

Using meaningful and stable categories to support exploratory web search: Two formative studies

Bill Kules and Ben Shneiderman

Department of Computer Science, Human-Computer Interaction Laboratory,
and Institute for Advanced Computer Studies

University of Maryland at College Park, College Park, MD 20742

{wmk,ben}@cs.umd.edu

Please send correspondence to Bill Kules (wmk@cs.umd.edu)

(301) 891-7271 (Kules)

(301) 405-2680 (Shneiderman)

(301) 405-6707 (fax)

ABSTRACT

Categorizing web search results into comprehensible visual displays using meaningful and stable classifications can support user exploration, understanding, and discovery. We report on two formative studies in the domain of U.S. government web search that investigated how searchers use categorized overviews of search results for complex, exploratory search tasks. The first study compared two overview conditions vs. a control (N=18). The overviews were based on the federal government organizational hierarchy. With the overview conditions, participants noticed missing results more often than participants in the control. They also found pages of interest

deeper within the results. The overview conditions received significantly higher subjective ratings. The second study compared an overview based on automated clustering vs. the government hierarchy overview (N=12), and the results suggest that domain knowledge and task influenced the preferred overview. The studies lend support to the use of compact overviews based on meaningful and stable categories tightly coupled with ranked result lists.

KEYWORDS

Web search; exploratory search; search user interfaces; categorization; categorized search results; search result visualization; information seeking; information access; information retrieval; graphical user interfaces; human-computer interaction.

1. INTRODUCTION

The World Wide Web creates tantalizing opportunities for learning and research. Every day, teachers, journalists, researchers and ordinary citizens search the web as they attempt to find, organize, understand, and ultimately learn from information on the web. Although search engines generate long lists of relevant results, the lack of effective overviews challenges users who seek to understand these results, especially for exploratory search tasks such as learning about a new topic, which require gaining overviews of and exploring large sets of search results, understanding document context and identifying unusual features (White, Kules, Drucker, & schraefel, 2006). These users struggle with information overload, coping with an overabundance of information that lacks a comprehensible organization.

This is particularly problematic when users undertake exploratory searches to satisfy information needs that are imprecise or evolving or when their domain knowledge is limited. Alternatively, they may have a clear question or goal, but are uncertain how to gather information to satisfy the goal. Incompletely formulated queries yield a plethora of potentially relevant search results, which must be examined and understood. The problem is exacerbated by the frequency of

short queries (Spink, Wolfram, Jansen, & Saracevic, 2001). Analysis of search goals suggest that between 20-30% of all web queries may be exploratory in nature (Rose & Levinson, 2004), which motivates study of this type of search.

Categorizing web search results into comprehensible visual displays using meaningful and stable classifications can support user exploration, understanding of large result sets, and discovery. Research prototypes and commercial search engines have incorporated category information, but (as discussed in our Related Work section) there have been few user studies of categorized overviews for exploratory web search, and there is little research explaining whether, why and under what circumstances they are effective. Research is needed to justify the entry and maintenance of category metadata and to guide the design of search engine interfaces.

This paper reports on two formative studies conducted on web search within U.S. government web sites. The purpose of both studies was to illuminate searchers' use of categorized overviews to explore and understand search results. The research goals motivating these two studies include:

- Identifying search tasks that benefit from categorized search result overviews
- Understanding how the visual presentation of the overview affects its utility
- Understanding how the classification (i.e. the set of categories) used for the overview affect its utility and the user's search experience

In study 1, we compared three presentations of results categorized into a 2-level government hierarchy. Two overview+detail interfaces (an expandable outliner and a treemap) allowed users to narrow the search results by categories, and a third interface (the control) provided a typical set of results with category information displayed below each result. In study 2 we investigated the affect of alternate classifications, one based on the government organizational hierarchy and the other based on Vivisimo's automated clustering. The information seeking tasks

used in the studies were motivated by our work with government agencies and our understanding of the challenge of finding government information and related publications. In this domain, web sites such as FirstGov (www.firstgov.gov), FedStats (www.fedstats.gov) and other specialized search engines provide some help for searchers. To our knowledge, no search engines currently provide overviews of search results categorized by government agency, even though studies have found that queries for governmental information comprised 1.5%-3.0% of all queries to general web search engines (Jansen, Spink, & Pedersen, 2005; Spink & Jansen, 2004).

These studies were conducted as part of a research program that is identifying design principles and developing prototypes for the visual display of and interaction with categorized search results. The study interfaces (except for Vivisimo) were developed in accord with six emerging principles (Kules & Shneiderman, in process), that draw on the fields of information science, information retrieval, human-computer interaction and information visualization:

- Provide overviews of large sets of results
- Organize results by meaningful classifications
- Tightly couple category labels to results list
- Arrange text for scanning/skimming
- Visually encode quantitative attributes on a stable visual substrate
- Support multiple visual presentations and classifications

The next section briefly discusses previous studies of categorized search results. Sections 3 and 4 describe the two formative studies, and section 5 discusses the study results. The paper concludes with a summary of our contributions and areas for future work.

2. RELATED WORK

For exploratory searchers, classifications, taxonomies and other knowledge structures support information organization and retrieval, provide semantic roadmaps to fields of knowledge, and improve learning (Soergel, 1999). There is growing use of thesauri on the web to support information retrieval (Shiri & Revie, 2000). Web directories such as Yahoo! (www.yahoo.com) and the Open Directory Project (www.dmoz.org) (DMOZ) catalog a small but important fraction of the Web, providing an overview of general Web content and enabling users to find information by browsing a familiar subject hierarchy. These knowledge structures can be used to categorize search results for presentation. The following sections briefly discuss studies of categorized search results for web and non-web search applications.

2.1. Studies of categorized search results for web search

Meaningful and stable categories have been found beneficial for presentation of web search results in the few studies conducted. Grouping search results by a two-level subject classification expedited document retrieval for informational tasks with a single correct answer (Dumais, Cutrell, & Chen, 2001). For question answering tasks, search results augmented with category labels produced the fastest performance and were preferred over results without category labels (Drori & Alon, 2003). The Cha-Cha system organized intranet search results by an automatically generated web site overview. Preliminary evaluations were mixed, but promising, particularly for what users considered “hard-to-find information” (Chen, Hearst, Hong, & Lin, 1999). The WebTOC system provides a table of contents visualization that supports search within a web site, although no evaluation of its search capability have been reported (Nation, Plaisant, Marchionini, & Komlodi, 1997).

Clustering web search results into dynamic categories, in which documents are grouped by similarity measures rather than explicit categorical attributes, has been investigated as an alternative to classification, and has been shown to improve on ranked lists for information

retrieval metrics such as precision and recall (Hearst & Pedersen, 1996; Käki, 2005; Marshall, McDonald, Chen, & Chung, 2004; Zamir & Etzioni, 1999; Zeng, He, Chen, Ma, & Ma, 2004) or task completion time (Turetken & Sharda, 2005). Chen, Houston, Sewell, & Schatz (1998) found that recall improved when searchers were allowed to augment their queries with terms from an thesaurus generated via a clustering-based algorithm. A one-level clustered overview was found helpful when the search engine failed to place desirable web pages high in the ranked results, possibly due to imprecise queries (Käki, 2005). The benefits of clustering include domain independence, scalability, and the potential to capture meaningful themes within a set of documents, although results can be highly variable (Hearst, 1999). Generating meaningful groups and effective labels is a recognized problem (Rivadeneira & Bederson, 2003).

