ABSTRACT

Title of Document: SUPPORTING EXPLORATORY WEB SEARCH WITH MEANINGFUL AND STABLE CATEGORIZED OVERVIEWS

William Michael Kules, III,
Doctor of Philosophy, 2006

Directed By: Professor Ben Shneiderman,
Department of Computer Science

This dissertation investigates the use of categorized overviews of web search results, based on meaningful and stable categories, to support exploratory search. When searching in digital libraries and on the Web, users are challenged by the lack of effective overviews. Adding categorized overviews to search results can provide substantial benefits when searchers need to explore, understand, and assess their results. When information needs are evolving or imprecise, categorized overviews can stimulate relevant ideas, provoke illuminating questions, and guide searchers to useful information they might not otherwise find. When searchers need to gather information from multiple perspectives or sources, categorized overviews can make those aspects visible for interactive filtering and exploration. However, they add visual complexity to the interface and increase the number of tactical decisions to be made while examining search results.

Two formative studies (N=18 and N=12) investigated how searchers use categorized overviews in the domain of U.S. government web search. A third study (N=24) evaluated categorized overviews of general web search results based on thematic, geographic, and government categories. Participants conducted four exploratory searches during a two hour session to generate ideas for newspaper articles about specified topics. Results confirmed positive findings from the formative studies, showing that subjects explored deeper while feeling more organized and satisfied, but did not find objective differences in the outcomes of the search task. Results indicated that searchers use categorized overviews based on thematic, geographic, and organizational categories to guide the next steps in their searches.

This dissertation identifies lightweight search actions and tactics made possible by adding a categorized overview to a list of web search results. It describes a design space for categorized overviews of search results, and presents a novel application of the brushing and linking technique to enrich search result interfaces with lightweight interactions. It proposes a set of principles, refined by the studies, for the design of exploratory search interfaces, including “Organize overviews around meaningful categories,” “Clarify and visualize category structure,” and “Tightly couple category labels to search result list.” These contributions will be useful to web search researchers and designers, information architects and web developers.

SUPPORTING EXPLORATORY WEB SEARCH WITH MEANINGFUL AND STABLE CATEGORIZED OVERVIEWS

By

William Michael Kules, III

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Professor Ben Shneiderman, Chair
Professor Dagobert Soergel
Associate Professor Douglas W. Oard
Assistant Professor Lise Getoor
Catherine Plaisant, Associate Research Scientist

Dedication

This dissertation is dedicated to

the memory of Abbott and Wanda Washburn,

and to Julia, Genna, and Ruby.

Acknowledgements

Researching and writing this dissertation ranks among the most satisfying intellectual activities I have been privileged to pursue. It would not have been possible without the support of family, friends, mentors, and colleagues.

Ben Shneiderman introduced me to the field of human-computer interaction research. He showed me a path that allows me to pursue my interest in technology while contributing – in a small but direct way – to humane ends. He has mentored me as I followed this path, provided financial support, made terrific opportunities available to me, and encouraged me when I questioned my ability to finish.

My committee members have challenged and supported me along the way. Doug Oard has consistently challenged me to sharpen my thinking, clarify my writing, and more deeply explore fundamental human-computer interaction issues. Lise Getoor has provided practical advice and feedback at critical junctures. Dagobert Soergel has pushed me to expand my understanding of information organization beyond data structures, to appreciate the human importance of classification much more deeply than I otherwise would. Catherine Plaisant has been a mentor since the beginning of my association with the Human-Computer Interaction Lab. She facilitated my early work with government agencies, provided detailed advice on my research, introduced me to the espresso machine on the fourth floor, and has always been available to bounce ideas around.

Other colleagues, mentors, and friends have helped in ways too numerous to recount. I am grateful for help from Alex Aris, Ira Chinoy, Gene Chipman, Abdur Chowdhury, Chip Denman, Jerry Fails, Kathleen Grathwol, Harry Hochheiser, Chang Hu, Hilary Hutchinson, Jack Kustanowitz, Tom Lalonde, Katy Lawley, Jaime Montemayor, Craig Murray, Anne Rose, Kiki Schneider, Greg Smith, Ryen White, Haixia Zhao, Julie Zito, the study participants, and many others in the Human-Computer Interaction Lab and beyond. The staff of the Computer Science Department and at the Institute for Advanced Computing Studies have provided valued administrative support. This research was supported in part by an AOL Fellowship in Human-Computer Interaction and National Science Foundation Digital Government Initiative grant (EIA 0129978) “Towards a Statistical Knowledge Network.”

