Evaluating Category Membership for Information Architecture

Craig S. Miller
Sven Fuchs
Niranchana S. Anantharaman
Priti Kulkarni
DePaul University
243 S Wabash Ave
Chicago, IL 60604
USA
cmiller@cs.depaul.edu
+1-312-362-5085

DRAFT as of 21 September 2006
DO NOT QUOTE!

ABSTRACT
We examine two approaches for evaluating category membership within an information taxonomy: card sorting and a simple statistical approach based on the word frequencies of content items. To assess these two approaches, we developed a category structure for a web site based on card sorting. Metrics from the card sorting were obtained to predict problematic tasks. Potentially problematic tasks were also identified using word frequencies of the content items under each category. We then conducted a web navigation study to compare actual user performance of tasks identified by the two approaches. Despite its simplicity, the word-based statistical approach produced marginally better predictions. In particular, it predicted problematic tasks that took nearly twice as much time on average to complete than the remaining tasks in the web navigation study. We review these results and discuss the general usefulness of card sorting and word-based statistics for designing information taxonomies such as those used in web sites and menu systems.

Author Keywords
Information architecture, card sorting, web navigation, evaluation methods.

ACM Classification Keywords
H5.4. Information interfaces and presentation (e.g., HCI): Hypertext/Hypermedia.

INTRODUCTION
As the size and number of repositories of digital content grows, users often find it difficult to find the item they seek. To some extent, this difficulty is being addressed with the help of powerful keyword search engines, where the user provides relevant terms or phrases and the search engine returns links to pages that hopefully contain items that the user is seeking. While keyword search can help users in some cases (e.g., when queries are well-defined [13]), other tasks may be more suited for browsing labeled options [15], or there may just be a user preference for one or the other [19]. As long as information navigation remains a prevalent strategy for accessing digital content, information architects need effective strategies for labeling and structuring content so that users can find the items they seek.

For creating effective categories and labels, participatory design methods in the form of card sorting are often used (see [11,18] for examples of use). To conduct a card sort, a practitioner asks representative users to organize digital content by placing items (often described on index cards) in categories. For an open card sort, the users create their own categories and group cards as they see fit. For a closed card sort, users are given a set of categories and labels and are then asked to place each card in what they see as the best category. The open card sort is useful for exploring possible organizations. The closed card sort is useful for validating an organization and learning where specific items should be placed. Both types are frequently used for designing the information architecture. Their application can range from informal data gathering to a rigorous quantitative analysis.

Informal approaches mostly rely on the practitioner's judgment to find trends and classification themes whereas formal approaches may use classification frequencies and cluster analysis tools to organize content [11].

Despite the common use of card sorting for creating navigation structures, few empirical studies or theoretical analyses have been conducted to assess the utility of card sorting methods. Our goal in this paper is to provide some assessments and determine the extent to which card sorting results can usefully predict navigation performance. For our assessments, we work with a real-world domain to conduct a card sorting study, use its results to create an information structure and then conduct navigation tests on the structure. While we use various card sorting techniques to create the structure in our study, our analysis focuses on results from a closed card sort. These results are compared with those from the navigation tests. In addition, we explore the viability of an automated alternative to card sorting. In this case, we employ a computer analysis based on word overlap between the target item and text that represents the category label. We show how these results relate to the card sorting data and the navigation data.

DOMAIN FOR THE INFORMATION STRUCTURE

The Media Relations department at DePaul University hosts an "expert database" with descriptions of university faculty members and their respective areas of expertise as a resource for journalists or other people in need for a subject-matter expert. The database appears online as a web site. Originally derived from a book, it indexes the experts through a list of approximately 50 topics, presented alphabetically. Some of the topics have subtopics. We adapted this information to serve as the information structure for our studies. As a working web site, its content provides a realistic, non-proprietary domain for applying common design methods in order to study their effectiveness.

The expert names and their descriptions were automatically extracted from the web site. Duplicate, incomplete and erroneous entries were removed, resulting in 970 unique descriptions consisting of an expert name and a short description of his or her expertise. On average, the description of the expert consisted of 14.0 words. Experts with multiple areas of expertise were sometimes represented multiple times with different descriptions.