2.2. Other studies of categorized search results

The Flamenco system (Hearst et al., 2002; Yee, Swearingen, Li, & Hearst, 2003) provided interfaces to specialized collections (art, architecture and tobacco documents), using faceted hierarchies to produce menus of choices for navigational searching. A usability study compared the interface to a keyword-based search interface for an art and architecture database for structured and open-ended, exploratory tasks (Yee et al., 2003). With Flamenco, users were more successful at finding relevant images (for the structured tasks) and reported higher subjective measures (for both the structured and exploratory tasks). The exploratory tasks were evaluated using subjective measures, because there was no (single) correct answer and the goal was not necessarily to optimize a quantitative measure such as task duration. The Dyna-Cat system organized medical search results by a taxonomy of question types (Pratt, Hearst, & Fagan, 1999). In a comparison with clustering and ranked list interfaces, Dyna-Cat helped searchers find more answers to general fact-finding questions within a fixed time. Searchers also felt that they learned more using Dyna-Cat. The SuperBook interface organized search results within a book according to the text's table of contents, expediting searches without loss of accuracy (Egan et al.,

1989). The GRiDL prototype displays search result overviews in a matrix using two hierarchical categories (Shneiderman, Feldman, Rose, & Grau, 2000). The List and Matrix Browsers provide similar functionality (Kunz, 2003). Informal evaluations of these two interfaces have been promising, although no extensive studies of the techniques have been published.

2.3. Summary

Few user studies have examined the use of meaningful and stable categories specifically for organizing web search results. User studies have investigated meaningful and stable categories for organizing database search results, and studies have been conducting using automated clustering of web search results to generate dynamic categories. Most studies have focused on non-exploratory tasks. This leaves a gap in the research, which these studies begin to address.

3. STUDY 1: EXPANDABLE OUTLINER VS. TREEMAP VS. CONTROL

3.1. Research Questions

This study investigated the first two research goals listed in the introduction: What tasks benefit from categorized overviews and the effect of the visual presentation of the overview. For the visual presentation of results, an overview+detail approach was consistent with our initial principles. We identified three tasks that we believe are common in exploratory search: finding groupings of information (based on departments and agencies) that have large numbers of results, identifying different aspects of or perspectives of a query topic, and identifying unusual results.

This study addresses three research questions:

- 1) Can an overview+detail display of search results based on a government hierarchy improve exploratory search success over the typical ranked list?
- 2) Can a graphical overview improve on a non-graphical overview?
- 3) What patterns of usage does the overview+detail approach induce?

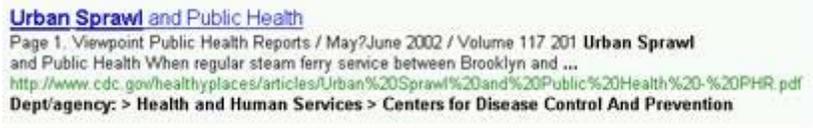


Figure 1. The control condition mimics a typical set of Google search results, adding the government department and agency.



Figure 2. Detail of the expandable outliner condition. The top 200 urban sprawl results have been categorized into a two-level government hierarchy, which is used to present a categorized overview on the left. The Interior Department, which has 20 results, has been expanded and the National Park Service has been selected. The effect on the right side is to show just the three results from the Park Service.



Figure 3. Detail of the treemap condition, which used nesting to show both top and second-level categories simultaneously. The set of results and the selected agency (NPS) is the same as in figure 2.

3.2. Experimental Conditions

This study compared presentations of search results with and without categorized overviews. We used the U.S. federal government organizational hierarchy as a meaningful and stable structure to categorize search results. Results were categorized into the leaf nodes of a broad, shallow, 2-level government agency hierarchy by matching the URLs to a database of federal government web sites. Although strictly a tree and not a true hierarchy (Kwasnik, 1999), it has many benefits: It is reasonably complete and comprehensive, the categorization rules are systematic and predictable; and a given result will (with very few exceptions) be found in a single category (mutual exclusivity).

The study used a 1x3 between groups design (N=18, 3 groups of 6), with interface type as the independent variable. The control condition (Figure 1) displayed search results in a manner similar to Google, adding the government department and agency, but it provided no categorized overview. Two experimental conditions used overview+detail interfaces (an expandable outliner or a treemap, figures 4 and 5 respectively) allowing participants to limit the displayed list of results by selecting (clicking on) a single category. The overview conditions allowed participants to show or hide empty categories, and the expandable outliner additionally allowed participants to display or hide the counts of results in parentheses after each category. Both quantitative and qualitative data were collected. Preliminary results were reported in Kules & Shneiderman (2004).

3.3. Hypotheses

In addition to collecting qualitative data, this study tested three hypotheses, based on the initial search result visualization principles:

1. Overview conditions will yield higher successful completion rates within a fixed time.
2. Overview conditions will be rated more favorably than the control.
3. Overview conditions (and particularly the treemap) will be judged as more complex than the control and more difficult to learn.
4. The results identified by participants using the overview conditions for exploratory tasks (B and C, below) will be more unique.

3.4. Scenario and task design

Scenarios and tasks were carefully constructed to provide a realistic exploratory search context, while constraining the search task to the examination of a constant (across participants) set of search results. We also desired to control – to the extent possible – for differences in interpretation of the exploratory search tasks (Järvelin & Ingwersen, 2004). Examining search results is a necessary step within a larger information seeking process, the objective of which is to satisfy a perceived information need or problem (Marchionini, 1995). In turn, the perceived information need is situated within a higher level social, cultural and organizational context and motivated by a higher-level work (or pleasure) objective (Byström & Hansen, 2002; Järvelin & Ingwersen, 2004). For these reasons, the task design for these studies considered multiple levels of context. Byström and Hansen (2002) proposed a three-level abstraction for task context which was adapted as a frame for these two studies.

The highest level of Byström and Hansen’s taxonomy is the work task. Work tasks are situated in the work organization and reflect the organizational cultural and norms, as well as organizational resources and constraints. In these two studies, the scenarios described a simulated work task, as advocated in Borlund (2003), which provided the “cover story” that encouraged participants to bring their own knowledge and experience (however limited) to the subsequent tasks. The scenarios provided a second level of context, the information seeking context, by

locating the searcher within the initial stages of an exploratory search task, equivalent to the pre-focus exploration stage of Kuhlthau's (1991) six stages or the pre-focus stage of Vakkari (2001). The scenarios described the participant (information searcher) as being at a "starting point" or "exploring topics and defining your paper's thesis." Within this stage, the third level of context was the information retrieval context, which placed the participants in the Examine Results stage of an information seeking session by indicating that they had just entered a pre-specified query. This allowed us to use a consistent set of search results across all participants.

The scenarios thus attempted to provide a set of situational and contextual cues to induce a realistic information need within each participant. Due to practical limitations on the software (search results had to be pre-processed), and the duration of experimental sessions, it was not practical to use real-life, participant-provided search tasks as recommended by Borlund (2003). Because these were formative studies, we chose to expose participants to three diverse scenarios, rather than a tailored scenario advocated by Borlund. These are, of course, imperfect and incomplete mechanisms, but have repeatedly proven to be effective in guiding our research.

The scenario content was motivated by our work on the challenges of finding government information and publications (Ceaparu & Shneiderman, 2004; Kules & Shneiderman, 2003; Marchionini, Plaisant, & Komlodi, 1998). Our work with statistical agencies generated 15 prototype scenarios (Ceaparu & Shneiderman, 2004). Many of these involved some aspect of learning about a general topic such as breast cancer, Alzheimer's disease, or soybean production. The statistical information seeking scenarios were readily generalized to the full government domain for these studies, with details such as age and location included to provide a plausible description.