My deepest gratitude goes to my family: my daughters, Genna and Ruby, who put up with a too-occasionally distracted dad, and to my partner and wife, Julia Washburn. Her unwavering support during a decade of graduate school made this possible.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation

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. These users struggle with information overload, coping with an overabundance of information that lacks a comprehensible organization. Search engines are effective at generating extensive lists of results that are highly relevant to user-provided query terms. For known-item queries, users often find the site they are looking for in the first page of results. However, a list may not suffice for more sophisticated exploratory tasks, such as learning about a new topic or surveying the literature of an unfamiliar field of research, or when information needs are imprecise or evolving (White, Kules, Drucker, & schraefel, 2006).

The lack of comprehensible overviews of web search results is particularly problematic when users initiate exploratory searches to satisfy information needs that are imprecise or evolving or when their domain knowledge is limited. Incompletely formulated queries yield a plethora of potentially relevant search results, which must be examined and understood. This is exacerbated by the frequent use of short queries (Spink, Wolfram, Jansen, & Saracevic, 2001 & Saracevic, 2001). Although it is difficult to quantify the prevalence of such exploratory searches, recent analysis of

search goals suggests that between 20-30% of all web queries may be exploratory in nature (Rose & Levinson, 2004), which motivates study of this type of search.

This dissertation explores the premise that organizing search results into comprehensible visual overviews using meaningful and stable categories can support user exploration and understanding of large search result sets. When searchers need to gather information from multiple perspectives or sources, categorized overviews can organize results from web or digital library searches. Categorized overviews can help searchers explore alternative sources, assess utility of results, and decide on next steps. When searchers' information needs are evolving or imprecise, categorized overviews help by stimulating relevant ideas, provoking illuminating questions, and guiding searchers to useful information they might not otherwise find.

Research prototypes and commercial search engines have incorporated categorized overviews, but (as discussed in the Related Work section) there have been few user studies of categorized overviews for exploratory web search, and there is little research explaining whether they are effective, why, and under what circumstances. Research is needed to understand how categorized overviews change the way users conduct web searches, to guide the design of search engine interfaces, and to justify the entry and maintenance of category metadata.

1.2 Illustrative example

A simple scenario, using the SERVICE search system (described in Chapter 4), illustrates the use of categorized overviews in a web search. Genna, who is 10, has

been assigned a homework problem to find the median value of a set of numbers. Her father wants to quickly find an age-appropriate definition and example for her. He isn't sure what query terms would best limit his query to age-appropriate definitions, so he types "median" into the search engine and peruses the results (Figure 1). The fifth item in the list looks promising, so he clicks on it to view the page, but it turns out to be too wordy. Placing the pointer over the Kids and Teens category pops up a list of its nonempty subcategories and highlights the two visible results that fall in the Kids and Teens category (Figure 2). These two items are in the Wikipedia, so they might be helpful, but he sees a subcategory called School Time that looks more promising and he decides to see the list of all the Kids and Teens results. Clicking on Kids and Teens yields a list of child-friendly web pages. "Lesson on the Median of a Set of Data" is no longer available, but "How to Calculate the Median Value" looks like what he wants. The snippet says it is for K-12 kids and uses easy language. He clicks on the result and finds exactly what he needs.

This example illustrates several common elements of exploratory search using categorized overviews. Genna's father did not know what term to use in his query to select for age-appropriate pages. He did know that there was a top-level category for Kids and Teens, because he had seen it on previous searches, so he was confident that he could use a broad query and then narrow his results if needed. After scanning the result list, he used the categories. The pop-up subcategories provided additional information that induced him to explore all the Kids and Teens results instead of clicking on the Wikipedia results. Finally, the desired item was ranked #29 in the

original list, so he would have had to scroll or page to the third screen before he would have found it without the category overview. For comparison, Figure 4 shows an automatically clustered overview from Clusty.com for the same query, which does not provide a meaningful cluster label for child-friendly pages.

