the number of attributes used to describe data led to an improvement in average reaction time, it did not take into account other underlying usability concerns for this kind of directory interface. For instance, having each piece of content available under three different hierarchies could confuse users if they assume that only one selection is correct, yet they see multiple possibilities. This usability concern will be the focus of the next study.

CONCLUSION

By using three metadata attributes to describe documents, this study found a 27% improvement in average subject reaction time compared to using just one attribute. Since users browsing for information will use the same cognitive map as when they conducted these trials, this improvement in information retrieval should carry over to a directory interface. However, there are other underlying usability concerns to take into consideration, and these will be addressed in the future.

ACKNOWLEDGEMENTS

I would like to thank fellow college scholar Adam Barth for the custom made three dimensional graphing application, as well as Stewart Whitman, Scott Lamb, Tom Fearon, and Ben Canon from Lehman Brothers. They helped me evaluate many of the ideas used in this study while I created

a proposal for redesigning the Lehman Brothers Intranet and deploying a firm wide metadata standard for online content.

REFERENCES

1. Peter Pirolli , Stuart Card, Information foraging in information access environments, Conference proceedings on human factors in computing systems, p.51-58, May 07-11, 1995, Denver, Colorado, United States.
2. Kevin Larson, Mary Czerwinski, Web page design: implications of memory, structure and scent for information retrieval, Conference proceedings on human factors in computing systems, p. 25-32, 1998 Los Angeles, California, United States.
3. Peter C. Weinstein, Ontology-based metadata, Proceedings of the third ACM Conference on digital libraries, p.254-263, June 23-26, 1998, Pittsburgh, Pennsylvania, United States.
4. Berners-Lee 1998a. Tim Berners-Lee. World Wide Web Design Issues: A Roadmap to the Semantic Web, [Online: <http://www.w3.org/DesignIssues/Semantic.html>], 1998.
5. George Robertson , Mary Czerwinski , Kevin Larson , Daniel C. Robbins , David Thiel , Maarten van Dantzich, Data mountain, Proceedings of the 11th annual ACM Symposium on User Interface Software and Technology, p.153-162, November 01-04, 1998, San Francisco, California, United States