Each scenario introduced a pre-specified query and a set of 200 search results for the queries "breast cancer", "alternative energy" and "urban sprawl":

Scenario 1 (Urban sprawl) - *Imagine that you are a 40-year old social activist in a rural town near the Washington, DC metropolitan area and have become increasingly concerned about the impact of urban sprawl on your town. You are planning to write a letter to your neighbors about the issue, and you would like to learn more about it. You are using the Web as a starting point, because you are not located near a major library. You are first interested in federal government information, and later you'll look at state and local information. You have just entered the search terms "urban sprawl" into a new search engine for government web sites.*

Scenario 2 (Breast cancer) - *You are a 30-year old journalist writing an article on breast cancer and what the federal government is doing about it. You are exploring the topic, starting by looking on the Web to find out what kind of information is available. You have just entered the search terms "breast cancer."*

Scenario 3 (Alternative energy) - *You are taking an undergraduate class in environment sciences, and preparing to write a term paper on government involvement in alternative energy technologies. Your first step is to get an overview from the web of the information available to identify potential topics. You have just entered the search terms "alternative energy."*

For each scenario the three tasks were described to the participants as:

Task A (Overview) - *Your first step is to get an overview of which federal agencies (the 2nd level organizations) have substantial amounts of information on this topic. This will help you decide where to focus your research efforts. What 3 agencies publish the most information about this topic? (Time limit: 3-4 minutes)*

Task B (Finding perspectives) - *The web contains a variety of sources, perspectives and viewpoints on almost any given topic, and this is true within the federal government. Find*

3 web pages providing different aspects of or perspectives on this topic. (Time limit: 3-4 minutes)

Task C (Finding unusual results) - *Spend a couple more minutes exploring these results. Do you notice any results that, at first glance, appear to be unusual, unexpected or surprising? If so, explain why they are unusual. (Time limit: 2-3 minutes)*

The tasks were time-limited to permit completion of the session within approximately one hour.

3.5. Materials and Procedure

After the participants signed an informed consent form, they completed a short demographic questionnaire, providing their age, gender, occupation, knowledge of federal government organization, web experience, search experience and search frequency. They were asked to talk-aloud (Ericsson & Simon, 1984) and ask questions throughout the session. Training was provided for the interface to be used, and they were encouraged to use it with sample search results (from the query “soybeans”) until they were comfortable. They were instructed to view just the results and categorized overview (when available). After participants were comfortable with the interface, the first scenario was presented, and they were asked to perform the three tasks. The tasks were presented in an order searchers would commonly follow in the exploratory search scenario. That is, they would start by seeking an overview of the results, then explore, and finally integrate and reflect on their findings, possibly identifying unusual results or yielding other insights. Following these tasks, each participant was asked for subjective ratings of the interface and an informal interview was conducted to elicit comments. These steps were repeated for the remaining two scenarios. The total session time was approximately one hour. The procedures and materials were pilot tested with four participants to refine scenarios, tasks and measures. The task time limits were adjusted to keep the sessions within the one-hour target while giving participants enough time to at least get a good start on each task.

3.6. Participants

Eighteen participants (11 male, 7 female) were recruited from university and professional contacts. They ranged in age from 22 to 54, with the average age being 35. Seven were students. A heterogeneous group was appropriate due to the explorative nature of the study. All reported some familiarity with the federal government. All had at least one year of experience with web search and reported searching at least once a week.

3.7. Results

A one-way analysis of variance (ANOVA) for 10 measures was performed using SPSS or Excel. The measures were a correctness score on task A plus nine subjective satisfaction measures. When the ANOVA indicated significant differences, post hoc analysis was performed using a Tukey test. For the perspectives task, the position of selected pages was measured, as well as the number of pages selected beyond the top 10. For the unusual results, the number of unusual results identified was measured. We also qualitatively analyzed the perspectives and unusual items identified. In addition, the comments of participants and the observer's notes were reviewed.

3.7.1. Correctness score

In task A participants were asked to find the three agencies that provided the most pages within the provided results. When several agencies were tied for third place, any of them were considered correct. The scores for all three scenarios were summed, yielding a total score in the range 0-9. Rank order was not evaluated for correctness. The ANOVA showed significant differences, $f(2, 15) = 6.74, p = 0.008$. Post hoc analysis showed significant differences between the control and expandable outliner and between the control and treemap, but not between the expandable outliner and treemap (Table 1). These results support our conjecture that meaningful categorical grouping benefit users.

Table 1. Mean correctness scores for each interface, with standard deviation in parentheses.

	Control	Expandable Outliner	Treemap
Correctness score	6.50 (1.38)	8.33 (1.21)	8.67 (0.52)

3.7.2. Perspectives found

The perspectives task required participants to identify three different perspectives on or aspects of the topic. We measured task completion rates, position of pages found and number of pages found beyond the top 10. The perspectives reported by participants are listed in the Appendix.

Task completion – With two exceptions, all participants completed all tasks. One member of the control group provided only one perspective for the Urban Sprawl scenario, and one member of the Expandable Outliner group provided only two perspectives for the Breast Cancer scenario.

Position of perspectives found –For each scenario, the median position of the identified perspectives was computed (Table 2), as well as the fraction and percent of perspectives that were identified from beyond the top 10 results (Table 3). The ANOVA showed significant differences, $f(2, 146) = 17.10, p \ll 0.01$. Post hoc analysis showed significant differences between the control and expandable outliner and between the control and treemap, but not between the expandable outliner and treemap.

Table 2. Median position of identified perspective, with standard deviation in parentheses

	Control	Expandable Outliner	Treemap
Position of identified perspective	4 (9.79)	38 (55.77)	18 (56.85)

Table 3. The fraction and percent of perspectives which were found beyond the top 10 results.

Scenario	Control	Expandable Outliner	Treemap	Over all conditions
Urban Sprawl	8/16 (50%)	10/18 (56%)	6/18 (33%)	24/52 (46%)
Breast Cancer	10/18 (56%)	10/17 (59%)	8/18 (44%)	28/53 (53%)
Alternative Energy	7/18 (39%)	14/18 (78%)	16/18 (89%)	37/54 (69%)
Over all scenarios	25/52 (48%)	34/53 (64%)	30/54 (56%)	

Category use – For the overview conditions, we computed the mean number of categories selected during the task (Table 4). Note that no top-level categories were selected within the treemap. We can conjecture two explanations for this. First, users may have preferred the specificity of the second-level categories (agencies) rather than the top-level (departments). The nature of the treemap layout, however, suggests another explanation. The top level categories are selected by clicking on narrow rectangles containing the labels, whereas the second-level categories are selected by clicking on the much larger color-coded rectangles. Users may not have noticed this distinction, and clicked second-level rectangles intending to select the top-level categories.

Table 4. Mean number of top-level and second-level categories selected during perspectives task for the overview conditions, with standard deviation in parentheses.

	Expandable Outliner	Treemap
Top-level categories	3.07 (2.76)	0.00 (0.00)
Second-level categories	2.07 (1.22)	2.22 (1.35)
Total	5.13 (2.85)	2.22 (1.35)

3.7.3. Unusual results task

We counted the number of participants who found something unusual for each condition and scenario (Table 5).

Table 5. Number and percent of participants who found something unusual by condition and scenario.

Scenario	Control	Expandable Outliner	Treemap
Urban Sprawl	4 (67%)	6 (100%)	5 (83%)
Breast Cancer	5 (83%)	5 (83%)	6 (100%)
Alternative Energy	4 (67%)	5 (83%)	6 (100%)

For each condition, we counted the number of times participants identified unusual items. The full tables for each scenario are in the Appendix. With six participants per condition and three scenarios each, any item could be identified at most 18 times. Two unusual items were notable, both related to the number of results found from a department or agency. The table shows the number of times participants identified these two items and the corresponding percent of the maximum possible.

Table 6. Number and percent of times a participant identified selected unusual items. Maximum possible was 18 (6 participants per condition, 3 scenarios each).

Unusual item	Control	Expandable Outliner	Treemap
Why so many from a department/agency	3 (17%)	4 (22%)	8 (44%)
Why so few from a department/agency	0 (0%)	9 (50%)	4 (22%)

During the experimental sessions, we noticed that many of the 12 overview participants spontaneously commented on the lack of results from an agency. As the comments in the following sections illustrate, this could be surprising and useful information. We had not anticipated this, so we performed a post-hoc analysis of the video of all sessions, and found that

only one of the six control participants indicated (at any time during the experimental session) that they found it surprising that an agency had few or no results, whereas nine of the 12 overview participants at some time found this surprising. Based on participant comments, we concluded that the display of agencies with zero results and the color coding contributed to the searchers making such observations.