Three of us (two HCI graduate students and an HCI professor) each clustered 50 randomly selected expert descriptions. Through consensus we consolidated our clusters and created preliminary category names. We then asked 8 media professionals to classify 50 randomly selected descriptions using our candidate categories. Based on their feedback, we obtained nine top-level categories (TCs).

CLOSED CARD SORT

We conducted a closed card sort to assign all expert descriptions to one of the top-level categories. To facilitate data collection, we developed a web application that automates the card sort by allowing participants to press a button representing the chosen category for each content item.

While we consulted media professionals for the initial design of the navigation structure, we recruited university students for empirically assessing closed card sorting. We used this population of users so that we could use the same population for both the card sorting study and the web navigation study, which we will later present.

Participants

We recruited 15 participants via flyers and class announcements at DePaul University.

Procedure

The 970 expert descriptions were randomly divided equally into two sets. Eight of the participants were randomly assigned to one set and the remaining seven were assigned to the other set. The closed card sort was conducted using a browser-based web application which presented one randomly chosen expert description at a time and nine buttons, each labeled with one top-level category (TC). Buttons were arranged in a 3x3 grid with their order randomized for every expert description to reduce learning effects and position biases. Participants were asked to read the description and then click the TC button that provided the best fit for this expert. Once a selection was made, it could not be changed and the next random item was displayed. TC selection and reaction time were collected along with the respective item and participant ID.

Results

The response times for each category selection were skewed to the right with a median of 4.6 seconds and first and third quartiles of 2.7 and 7.9, respectively. On average, the category receiving the most selections for each classified item was selected 69.6% of the time. Alternately phrased, if the category with the most selections is considered the "correct category," participants selected the correct category 69.6% of the time, producing an error rate of 30.4%.

Discussion

The error rate of 30.4% gives some indication of the difficulty of the domain. As a point of comparison, Dumais and Landauer [9] note that some studies show error rates up to 35%-50% for selecting categories at top levels. We will present additional results and discussion of the card sort after we have described the study for assessing its results.

CREATING THE NAVIGATION STRUCTURE

Our design process used a combination of methods with varying degrees of formality (see Kuniavsky [11], which

presents card sorting using both informal and formal analyses). While we do not claim that our process produces an optimal structure, we believe that our process uses methods that are commonly practiced and adhere to realistic time constraints and participant availability. For our process, top-level organization used the results of the closed card sort performed by the participants we recruited. We created the remaining organization by clustering content ourselves.

Top-level organization

At top-level, content items (i.e. the expert descriptions) were moved to the TC that had more selections by participants than any other TC. Note that the chosen TC did not necessarily receive an outright majority of selections. We resolved ties by choosing through consensus. Table 1 shows the distribution of the content items among the nine TCs. No category had more than 20% of the content items and all categories had at least 4% of the content items.

Category	Count	Percent
Arts & Literature	168	17.32%
Business & Economics	190	19.59%
Education	74	7.63%
Health & Medicine	57	5.88%
Law & Legal Issues	94	9.69%
Politics & Public Policy	102	10.52%
Religion	45	4.64%
Science & Technology	88	9.07%
Society & Culture	152	15.67%
Total	970	100.00%

Table 1 Distribution of Content Items by Category

Sub-level organization

For each TC, we created subcategories based on open card sorts. Three of us each clustered expert descriptions belonging to each of the TCs. Through consensus, we consolidated our individual clusters to create a second tier of categories. Similarly we created additional tiers of categories for SCs whenever a category contained a substantial number of items (typically more than a dozen) and afforded some natural subcategories. We used category names from the original web site to the extent that they fit our new structure. The resulting information structure had a maximum depth level of four. That is, content items could be reached with a maximum of three category selections.

STATISTICAL MEASURE OF CATEGORY MEMBERSHIP

The closed card sort potentially predicts which top-level category a user will select when looking for a particular item. We were also interested in learning how well we could predict category selection based on a statistical measure of category membership. In our case, we consider the similarity between the words in the targeted item description and words associated with the category labels. Kaur and Hornof [10] have used semantic similarity

Task description: Manages one of the largest accountancy programs in the United States. Has served as director of auditing research for the American Institute of Certified Public Accountants and as a senior auditor for KPMG Peat Marwick.