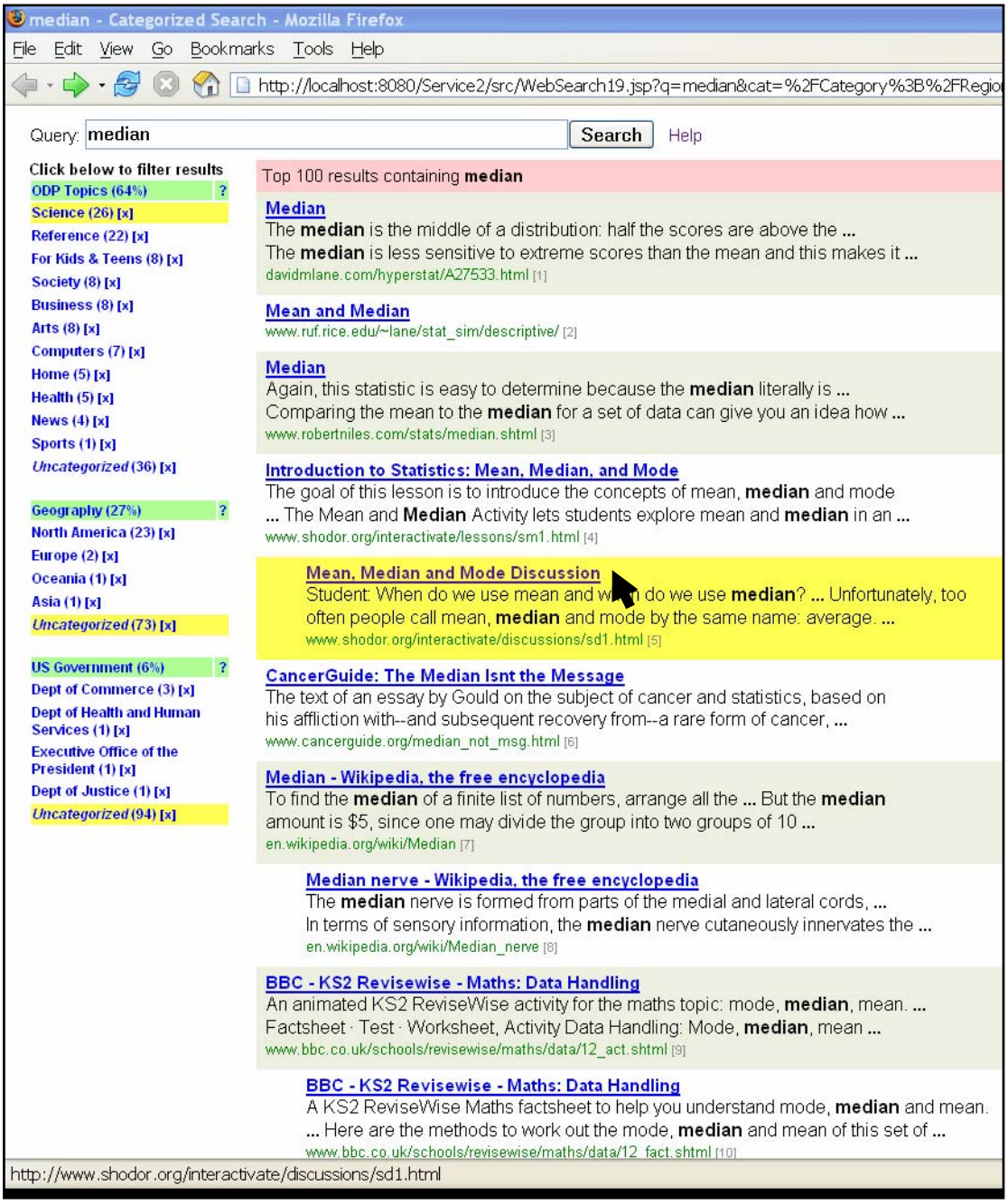

Figure 1. Search results for the query "median" are coupled with a categorized overview.