3.7.4. Subjective satisfaction measures

The subjective satisfaction questionnaire used a nine-point scale for all nine questions. Participants were asked to circle the number that most closely reflected their impression of the software. Five questions measured ranges between two assessments (1 = left-hand side, 9 = right-hand side):

1. Confusing...Understandable
2. Unhelpful...Helpful
3. Complex...Simple
4. Easy...Difficult
5. Frustrating...Satisfying

Four questions assessed agreement with the following statements (1 = disagree, 9 = agree):

6. Overall, I was able to get a good overview of the available search results for the tasks
7. For the first task in each scenario, I am confident that I found the agencies with the most pages in the search results
8. For the second task in each scenario, I am confident that I found good examples of web pages that represent different perspectives or viewpoints in the search results
9. For the third task in each scenario, I was able to find unusual results effectively

For all questions except number 4, higher values indicate higher satisfaction ratings.

Table 7. Mean subjective satisfaction measures, 1=poor, 9=good, except for #4 (Difficulty) which is reversed. Standard deviations are shown in parentheses with ANOVA degrees of freedom, F values and significance. Significant differences are shown in bold.

	Control	Expandable Outliner	Treemap	ANOVA		
				df	F	sig
1. Understandable	6.50 (1.34)	8.33 (1.21)	8.67 (0.52)	2,15	1.985	.172
2. Helpful	6.00 (1.27)	8.33 (0.52)	7.50 (0.84)	2,15	9.805	.002
3. Simple	7.50 (0.55)	7.50 (1.05)	7.50 (1.04)	2,15	0.000	1.000
4. Difficult	5.50 (0.55)	2.33 (2.34)	3.00 (1.55)	2,15	6.143	.011
5. Satisfying	5.17 (1.83)	7.83 (0.98)	6.78 (1.73)	2,15	6.698	.008
6. Overview	6.17 (2.14)	8.50 (0.84)	7.83 (0.75)	2,15	4.457	.030
7. Most pages	5.33 (1.97)	7.50 (1.38)	8.00 (2.00)	2,15	3.703	.049
8. Perspectives	6.33 (1.21)	8.33 (0.52)	7.83 (0.98)	2,15	7.222	.006
9. Unusual	4.83 (2.79)	7.33 (1.21)	6.17 (1.83)	2,15	2.235	.141

The ANOVA analyses show significant differences for questions 2 and 4-8. For these questions, the post hoc analysis shows significant differences between the control and each overview condition, but not between the two overview conditions. Table 7 shows satisfaction values with standard deviation in parentheses and ANOVA degrees of freedom, F values and significance. Users with an overview had higher satisfaction.

3.7.5. Observations and participant comments

Task A (Overview) – Most users of the control interface linearly scanned the list to get a rough idea of the top agencies. They usually scanned the list once and produced an educated guess. Several particularly motivated participants scanned the entire list twice, once to get a rough

idea of the top agencies and a second time to confirm their initial estimate by counting (spending much more time on the task). Users of the expandable outliner interface typically scanned the top-level departments, and then drilled down into the agency level. The implementation only showed one open department at a time, and participants often had to re-open a department several times to compare counts between agencies. Users of the treemap interface appeared to use the color-coding more than the expandable outliner users, and then they would scan for the counts. When the counts were not displayed (which occasionally occurred due to a programming error) they would move their pointer over the node to view the pop-up details. Many participants were puzzled or frustrated by this obvious usability flaw and commented on it. Several users of the treemap suggested that a color gradient could be used to show more detail. In both overview interfaces, some participants used the “Hide empty categories” feature extensively. The readability advantage that this provided was particularly noted in the treemap interface. In both overview conditions, several participants asked if there was a way to sort the overview by the result count.

Task B (Finding perspectives) – The control group typically scanned the results linearly until they had found three satisfactory perspectives. A few participants would scan down one or two pages, and then scan up from the bottom, stating that they expected the lower-ranked results would produce different perspectives. Most participants scanned either the title only or title and snippet. Very few of these participants appeared to use the department/agency name. The overview groups, however, often immediately clicked on a department or agency node. When asked to explain this behavior, they typically replied that their knowledge of the agency or the large number results from that agency led them to believe they would get a certain perspective by doing so. A few indicated that they just picked agencies randomly with a similar expectation. After selecting an agency, some participants would exhaustively scan the restricted list of results

before selecting another agency, while others would find an acceptable page and immediately select another agency.

Task C (Finding unusual results) – Participants typically used similar tactics as for task B. The control group participants often satisfied after a few pages. As with task B, the findings varied widely among all participants and within groups. Several participants commented:

What I found informative was... what didn't show up, which I wouldn't know if the hierarchy wasn't there.

The biggest surprises are the ones that are red [have the most results] and black [have no results]...

This participant added that if he were surprised to see an agency with no results, he would look at the uncategorized results:

I would... go to the uncategorized and see what I find there. When that was the case [it would be] frustrating that there were 70 results, but... 70 is a whole lot better than 200, and look how much I can cut out.

Several participants indicated that they selected an agency that had results but which they believed was unrelated to the topic to look for a surprising result.

For both tasks B and C, participants occasionally asked for clarification of the task or expressed concerns that they weren't sure that they were doing what had been requested.

4. STUDY 2: AUTOMATED CLUSTERING VS. GOVERNMENT HIERARCHY

4.1. Research Questions

The second study focused on the first and third research goals listed in the introduction: What tasks benefit from categorized overviews and the effect of the classification used for the overviews – using two different hierarchical classifications. Our emerging principles asserted that

search results should be organized by meaningful, stable classifications, but the variable categories returned by clustering search engines (e.g. Vivisimo) have been found helpful, even though participants sometimes fail to understand the clusters or their labels. Therefore we wished to investigate how clustered overviews supported user examination of search results. For this study, we identified two new tasks, idea generation and resource finding, as examples of more complex exploratory search tasks than were used in study 1. In particular, we were interested in three questions:

1) How can overviews based on variable categories (automated clustering) vs. stable categories (government hierarchy) affect user examination of search results with respect to domain and classification knowledge?

2) How can overviews based on variable categories (automated clustering) vs. stable categories (government hierarchy) affect user examination of search results with respect to the type of search task?

3) How can overviews based on variable categories (automated clustering) vs. stable categories (government hierarchy) affect user perceptions of search processes and outcomes?

4.2. Experimental Conditions

A within-subject experimental design (N=12) with subjective measures and qualitative observation was used to address these questions. Two experimental conditions were used by each participant: Condition 1 used the Vivisimo search engine (Figure 4), as an example of an interface using variable categories to provide an overview. Vivisimo uses a form of automated document clustering that generates hierarchies of concisely labeled clusters. Vivisimo's hierarchies satisfy many of Kwasnick's (1999) criteria, but the inclusion rules are based on mathematical relationships of surface terms rather than the human-assigned concepts used in the government hierarchy. The cluster labels are displayed using an expandable outliner to provide an overview of

the search results. Condition 2 used the expandable outliner interface from the previous study, in which results were organized by government department and agency. This experimental design unavoidably conflated several search engine and interface design issues with the classification. In addition to the different presentation style of the results, the search results for condition 1 were computed prior to the start of the experimental sessions, whereas Vivisimo was used on-line with live results. This was acceptable for our purposes, because a) the basic layout of results and interaction styles were consistent, b) we were not seeking specific quantitative measures that would be affected by these differences, and c) our focus was on subjective satisfaction measures and observation. The order of interface presentation was counterbalanced; half the participants used the Vivisimo interface first, and half used the government hierarchy first. Two of the three scenarios were used for each participant, one for each interface, allowing us to collect data from each scenario eight times over the course of the study.