Category	Similarity Score
Arts and Literature	0.128
Business and Economics	0.251
Education	0.079
Health and Medicine	0.082
Law and Legal Issues	0.077
Politics and Public Policy	0.153
Religion	0.063
Science and Technology	0.133
Society and Culture	0.096

Table 2. Similarity Scores for an Example Task

models, such as Latent Semantic Analysis (LSA [6]) and Pointwise Mutual Information-Information Retrieval (PMI-IR [21]), to predict which link label users select when they have a particular navigation goal. Unlike the closed card sort, an approach based on computed similarity scores does not require collecting results from human participants.

In order to explore the viability of using a statistical measure of category membership as an alternative to card sorting, we used a method based on a simple similarity metric. This method makes predictions based on the amount of common words in the targeted content item (user goal) and the words in the content items found in the structure under the top-level category. Note that this method does not quite correspond to the closed card sorting task since the human participants had to rely on the category labels rather than the text descriptions that were organized with the labels. In this way, our method uses the content descriptions as a proxy for the information in the category labels. We will later discuss practical implications of this approach.

The implementation of our method was customized from software that implements a vector space search engine [4]. Term vectors were created for each user goal and for each category. The term vector for each user goal consists of the word frequencies in each task description (i.e. the text of the targeted content item). The term vector for each category consists of the word frequencies for the text of all of the content items organized within the respective category. We excluded 68 stop words such as articles (e.g. "a", "an", "the") and common prepositions (e.g. "in" and "to") from the word list. The porter stemming algorithm [16] was applied to reduce each word to a root form, thus allowing words such as "audit" and "auditing" to count as the same word. Similarity between the task description

vectors and category vectors was calculated using cosine similarity, which is the cosine of the angle between the two vectors. Texts with no words in common have a cosine of zero and texts with the same word frequencies will have a cosine of one. This measure of similarity is a popular method for comparing the documents and queries of information retrieval tasks [22].

Here we present an example. Table 2 presents a user goal specified as a task description. This task description is one of the content items in the information structure. Cosine similarity scores are generated by how well the combined task descriptions in each of the top-level categories match this task description. The table shows the cosine similarity scores for each top-level category in the information structure we created. The "Business and Economics" category has the highest score, presumably because words in its task descriptions share the greatest number of words with that in the task description serving as the user goal. In this example, the model predicts that users are likely to select this category.

Note that our measure of category membership uses a similarity score based on frequencies of the same words. Different words with similar meanings do not contribute to the similarity score. This simple approach contrasts with potentially more powerful approaches like LSA and PMI-IR, which make use of training documents to learn similarities between different words as determined by the extent to which words co-occur in the training documents.

WEB NAVIGATION STUDY

We are interested in how well results from closed card sorting and similarity metrics assess the categories and their structure used in a web site. Here we present a study where participants performed navigation tasks in a web site constructed from the categories and structure described above. We will principally examine navigation performances and compare them to predictions obtained from the card sorting and similarity metrics. The closed card sort indicates problematic tasks if the item's categorization in the web site does not match the category where card sorting participants generally placed the item. Our similarity-based method indicates problematic tasks if the item's categorization in the web site does not match the category receiving the highest similarity score.

For our navigation study, we wanted good estimates of mean navigation times for unpracticed users in order to evaluate the web structure while minimizing learning effects. To this end, we created simple layouts where the presentation order of the links was randomized with each task. By restricting browser controls to link selection and pressing the back button, participants were required to find items using the most common actions for web navigation [3]. While not the focus of this article, top level categories were presented with and without a few exemplar subcategories for two different conditions. Additionally, open-ended scenario tasks were also presented to participants but will not be used in the analysis here. Nevertheless we present the experimental design in its entirety here.

Participants

We recruited 35 participants via flyers and class announcements at DePaul University.

Instrumentation

A web application was programmed to present participants with a sequence of navigation tasks. The application dynamically generated web pages that allowed the participant to navigate through the web structure. The browser-based application featured a back button which returned to the previously displayed page. Advanced browser features found in the toolbar or the browser menus (e.g. keyboard shortcuts, keyword search, context menu, etc.) were hidden during the experiment and users were instructed not to use these features to access browser navigation functions. The web application recorded and time-stamped every selection performed by the participant.