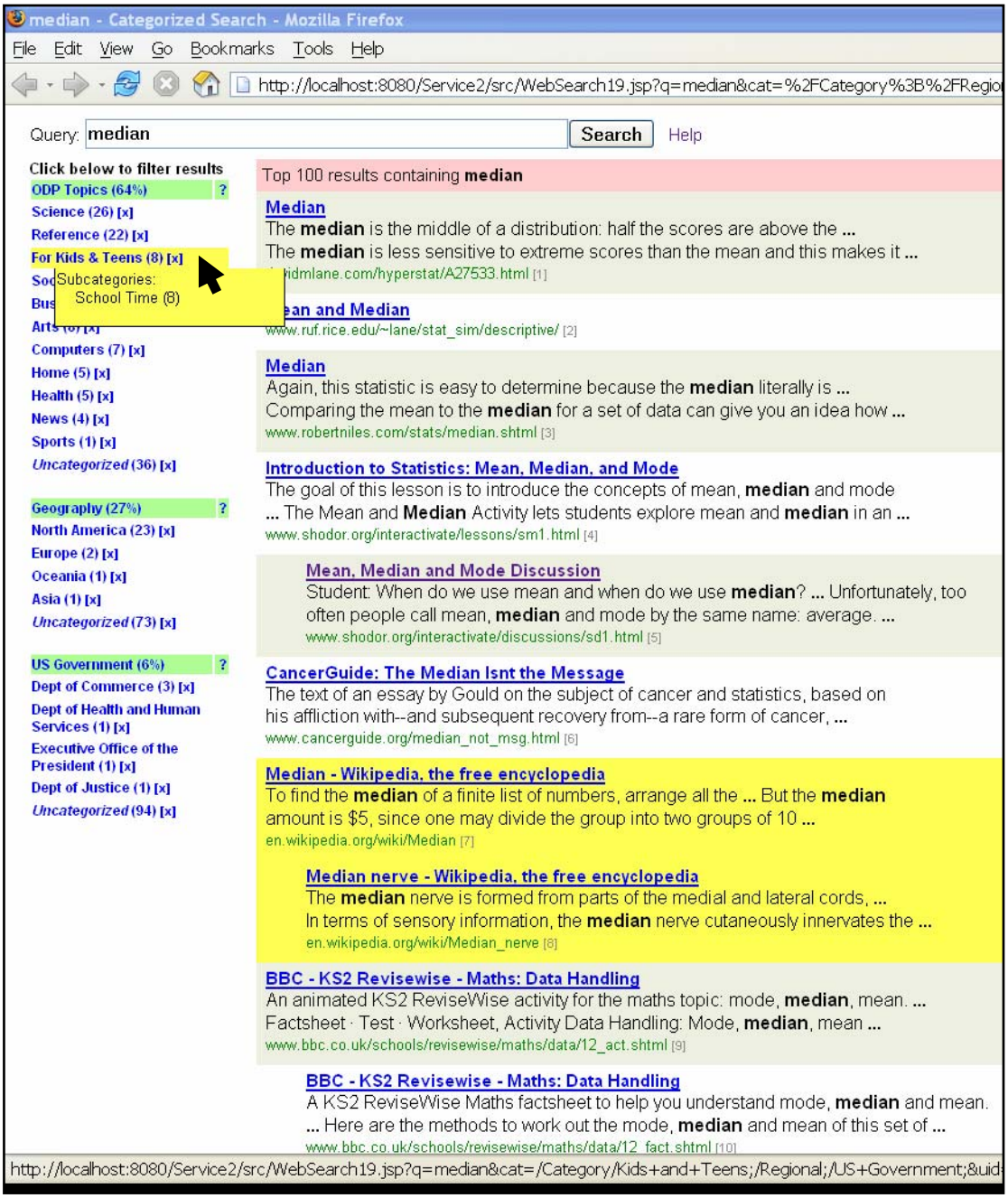

Figure 2. Placing the pointer over the Kids and Teens category pops up a list of its nonempty subcategories and highlights the visible results in the Kids and Teens category.