4.3. Scenario and task design

As we argued earlier, the exploratory search tasks must be placed in the context of realistic higher level information seeking and work scenario to motivate the specific tasks and control for how participants interpret the search tasks. The three scenarios from study 1 were revised and adapted to more clearly specify a high-level information need and to provide a stronger indication of the organizational context. The age element was removed because it was not judged helpful in setting the context in the first study. The revised scenarios were:

Scenario 1 (Breast cancer) - *Imagine that you are a Washington Post reporter who writes about government affairs. You have been asked to research a special series of articles for the Health section on what the federal government is doing about breast cancer. You have just entered the search terms “breast cancer” in a new government search engine.*

Scenario 2 (Alternative energy) - *Imagine that you are a Senate staffer. You have been asked to write a summary of government activity on wind power as an alternative energy source as background for a comprehensive legislative funding initiative. The summary will be read by the senators and other legislative staff. It will overview federal government activities, without advocating particular actions or expressing specific opinions. As a starting point, you are using a new government search engine to gather information. You have just entered the search terms “alternative energy wind power”.*

Scenario 3 (Urban sprawl) - *Imagine that you are an undergraduate student taking a class on Science and Public Policy. Your professor has assigned a 20-page term paper on the federal government’s role in addressing urban sprawl. (Urban Sprawl is low density, automobile dependent development beyond the edge urban areas.) You are at the stage of exploring topics and defining your paper’s thesis. As a starting point, you are using a new government search engine to gather information. You have just entered the search terms “urban sprawl”.*

Within each scenario, participants were asked to perform 3 tasks:

Task A (Overview) – *Please spend 2-3 minutes exploring these search results to find out what kind of information is available.*

Task B (Idea generation) – The wording of this task was customized for each scenario (see discussion in section 4.6.1):

Scenario 1 - *Please spend 4-5 minutes using these results to formulate 2 story ideas that could be developed into a series of articles. State each story idea in a single sentence. Bookmark the pages that contribute to the ideas.*

Scenario 2 - *Please spend 4-5 minutes using these search results to find 3 examples of important programs, studies, activities, etc. that should be*

considered by anyone interested in this legislation. You should try to find the 3 most important examples within these results. Bookmark the pages.

Scenario 3 - Please spend 4-5 minutes using these results to identify 3 possible paper topics. State each topic idea as a single sentence. Bookmark the pages that contribute to the topic.

Task C (Finding resources) – Please spend 2-3 minutes using these search results to find 3 web pages likely to list sources (people or organizations) you would like to contact. Bookmark the pages you found.

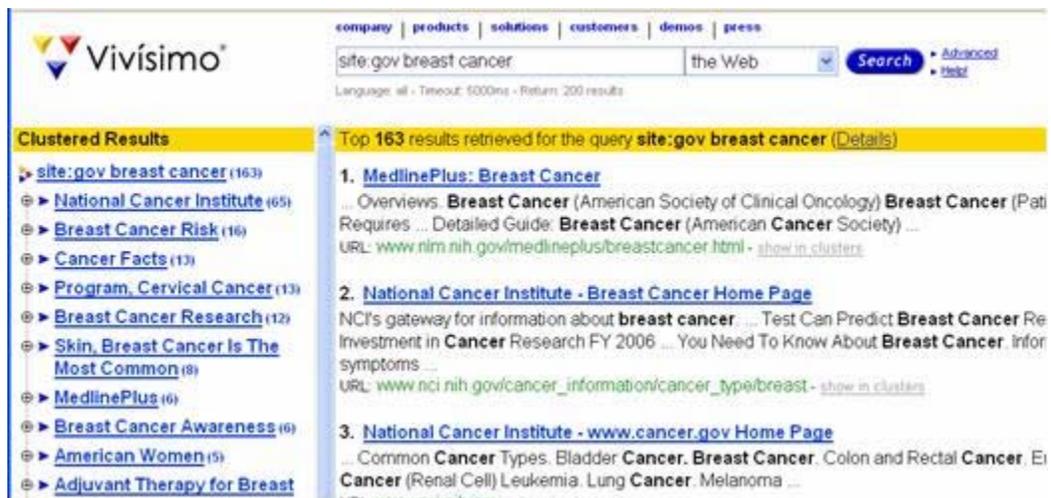


Figure 4. The Vivísimo search engine was used for the clustered hierarchy condition.

4.4. Materials and Procedure

After the participants signed an informed consent form, they completed a short demographic questionnaire, providing their age, gender, occupation, knowledge of federal government organization, web experience, search experience, search frequency and whether they had participated in study 1. The two hierarchical overviews were described and they were given a

sample task to try with both interfaces. They were encouraged to think aloud as they attempted the sample tasks, and any questions were addressed. As in the first study, participants were instructed to view just the results and categorized overview (when available). When they were comfortable with the interfaces, the first scenario was presented, and they performed the three tasks and completed a short subjective questionnaire. These steps were repeated for the second scenario. After the second scenario, participants completed another short questionnaire comparing the two interfaces and an unstructured interview was conducted to collect additional user comments. The audio and screen video for the session was captured using Camtasia (about 8 hours total). Sessions lasted approximately one hour.

The procedures and materials were pilot tested with 2 participants to clarify the scenarios and task descriptions and to streamline the questionnaires. We also clarified instructions so that participants would avoid Vivisimo's sponsored links and the "Find in clusters" feature, which was not available in the government hierarchy interface.

4.5. Participants

Twelve participants (6 male, 6 female) were recruited from university and professional contacts. They ranged in age from 22 to 58, with the average age being 42. Three were students, and six had some strong connection to the federal government, either being employees or working closely with a department or agency. All had at least a year of experience with web search and reported searching at least once/week. All except one participant reported some familiarity with the federal government. Three participants in the previous study were recruited because we were interested in whether their experience would differ from others.

4.6. Results

4.6.1. Subjective Measures

Post-scenario questionnaires - After each scenario, participants were asked to complete a short questionnaire in which they provided subjective ratings for their experience with that interface (Table 8).

Table 8. Mean differences in subjective ratings between conditions (standard deviation in parentheses). These questions were asked immediately after each scenario.

Question	Mean difference (std dev)	
	Favors automated clustering	Favors government hierarchy
Q1. Prior familiarity with topic		1.00 (3.61)
Q2a. Stressful/relaxing	0.67 (1.43)	
Q2b. Interesting/boring	0.33 (0.98)	
Q2c. Tiring/restful		0.33 (1.50)
Q2d. Easy/difficult	0.17 (1.33)	
Q3. Tried to only view related information	0.83 (0.79)	
Q4. Got a good overview of results		0.58 (3.06)
Q5. Usefulness of hierarchy for general exploration task	0.75 (4.14)	
Q6. Usefulness of hierarchy for ideas/examples task	0.83 (2.25)	
Q7. Usefulness of hierarchy for finding resources task		0.58 (2.97)
Q8. Noticed something unusual/surprising	0.08 (0.67)	
Q9. Confidence that S found good resources		0.75 (1.54)
Q10. Confidence that S generated good ideas	0.25 (2.30)	

Exit questionnaires – A post-session questionnaire solicited, participant preferences (Table 9). One participant did not answer these questions. Mean preferences are also shown with participants segmented by whether they were associated with the federal government (participants were evenly divided).

Table 9. Mean preferences for each task by all participants, participants associated with federal government and participants not associated with federal government (1 = preferred automated clustering, 9 = preferred government hierarchy).

Question	Mean preference (std dev)		
	All participants	Associated with federal government	Not associated with federal government
Q1. Preferred condition for general exploration task	3.82 (2.68)	4.00 (3.16)	3.60 (2.30)
Q2. Preferred condition for ideas/examples task	4.27 (2.45)	4.38 (2.56)	3.60 (2.40)
Q3. Preferred condition for finding resources task	6.00 (2.79)	6.67 (2.66)	5.20 (3.03)

Based on participant comments and a post-hoc review, we determined that generating ideas (scenarios 1 and 3) and finding examples (scenario 2) were not the same type of tasks. When the analysis was limited to the 4 cases in which scenarios 1 and 3 were both used, the mean preference value for question 2 was 3.25 (standard deviation 2.06), suggesting a stronger preference for the clustered hierarchy for the task of generating ideas .