Tasks

Each participant performed 16 tasks consisting of two types:

- "Exact description tasks": Participants were asked to find a predetermined target within the structure. Example: "Find the expert with this description: Women's history, especially in the United States, race, ethnicity and immigration." If a wrong target was selected, the system would prompt the user to continue the search. If the right target was located, the system would move on to the next task.
- "Scenario tasks": Participants were given a scenario and asked to identify an appropriate expert for this problem within the structure. Example: "Find an appropriate expert for this scenario: A study is released about how working parents affect the psychological outlook and the educational achievement of their kids. Find an expert to analyze it." Once a target item was selected, the system moved on to the next task.

All participants received the same 6 scenario tasks. However, only 2 exact description targets were pre-selected and included in every user's set. Another 8 target descriptions were randomly selected for each user from the entire pool of a total 970 content items (i.e., expert descriptions). The order of the tasks was randomized for each participant.

Display

The start page (level 1) consisted of categories only; on lower levels (2, 3 and 4), pages could contain expert descriptions as well as subcategories (SCs). For half of the participants (randomly designated), the start page consisted of only top-level categories (TCs). For the other half, their start page consisted of the TCs, each with a few exemplar

SCs listed below them. The exemplar subcategories were chosen as those with the highest number of expert descriptions. We call the condition without subcategories the **Simple** display and the condition with subcategories the **Exemplar** condition. Pages on deeper levels were identical in both conditions. The order of the categories was randomized with each task. The order of the expert descriptions, when they appeared, was also randomized. At the top of each page, the current task goal was permanently present for review. Also presented was the category path to the current page (often called "breadcrumbs").

Procedure

We provided online instructions prior to the session. Experimenters also encouraged the participants to ask questions to ensure they understood the task completely. Upon the start of the experiment, the first target was presented along with a "continue" button which – when pressed – displayed level 1 of the category structure and started the timer. Upon completion of a task (i.e. finding the target) or after four minutes, the application presented the next task.

ANALYSIS AND RESULTS

Data Preprocessing

We restrict our analysis to the exact description tasks, which have a clear criterion for correctness. During the study, we observed that a few participants had difficulty understanding the exact description task. In particular, they selected expert descriptions even though they did not match the task description. Participants who frequently selected non-matching descriptions were identified as those whose mean selection count exceeded 1.5 times the inter-quartile range (IQR) of mean selections. Observations from these participants were removed from the data set. In addition, individual tasks exceeding 1.5 times the IQR of description selection counts for all individual exact-match tasks were also removed from the analysis. After preprocessing, the data consisted of 271 tasks from 30 participants.

Tasks that exceeded the time limit of 240 seconds (4 minutes) were coded with a completion time of 240 seconds. Based on a two-tailed t-test, it was found that the mean navigation time for participant means in the Exemplar start page condition (M=78.1s, SD=20.4s) was not significantly different from that in the Simple start page condition (M=70.1, SD= 24.4s), $t(28)= 0.89$, $p=0.38$. The remaining analyses combine the navigation times of both conditions.

Below we report the effects of task factors indicated by card sorting, statistical category membership and the initial selection in the web navigation study. Since tasks were blocked by participant, we performed mixed model analyses where the participants were modeled as a random effect [19]. We use this model to test the significance of the task factors and estimate the standard error for the mean navigation times.

Card sort predictions

We identified expert descriptions that were placed in a particular category a majority of the time during the closed card sort. Expert descriptions with majority category selections matched the actual category in the navigation structure for 82% of the navigation tasks (N=223). The mean navigation time for these cases was 69.3 seconds (SD=71.3, SE for the mean = 4.9). For task descriptions without majority selections or where the majority selection did not match the location in the web structure (N=48), the mean navigation time was 100.5 seconds (SD=79.5, SE=10.5). The difference between these two cases as predicted by card sorting is significant, $F(1, 240) = 7.27$, $p=0.0075$. The first pair of bars in Figure 1 shows the mean navigation times with standard error bars for the conditions indicated by card sorting.