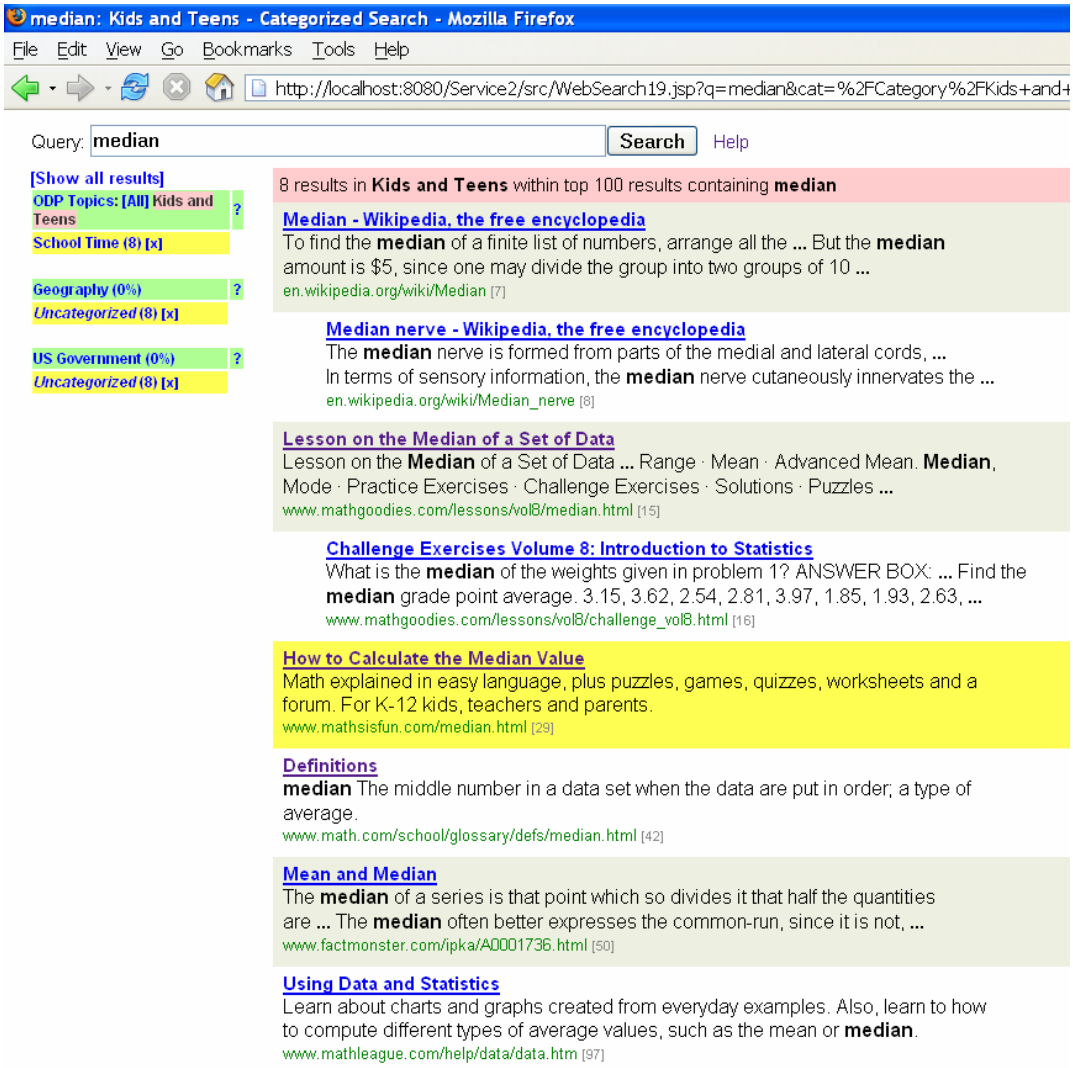

Figure 3. Selecting the Kids and Teens category filters the results to just that category.