4.6.2. Observations and Participant Comments

The observed interactions varied widely between participants, reflecting personal preferences, skills, knowledge, motivation and attitude. They suggest interactions between domain knowledge, task and the classification scheme.

Domain and classification knowledge – Participants applied their government knowledge to both interface conditions, but particularly to the government hierarchy:

Now I definitely want to go over here, because we're talking energy... go to DOE [Department of Energy]... you're saying wind energy... important to DOE... what other government agency?... well nothing showed up under defense, that's interesting... go to Uncategorized... The other one where wind energy might be important might be Commerce, but let's look at Energy first.

They also used opinions and biases to guide their exploration, as another participant admitted:

The fact that I have feelings about how HUD works... (laughs) and there was a subcategory that said Independent Agencies appealed to my revolutionary spirit... I said alright well who's trashing these guys...and that probably played some role...

They occasionally chose the wrong category based on incorrect domain knowledge:

Well I know that NASA is under commerce [clicks Commerce]..., oh I'm not even clicking on NASA. Is NASA part of Commerce? No, maybe it's not. It's its own independent agency [clicks Independent Agencies]. There you go, I was looking at NOAA.

For at least one participant, the utility of the government hierarchy also depended on his specific knowledge of the government relative to the scenario topic. He commented:

What you bring to it becomes a very powerful factor. The fact that I know the agencies with respect to this topic made this a snap which wasn't the case with the other one.

When using the clustered hierarchy, participants occasionally expressed confusion when they noticed that government agencies were not organized in a manner consistent with their understanding of the U.S. government's organization.

Classification and task – Participants expressed a variety of opinions on the applicability of each classification (the government hierarchy or the Vivisimo clustered hierarchy) to the different tasks (ideas versus resources). Comments included:

If I was just looking for sources of people to talk to I might prefer [the government hierarchy], but if I'm looking for ideas, stories [the clustered hierarchy] is probably more useful.

For what I do I would prefer the government thing, because at my level what I care about are finding data, but the data that I find, but the data I use has to be "blessed"... has to come from BLS... if I'm using statistics on agency size, if I want to know how big homeland security is, I got to get it from Homeland, or OMB or OPM or something like that.

One user initially found the clustered hierarchy too complex, but after using it commented:

It's sort of set up posing a question. If you want cancer facts, do you want this aspect or that? It's sort of leading you down a path. It's helping you ask the questions you need to ask, whereas you're sort of asking them intuitively, it's doing that in sort of a logical path. I like that. It's helping you burrow down into your search strategy.

But another participant was wary of the level of detail in the clustered hierarchy:

Sometimes, particularly when I'm looking for ideas, having stuff – this is the nature of the digital age – having stuff broken down too finely makes thinking more difficult, makes search for stuff more efficient but makes thinking about stuff more difficult for me... it's a lot easier for me to think in a category that talks about the statements of independent agencies... as opposed to going through [the clustered hierarchy]. I'm not necessarily looking for something that's that efficient.”

The same participant found using the clustered hierarchy condition to induce “a more deliberative process... it requires me to put a lot more into this thing.”

Category labels – Participants would often look at categories without selecting them. They expressed two reasons for this. First the category label might be meaningful but not relevant. Second, the category label might not be meaningful in the context of the scenario. As one participant commented about the labels used for the clustered hierarchy:

Stuff like 'Green' is useless to me. 'Renewable and Alternative'... is what it and a hundred other things are... doesn't save me time.

Several participants compensated for this by expanding each of those categories. This often revealed more interesting subcategories:

The refinements were more useful than the major subject headings. They get down to a level of detail that is more useful. I'd have to look and see how well that correlates... the breakdowns are actually a whole lot more useful. The next time through I'd use them more aggressively.

Assessing search results – When assessing the relevance of search result items or categories, participants commented on multiple facets, including topicality, pertinence, utility, document quality and source credibility. They often expressed skepticism about the results they found, because they were not able to view the individual web pages (due to the experimental procedure). As two participants noted:

I find a web site that seems to have a lot of really interesting stuff [in the search result list] and then find it... is sponsored by the nuclear industry and everything is powerfully skewed... or some rant by some lunatic...with federal sites in particular they have this laundry list of what they're responsible for... but it ends up so sanitized...

I'd have to see if this stuff is substantive or not... so much of this stuff is window dressing.

Acronyms appeared to be widely problematic, although we did not have quantitative measures for them. This was particularly noticeable within category labels. Even experienced government participants had puzzling encounters with unknown agency or project acronyms.

Usability of the expandable outliner – Participants found both interfaces quite understandable and quickly became comfortable with the expandable outliner. Most participants became comfortable alternating between the outliner, selecting a category, and then scanning the

search result list. Several usability issues were observed or noted by participants. The small size of the expander (a plus sign) in both interfaces caused several participants to initially overlook this capability ("I sort of forgot about this little plus thing"). One participant was irritated by the fact that in the Vivisimo interface the overview pane scrolled back to the top whenever a category is expanded.

5. DISCUSSION OF BOTH STUDIES

These two studies begin to answer the research questions posed at the beginning of this paper and suggest additional insights. They corroborated several of the emerging principles and entailed revisions to others, as discussed in the following sub-sections.

5.1. Benefits of categorized search result overviews

Study 1 confirmed that the overview conditions (the expandable outliner and the treemap) produced significantly higher successful completion rates for the task of identifying the agency with the most pages (hypothesis 1). The subjective measures showed that the overview treatments were preferred (hypothesis 2) and this was supported by user comments. Participants found the overviews significantly easier to use, more helpful, and more satisfying than the control (the standard Google interface), and they were more confident of their own success. They agreed more strongly that they had gained a good overview and found good examples of different perspectives. There was no significant difference between the three interfaces on the question of whether they had found unusual results effectively, although the difference in means is suggestive. This task was the most open-ended and most subject to interpretation by participants, and this was reflected in the subjective measure variability as well as the questions participants asked to clarify the task.

The results support our belief that the overview interfaces are seen as simple, understandable and easy to learn (i.e., hypothesis 3 of study 1 was not supported). We note, however, that participants were provided brief training in the use of the treemap.

During the perspectives task, participants found their perspectives significantly deeper in the ranked list of results. This is consistent with results reported in Käki (2005). Participants using the expandable outliner found more of their perspectives beyond the top 10 results than did participants using the control, but the treemap outcomes were mixed. Participants may have taken longer to become comfortable with the treemap interface. We observed a large variation in how participants interpreted this task.

Having the overview available helped participants to notice areas particularly well-covered and not well-covered by the search results. We attribute this to the use of the meaningful and comprehensive hierarchy, which allowed users to make inferences and draw conclusions. During the entire experimental session, only one of the six control participants found it surprising that an agency had few or no results, whereas nine of the 12 overview participants at some time found this surprising. During the Unusual results tasks, treemap users particularly noted agencies that they had not expected to have results (but that did), while expandable outliner users noticed the opposite, i.e., those agencies with few or no results. This difference might be explained by the large, colored rectangles used for the treemap (thus drawing attention to agencies with results) and the expandable outliners linear arrangement of text (which encouraged scanning of agency names). This was echoed in the participant comments and suggests that color coding might be more useful in the expandable outliner if used more extensively.

5.2. Effect of visual presentation of overviews

The appeal of both overviews was confirmed by the lack of statistically significant differences between the expandable outliner and the treemap. Most participants preferred the expandable outliner, although several participants found the graphical nature of the treemap more

appealing. The participant comments suggest that additional user control of the overview would be desirable. This included allowing participants to select the desired presentation, as well as creating or selecting the categorization scheme used.

5.3. Effect of classification used for overviews

When the overview was available participants took advantage of it, even when the organizing structure was not optimal for the task. Observations and participant comments indicated that participants used their prior knowledge of the classification to interpret search results. Participants indicated that they became more familiar with the government hierarchy over the course of the experiment. Because the government hierarchy is stable, this familiarity may be beneficial in successive searches.