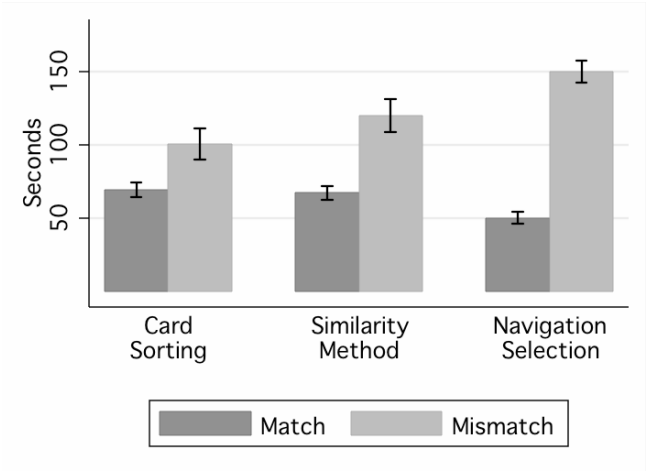

Figure 1: Mean Navigation Times by Indicator

Similarity measure predictions

We identified the categories with the highest cosine similarity scores for each expert description. The categories with the highest values matched the actual categories in the navigation structure for 86% of the navigation tasks (N=232). The mean navigation time for these cases was 67.2 seconds (SD=68.6, SE=4.7). For the tasks where the category with the highest similarity score did not match the actual category in the navigation structure (N=39), the mean navigation time was 119.8 seconds (SD=86.5, SE=13.8). The difference between these two cases as predicted by this method of using similarity scores is significant, $F(1, 240) = 18.07$, $p<0.0001$. The second pair of bars in Figure 1 shows the mean navigation times with standard error bars for the conditions indicated by the similarity measure.

Effect of selecting the wrong category

We identified the first categories that users actually selected during the web navigation study. Participants initially selected the correct category at the top level for 75% of their tasks (N=204). In these cases, the mean navigation time was 50.2 seconds (SD=51.5, SE=4.1). For initially

selecting the incorrect category (N=67), the mean time was 149.8 seconds (SD=80.1, SE=7.3). The difference between the means of these two navigation situations is significant, $F(1,240)=139.8$, $p<0.0001$. The third pair of bars in Figure 1 shows the mean navigation times with standard error bars for cases when the user initially selects the correct category (match) and when the user initially selects an incorrect category (mismatch).

Predicting category selection during navigation

For card sorting, we considered the category receiving *the most* selections by participants even if no category received a majority of selections. Ties for most selections were chosen at random. These selected categories predicted the selected category during web navigation for 76% of the tasks (N=206). For the automated similarity method, the category with the highest similarity score predicted the selected category during web navigation for 73% of the tasks (N=198).

DISCUSSION

Both card sorting and our similarity method identified problematic tasks that were significantly more difficult as indicated by navigation times. The similarity method performed slightly better in distinguishing between easy and difficult tasks, where the mean navigation time of the difficult tasks is nearly twice as large as that of the easy tasks. On the other hand, card sorting performed slightly better in predicting which categories participants would select in the navigation task.

Our assessment is based on the methods' ability to identify problematic tasks. This ability has several immediate uses for information architects. If the number of problematic tasks is small, placing their respective content items in multiple places in the structure is one immediate solution for improving users' ability to find the items. Both card sorting and statistical category membership provide guidance for determining the second location. Also, tasks are associated with particular navigation pages and category labels. The navigation pages can then be flagged for further study, possibly involving limited usability tests if resources allow it. Category labels associated with problematic tasks can be reworded for greater clarity, ideally improving coherence within the categories.

While we employed informal card sorting methods to create subcategories and labels, our formal analysis focuses on categories and labels at the top level. With nearly one thousand content items and over a hundred subcategories at various levels, the time needed for exhaustive card sorting would have more than doubled. Since practitioners often have limited time and participant availability, they must selectively choose which parts of a structure they will design using formal methods. The navigation results support our choice of focusing on top-level categories. When participants selected the wrong category at the top level, they took nearly three times as long on average to find the item. Selecting the wrong top-level category incurs substantial cost in navigation time as users may need to backtrack multiple levels before finding the correct path from the top page. Even if the design process permits extensive data collection, there is also the possibility that formal empirical results on subcategories would become obsolete if extensive reorganization at the super-ordinate level is required. With these considerations, a reasonable strategy for information architects is to focus on the creation and evaluation of the top-level categories and their labels. Later in the design process, organization of subcategories can be informally assessed and revised, usually without requiring extensive redesign and without major consequences to the overall effectiveness of the navigation structure.