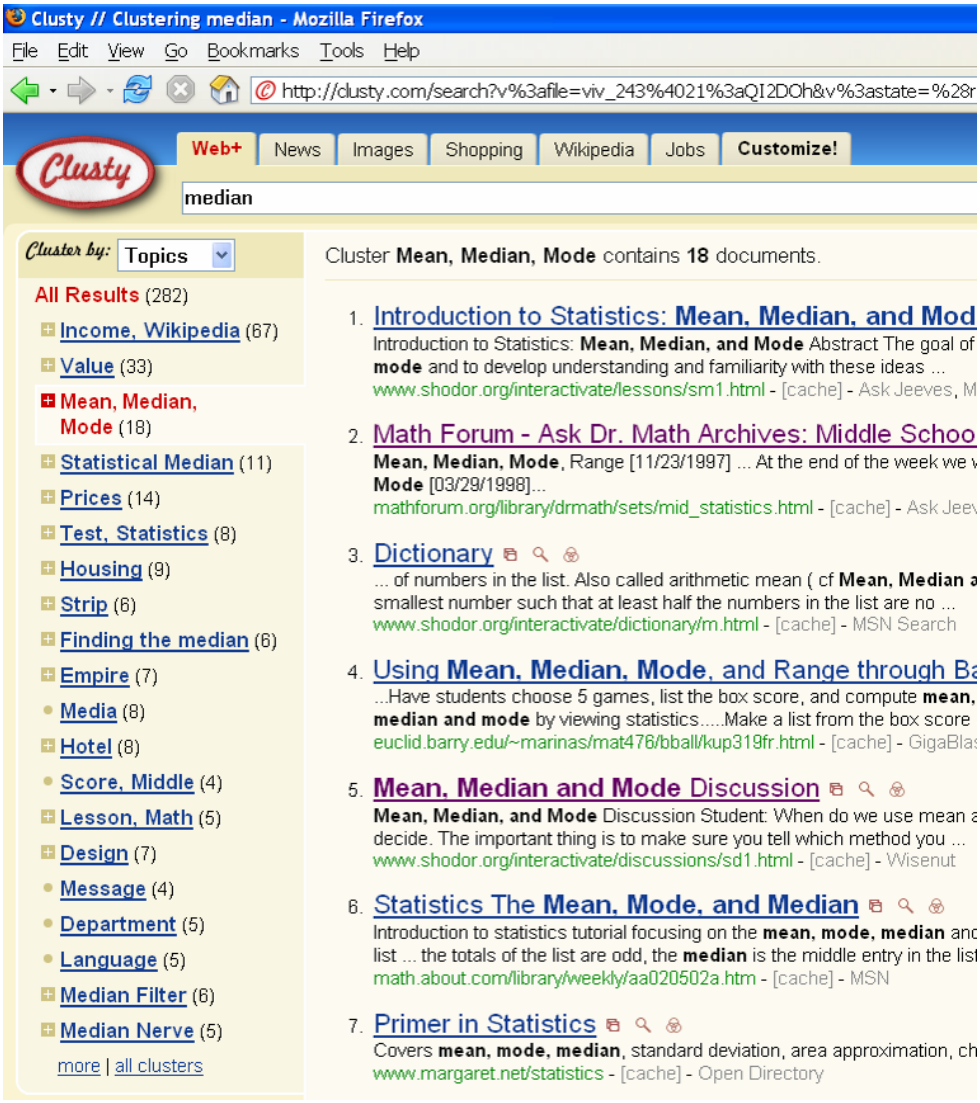

Figure 4. This automatically clustered overview for the same query, from Clusty.com, does not provide a meaningful cluster label for child-friendly pages.

1.3 Research contributions

This dissertation investigates the use of categorized overviews based on meaningful and stable categories to support exploratory search. It makes three contributions. First, it presents an analysis of search with categorized overviews, particularly focusing on how searchers evaluate their search results and decide their next move

(e.g. scroll/page for more results, refine their query, revise their conception of the information need, etc.).

The analysis provides theoretical support for the second contribution, a set of principles for the design of search interfaces to support exploratory search. The principles, refined and validated by empirical studies, complement and extend general human-computer interaction, web design, information architecture, and information visualization principles. They will be useful for search interface designers, because they provide guidance for the appropriate integration of visual overviews with search result lists, and particularly for the textual surrogates embedded in result lists. These principles represent a strong call for exposing meaningful structure – which is often used internally by search engines, but less often visible at the user interface – without abandoning the tried and true value of text.