In study 2, the participants appreciated the dynamically generated hierarchy for the ideas task. Its statistically based clustering yielded labels that they found suggestive of topic ideas. Other participants felt strongly that the government hierarchy helped them explore and understand the results more effectively. The inclusion rules were more transparent and predictable to users for the government hierarchy than for the Vivisimo hierarchy, providing a more comprehensive overview. Based on the results of study 2, we revised our second principle (which was originally “Organize results by meaningful, stable classifications”) to reflect the complementary nature of stable and dynamically generated classifications. Together, they supported a variety of exploratory search sub-tasks.

Individual user characteristics as well as task type appeared to affect user preferences for the classification hierarchy, suggesting that searchers be allowed to select from multiple organizational schemes. Several participants commented that they would like the ability to organize results in multiple ways, possibly customizing their own organization scheme. This buttresses principle six (Support multiple visual presentations and classifications), suggesting that

the faceted category approach (Yee et al., 2003) could be beneficial for organizing web search results. There may also be value in user-created or customizable taxonomies.

5.4. The importance of text

Observations and participant comments confirmed that text was important, even with the overviews available. As one person noted, the overview was a starting point. But searchers still needed to scan substantial amounts of text. This was particularly noticeable with those participants who interpreted the tasks more realistically, requiring in-depth evaluation/assessment. This bolstered our confidence in principle four (Arrange text for scanning/skimming).

5.5. Other findings

Government agency acronyms were problematic for all participants, particularly within category labels. A simple capability to perform a glossary lookup would probably be very helpful. Using hover text could allow searchers to pause the pointer over unfamiliar acronyms to see the full name of the agency or department.

Participants rarely commented on the need to scroll within either the overview or results list. This suggested that it is a very lightweight action, and may not substantially affect the searcher's cognitive process. It further suggests that larger sets of results (at least 100-200) can be usefully accommodated on a single page. Google, Yahoo!, and Vivisimo can return 100 results per page (with typical load times less than 5 seconds on a broadband network), so this is technically feasible.

5.6. Limitations of these studies

These studies were exploratory in nature, and the results must be interpreted within the context of the specified tasks and domain. We used a small sample of subjects, who were presented with pre-defined scenarios, queries and tasks. The government hierarchy was limited in

size and the specific tasks represented only a small slice of the tasks searchers perform in real-world topic searches. But, based on participant comments, the scenarios appeared to evoke a realistic information need in the subjects, and we used tasks that exploratory searchers really do perform. Examining large numbers of results and evaluating them in the context of current knowledge are characteristic of exploratory search tasks. By focusing on a specific domain (government web search), we limited the immediate scope of our findings in return for gaining a deeper understanding of how searchers used categorized search results within that domain.

6. CONCLUSIONS AND FUTURE WORK

The results of these formative studies suggest answers to our original three research goals: Exploratory search tasks can be supported by categorizing search results into comprehensible visual overviews using meaningful classifications. Stable classifications and dynamically generated classifications can be complementary ways to organize results. The use of stable hierarchies helped participants notice missing information, and the dynamically generated classifications were found useful for generating topic ideas. The study results also motivated several new requirements: user-selectable classifications and a lightweight mechanism for customizing hierarchies. To better support user-selectable classification, we are investigating the efficacy of “lightweight” classifications, which use simple schemes (e.g. DNS domain, the last time the document was viewed, or document size) to organize results into easily understood categories.

The studies were used to refine two of the six principles and they reinforced our confidence in three others. They raised the question of which tasks are best supported by stable categories vs. dynamic categories. Many important issues remain, such as spatial layout, textual elements and dynamic interactions of categorized search result visualization and optimal characteristics (e.g. breadth and depth) of hierarchy. Additional research is needed to model the specific strategies and tactics that searchers apply within exploratory search sub-tasks, and to

more fully explicate their use of domain, classification and interface knowledge. This could inform the development of a cognitive model of user exploration and understanding of categorized search results.

Situating the study tasks within the specific domain of government web search, and within higher level work tasks, reduced variation in participants' perception of the tasks without resorting to known-item search tasks. It allowed us to collect a rich set of observations about how searchers use categorized search results. Studies in other domains and with other classifications are needed to confirm the findings.

These formative studies are one step toward a better understanding of how exploratory searchers use categorized overviews. As the principles are refined and extended, they can be used by practitioners – the designers and developers of Web search engines – to help realize more effective interfaces for learning and research on the Web. Categorizing search results using meaningful and stable categories is a promising way to alleviate information overload while supporting user exploration and understanding of large sets of search results.

7. ACKNOWLEDGEMENTS

We would like to thank Alex Aris for his help during the first study. Doug Oard, Ryan White, Catherine Plaisant, Haixia Zhao, Marti Hearst and the anonymous reviewers provided many helpful comments and suggestions. This research was supported by an AOL Fellowship in Human-Computer Interaction and National Science Foundation Digital Government Initiative grant (EIA 0129978) “Towards a Statistical Knowledge Network.”

REFERENCES

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3).

- Byström, K., & Hansen, P. (2002). Work tasks as units for analysis in information seeking and retrieval studies. In H. Bruce, R. Fidel, P. Ingwersen & P. Vakkari (Eds.), *Emerging Frameworks and Methods* (pp. 239-251). Greenwood Village, CO: Libraries Unlimited.
- Ceaparu, I., & Shneiderman, B. (2004). Finding governmental statistical data on the Web: A study of categorically organized links for the FedStats topics page. *Journal of the American Society for Information Science and Technology*, 55(11), 1008 - 1015.
- Chen, H., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7), 582-608.
- Chen, M., Hearst, M., Hong, J., & Lin, J. (1999, October 11-14, 1999). *Cha-Cha: A system for organizing intranet search results*. Paper presented at the Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems, Boulder, CO.
- Drori, O., & Alon, N. (2003). Using Documents Classification for Displaying Search Results List. *Journal of Information Science*, 29(2), 97-106.
- Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 277-284.
- Egan, D. E., Remde, J. R., Gomez, L. M., Landauer, T. K., Eberhardt, J., & Lochbaum, C. C. (1989). Formative design evaluation of superbook. *ACM Trans. Inf. Syst.*, 7(1), 30-57.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.
- Hearst, M., Elliot, A., English, J., Sinha, R., Swearingen, K., & Yee, P. (2002). Finding the flow in web site search. *Communications of the ACM*, 45(9), 42-49.
- Hearst, M. A. (1999). The Use of Categories and Clusters for Organizing Retrieval Results. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval*. Boston: Kluwer Academic Publishers.

- Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 76-84.
- Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology*, 56(6), 559-570.
- Järvelin, K., & Ingwersen, P. (2004). Information seeking research needs extension towards tasks and technology. *Information Research*, 10(1).
- Käki, M. (2005). Findex: search result categories help users when document ranking fails, *Proceeding of the SIGCHI conference on Human factors in computing systems*. Portland, Oregon, USA: ACM Press.
- Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5), 361-371.
- Kules, B., & Shneiderman, B. (2003). Designing a Metadata-Driven Visual Information Browser for Federal Statistics. *Proceedings of the 2003 National Conference on Digital Government Research*, 117-122.
- Kules, B., & Shneiderman, B. (2004). Categorized Graphical Overviews for Web Search Results: An Exploratory Study Using U.S. Government Agencies as a Meaningful and Stable Structure. *Proc. Third Annual Workshop on HCI Research in MIS*, 20-24.
- Kules, B., & Shneiderman, B. (in process). Design principles for categorized web search results: an interdisciplinary approach.
- Kunz, C. (2003). SERGIO - An Interface for context driven Knowledge Retrieval. *Proceedings of eChallenges, Bologna, Italy, 2003*.
- Kwasnik, B. H. (1999). The Role of Classification in Knowledge Representation and Discovery. *Library Trends*, 48(1), 22-47.