Another issue for practitioners is how our results might compare to user performance for their navigation structures. The first point of consideration is the size of the information structure. The design for our study consisted of 970 content items placed at depths of two to four levels within the navigation structure. When designing structures of comparable size, our results indicate what practitioners can expect when selection errors occur at the top level. For larger structures, selection errors may have greater consequences if users cannot quickly determine when they have made the wrong selection. For smaller structures, the navigation cost of selecting the wrong category is limited. For these cases, it is possible that top level performance is a less important indicator of navigation times. Finally, practitioners should be mindful that our results pertain to unpracticed usage of a particular information structure. We deliberately designed our study to minimize learning effects across tasks. In contrast, with repeated usage of a particular web site or menu system, users can learn the categories and memorize the location of many content items. The performance of practiced users is thus less dependent on the semantics of the categorical labels. For these cases, engineering models of expert usage, such as the GOMS Keystroke-Level Model [4], would be more suitable for predicting human performance.

Both approaches identify problematic tasks by predicting when users are likely to select the wrong top-level category. As a point of comparison, participants took nearly three times as long to finish a task when they actually selected the wrong top-level category during the navigation task. This difference suggests a best-case performance if a method could perfectly predict the categories users select at the top level. Below we review shortcomings of the two methods we analyzed here.

The closed card sorting task is almost identical to the web navigation task up to the point where the participant selects a category. In this sense, we would expect the card sort to be a strong predictor of category selection during navigation. Sources of error include sampling error, especially with a small number of card sorting participants (a set of content items sorted by no more than 8

participants). Also, category buttons for the card sort were randomized. It is highly unlikely that the categories appeared in the same order for both card sorting task and the web navigation task. Finally, our identification procedure only considers the majority category or the category with the most selections. We have not considered the proportion of selected categories, which might be useful in predicting the proportions during the navigation task.

While open card sorting is useful for creating an initial set of categories and suggesting labels for them, we have not formally assessed its utility here. It is less clear how results from open card sorting can predict problematic tasks for navigation. For open card sorting, participants examine the content items when deciding how to place them together. Since content items are not visible during navigation tasks, users must fully rely on the category labels to select a menu category. Additional studies are needed to assess the utility of open card sorting for designing information structures and for predicting performance on navigation tasks.

Like open card sorting, our statistical measure of category membership uses information not immediately available to users performing a navigation task. Rather than comparing the navigation goal to each category label, our method compares the goal's text to all of the content text classified with the label. Its effectiveness for predicting category selection thus depends on how well the content beyond the label represents a user's understanding of the label itself. For our Experts web structure, we employed common design methods to produce relatively coherent categories with meaningful labels. It is possible that our similarity method would have performed much worse on a poorly designed structure with highly misleading labels. Ideally, the similarity measure would operate on the navigation goal and the category labels. While labels do not offer enough text to support our simple measure, more powerful similarity measures would provide more meaning ratings of label relevance. For example, CoLiDeS uses LSA to compare the navigation goal to the category [2]. Blackmon et al [1] report studies with CoLiDeS that demonstrate its effectiveness in identifying website navigation problems. Other approaches to automated methods for identifying navigation problems include systems that employ strategies that simulate user navigation [16,7,14,12]. Combining similarity measures such as LSA and PMI-IR, they show likely navigation paths users and thus identify targets that users would have difficulty reaching. While these approaches generally assess information scent based on labels, Chi et al. [6] report one version of a navigation model that calculates category relevance (information scent) based on content lying beyond the local labels. Their calculation of "distal scent" is similar to the approach taken here.

CONCLUSION

While further development of automated methods show promise, we have chosen to focus on simple methods based on card-sorting and a common measure of similarity. Despite their simplicity, our results indicate that these simple approaches can effectively identify problematic navigation tasks within a realistic information structure. Because of their simplicity, the methods present little difficulty in practicing them. In the case of card sorting, index cards and simple statistical calculations are all that is needed. The method based comparing content items using cosine similarity requires some programming but uses routines freely and publicly available. In addition, we have emphasized methods that are practiced early in the design process, before any visual design and implementation. This emphasis contrasts with the operation of most automated methods, which typically run on working web sites. While there is a place for evaluating navigation on functioning systems, early assessment of labels and structure allows for feedback before putting much effort into implementation. Finally, by assessing simple methods and reporting their performance, we have provided some minimal expectations for more sophisticated approaches to automatically assessing navigation.

ACKNOWLEDGMENTS

We will add acknowledgements here later.

REFERENCES

1. Blackmon, M. H., Kitajima, M., & Polson, P. G. Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. In *Proc. CHI 2005*. ACM press (2005), 31-40.
2. Blackmon, M. H., Polson, P. G., Kitajima, M., and Lewis, C. Cognitive walkthrough for the web. In *Proc. of CHI 2002*. ACM press (2002), 463-470.
3. Byrne, M. D., John, B. E., Wehrle, N. S., and Crow, D. C. (1999). The tangled web we wove: A taskonomy of WWW use. In *Proc. CHI'99*. ACM press (1999), 183-190.
4. Card, S. K., Moran, T. P., and Newell, A. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1983.
5. Ceglowski, M. Building a Vector Space Search Engine in Perl. http://www.perl.com/lpt/a/2003/02/19/engine.html.
6. Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. Using information scent to model user information needs and actions on the web. In Proc. CHI 2001. ACM Press (2001), 490-497.
7. Chi, E. H. et al. The Bloodhound project: Automating discovery of web usability issues using the InfoScent simulator. In *Proc. CHI 2003*. ACM Press (2003), 505-512.
8. Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S. and Harshman, R. Using latent semantic analysis to improve access to textual information, *Proc. CHI'88*, ACM Press (1988), 281-285.

9. Dumais, S. T. and Landauer, T. K. Using examples to describe categories. In *Proc. CHI '83*, ACM Press (1983), 112-115.
10. Kaur, I and Hornof, A. J. A comparison of LSA, wordNet and PMI-IR for predicting user click behavior. *Proc. CHI 2005*. ACM Press (2005), 51-60.
11. Kuniavsky, M. *Observing the User Experience: A Practitioner's Guide to User Research*. Morgan Kaufmann, San Francisco, 2003.
12. Juvina, I, and van Oostendorp, H. Bringing Cognitive Models into the Domain of Web Accessibility. In *Human-Computer Interaction International 2005*, 2005.
13. Marchionini, G. Information-seeking strategies for novices using a full text electronic encyclopedia. *Journal of American Society for Information Science*, *40*, 1 (1989), 54-66.
14. Miller, C. S., & Remington, R. W. Modeling information navigation: Implications for information architecture. *Human-Computer Interaction*, *19*, 3 (2004), 225-271.
15. Olston, C. and Chi, E.H.. ScentTrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction (TOCHI) 10*, 3 (2003), 177-197.
16. Pirolli, P., Fu, W., Chi, E., and Farahat, A. Information scent and web navigation: Theory, models and automated usability evaluation. In *Human-Computer Interaction International 2005*, 2005.
17. Porter, M.F.. An algorithm for suffix stripping, *Program*, *14*, 3 (1980), 130-137.
18. Rosenfeld, L. & Morville, P. *Information Architecture for the World Wide Web*. O'Reilly & Associates, Sebestapol, CA, 2002.
19. Singer, J. D. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics 24*, 4 (1998), 323-355.
20. Spool, J.M.. Are There Users Who Always Search? http://www.uie.com/articles/always_search/
21. Turney, P. D. Mining the web for synonyms: PMI-IR versus LSA on TOEFL, *Proc. 12th European Conference on Machine Learning*, (2001), 491-502.
22. van Rijsbergen, C. J. *Information Retrieval*. Butterworths. London, 1979. http://www.dcs.gla.ac.uk/Keith/Preface.html.