- Marchionini, G. (1995). *Information Seeking in Electronic Environments*: Cambridge University Press.
- Marchionini, G., Plaisant, C., & Komlodi, A. (1998). Interfaces and Tools for the Library of Congress National Digital Library Program. *Information Processing & Management*, 34(5), 535-555.
- Marshall, B., McDonald, D., Chen, H., & Chung, W. (2004). EBizPort: collecting and analyzing business intelligence information. *Journal of the American Society for Information Science and Technology*, 55(10), 873-891.
- Nation, D. A., Plaisant, C., Marchionini, G., & Komlodi, A. (1997). Visualizing websites using a hierarchical table of contents browser: WebTOC. *Proceedings of the 3rd Conference on Human Factors and the Web*.
- Pratt, W., Hearst, M. A., & Fagan, L. M. (1999). A knowledge-based approach to organizing retrieved documents. *Proceedings of the 16th national conference on Artificial intelligence and the 11th Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, 80-85.
- Rivadeneira, W., & Bederson, B. B. (2003). A Study of Search Result Clustering Interfaces: Comparing Textual and Zoomable User Interfaces. *University of Maryland HCIL Technical Report HCIL-2003-36*.
- Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search, *Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM Press.
- Shiri, A. A., & Revie, C. (2000). Thesauri on the web: current developments and trends. *Online Information Review*, 24(4), 273-279.
- Shneiderman, B., Feldman, D., Rose, A., & Grau, X. F. (2000). Visualizing Digital Library Search Results with Categorical and Hierarchical Axes. *Proc. 5th ACM International Conference on Digital Libraries (San Antonio, TX, June 2-7, 2000)*, 57-66.

- Soergel, D. (1999). The Rise of Ontologies or the Reinvention of Classification. *Journal of the American Society for Information Science and Technology*, 50(12), 1119-1120.
- Spink, A., & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. New York: Kluwer.
- Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 52(3), 226-234.
- Turetken, O., & Sharda, R. (2005). Clustering-based visual interfaces for presentation of web search results: An empirical investigation. *Information Systems Frontiers*, 7(3), 273-297.
- Vakkari, P. (2001). A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of Documentation*, 57(1), 44-60.
- White, R. W., Kules, B., Drucker, S. M., & schraefel, m. c. (2006). Supporting exploratory search. *Communications of the ACM*, 49(4).
- Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 401-408.
- Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results, *Proceeding of the eighth international conference on World Wide Web*. Toronto, Canada: Elsevier North-Holland, Inc.
- Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., & Ma, J. (2004). Learning to cluster web search results, *Proceedings of the 27th annual international conference on Research and development in information retrieval*. Sheffield, United Kingdom: ACM Press.

8. APPENDIX

The following three tables list the perspectives identified for each scenario in study 1, and the number of times each was identified within each condition.

Table 10. Perspectives identified for the Urban Sprawl scenario.

Perspective	Rank in results	Control	Expand-able Outliner	Tree-map	Total
Health-public health	2	4	1	3	8
NASA-satellite mapping	6	3	2	3	8
other-Interior Dept.		1	2	1	4
Health-obesity	8		2	1	3
overview-Definition of urban sprawl	9			3	3
environmental		2	1		3
Health-NIH		1	1	1	3
environmental-agricultural impact	3		2		2
autos/traffic		1		1	2
economic factors			2		2
environmental-air pollution			1	1	2
overview-big picture	1			1	1
assessing	3			1	1
other-Michigan	5	1			1
development-brown fields				1	1
development-coastal			1		1
development-density				1	1
development-Smart growth			1		1
environmental-photosynthesis		1			1
environmental-water resources			1		1
Health-CDC		1			1
NASA			1		1
NASA-scientific		1			1
Total		16	18	18	

Table 11. Perspectives identified for the Breast Cancer scenario.

Perspective	Rank in results	Control	Expand-able Outliner	Tree-map	Total
other-male BC	2	4	3	1	8
research-NASA/space based		1	2	2	5
general info-self-detection, diagnosis, screening	5, 7	1	1	2	4
general info-what you need to know	4,3	2	1	1	4
risks-assessment	10		2	2	4
legislation-senate		1	1	1	3
reports-medline	1	1		2	3
research-genes			3		3
general info-treatments		2			2
legislation		1		1	2
other-NIH				2	2
other-NIH-NCI	3			2	2
risks-heart		1		1	2
general info-cancer types		1			1
general info-early detection		1			1
general info-facts		1			1
other-NOAA				1	1
other-pre-knowledge/post-knowledge			1		1
reports-news/scientific		1			1
research-studies-biggest is NIH			1		1
risks-anti-perspirant			1		1
risks-environmental			1		1
Total		18	17	18	

Table 12. Perspectives identified for the Alternative Energy scenario.

Perspective	Rank in results	Control	Expand-able Outliner	Tree-map	Total
agriculture	5	4	2	1	7
legislation-presidential initiative			1	3	4
mailing list	1	3			3
promotion-benefits	2	2	1		3
legislation-house		1		1	2
legislation-tax code			1	1	2
lists of technology	6	1	1		2
medical use		1		1	2
sustainable				2	2
who [agency] is dealing with it				2	2
coast guard			1		1
economic-energy futures				1	1
Economic-hydro power-cost			1		1
environmental-climate change				1	1
environmental-conservation			1		1
environmental-green communities	9	1			1
form			1		1
halogen alternatives			1		1
info		1			1
info-overview		1			1
land management				1	1
legislation-senate			1		1
microbial				1	1
NOAA-current law				1	1
products of process			1		1
promotion-educational		1			1
prototypes			1		1
renewable	4			1	1
reporting-statistics		1			1
source			1		1
source-biomass energy			1		1
source-fuel cells		1			1
source-fuels/crops				1	1
source-solar power			1		1
studies-DOE labs			1		1
Total		18	18	18	

The following three tables list the unusual results identified for each scenario in study 1. If a participant identified multiple instances of the same value within the scenario, that was counted as one instance, i.e., noticing missing results from two agencies within the Urban Sprawl scenario would be coded as one instance. The user's first reaction was counted, even if they subsequently explained the instance and/or changed their mind.

Table 13. Unusual results identified for the Urban Sprawl scenario.

Unusual-1 (Urban Sprawl)	Control	ExpOut	TM	Total
why not more from agency		3		3
why so many/why any at all from agency		2	1	3
NASA-why/satellite images	1	1	1	3
Myths	1	1		2
Obesity	1			1
library of Michigan	1			1
desert blooms-guide to plants	1			1
miscategorized page		1		1
aggressive driving		1		1
measuring heat		1		1
why does lab link urban sprawl with natural disasters			1	1
invalid titles			1	1
coastal growth			1	1
hadn't clicked on that yet			1	1
Total	5	7	6	

Table 14. Unusual results identified for the Breast Cancer scenario.

Unusual-2 (Breast Cancer)	Control	ExpOut	TM	Total
why not more from agency		2	2	4
why so many/why any at all from agency	1	1	2	4
NASA-space based research	1	1	2	4
Male BC	3	1		4
myths	1	1	1	3
simulations of BC	1	1		2
FAQ on hereditary	1			1
hawaii	1			1
new gene found	1			1
CBCTR	1			1
SPORES project	1			1
URL changed		1		1
Defense bill		1		1
surveillance		1		1
expected general pages to be ranked higher		1		1
LOC/tracer bullets			1	1
economic statistics			1	1
Total	12	11	9	

Table 15. Unusual results identified for the Alternative Energy scenario.

Unusual-3 (Alternative Energy)	Control	ExpOut	TM	Total
why so many/why any at all from agency	2	1	5	8
why not more from agency		4	2	6
atrial defibrillation/AW for medical use	1	2		3
mailing list	2			2
student congressional town meeting		1	1	2
health	1			1
titles not helpful	1			1
photosynthesis	1			1
north korea	1			1
how few provide overviews	1			1
USAID & Brazil	1			1
Yurok	1			1
climate change	1			1
miscategorized page		1		1
homeland security		1		1
computer aided manufacturing		1		1
why not more wacky sites			1	1
Total	13	11	9	