The final contribution of this dissertation research is the SERVICE (SEarch Result Visualization and Interactive Categorized Exploration) architecture and implementation technology, illustrated with two working categorizing search interfaces: AOL music search and general web search. The ideas embedded in the user interface will be useful to designers of other search interfaces. The SERVICE system will be a flexible, extensible platform for additional research in categorizing search interfaces.

1.4 Terminology

In this dissertation, the term *category* is used to designate a concept (with an associated label) for grouping entities such that all of the entities that are members of that group share a common attribute. A category may be drawn from a formally defined classification or ontology with controlled vocabulary or indexing language. Alternately, it may come from an informal grouping that is simply meaningful within a context of use. This broad definition glosses over differences between categorization and classification systems, and between different types of classifications (Jacob, 2004; Soergel, 1974; Taylor, 1999). For this work, the most important characteristic of a set of categories is that the categories provide *some way* to organize and filter search results that is meaningful, and ultimately useful, to information seekers.

Chapter 2: Related work

Exploratory search is a sub-task in the context of a higher-level information seeking task, which is in turn motivated by a perceived information need. Searchers interact with search engines or search systems to formulate and execute queries, examine results, and browse for information to satisfy their information need. Categories may be used to organize results, which are then visualized for searchers to examine and use. This chapter presents a review of three areas of work related to this dissertation: information seeking (section 2.1), the use of categories to support information seeking (section 2.2), and the visualization of search results (section 2.3).

2.1 Information seeking – theory, studies and systems

Evolving information needs form a core motivation for information seeking. Dervin and Nilan (1986) consider user needs in the context of a sense-making theory of human behavior. Gaps in knowledge are conceptualized as questions, which can motivate a person to seek information. Belkin (1980) developed the Anomalous States of Knowledge model to explain information seeking behavior on open-ended questions. The model addresses iteration and refinement of the seeker's knowledge, specification of the problem, and an evolving ability to articulate requests. Kuhlthau's model of the stages of the information seeking process tracks cognitive and affective states in a constructive knowledge acquisition process such as writing a paper (Kuhlthau, 1991). Particularly in the latter two models, users' information needs are initially ill-defined, requiring a process of refinement. Marchionini's electronic browsing model includes problem definition and refinement in a seven-stage process

(Marchionini, 1995). Choo, Detlor and Turnbull (2000) develop a behavioral model of organizational information seeking on the web by integrating Ellis' (1989) six stages of information seeking (starting, chaining, browsing, differentiating, monitoring, and extracting) with Aguilar's (1988) four modes of scanning (undirected viewing, conditioned viewing, informal search, and formal search).

Problem refinement is inherent in each of these models, as users struggle to understand available information, refine the information need, and find new information. There has been growing interest in successive searching on the Web, in digital libraries, and in online public access catalogs (OPACs). Studies have found that users perform repeated searches on similar topics over a period of time. Spink, Bateman and Jansen (1999) surveyed users of the Excite search engine and found that two-thirds performed successive searches, with 30% searching at least 6 times on one topic. Spink, Wilson et al. (2002) found that successive searches often involved refining or extending previous searches in response to changes in understanding and evaluation of previous results. Vakkari (2000) studied 11 students who attended a two-semester proposal writing seminar, and found that as students progressed, they used more search terms and the search terms were more specific.

Many information seeking environments have been developed. The Digital Library Integrated Task Environment (DLITE) supports interaction with multiple search services while developing bibliographic citations (Cousins, Paepcke, Winograd, Bier, & Pier, 1997). It supports iterative searching by providing a persistent desktop on

which queries, results and services are maintained. The SketchTrieve system provides a similar information seeking environment, with an emphasis on allowing the user to connect services to generate search results, then place and annotate them (Hendry & Harper, 1997). The NaviQue workspace supports information seeking using a navigational perspective, based on a zooming user interface (Furnas & Rauch, 1998). More recently, researchers have advocated embedding the search function into application environments to support task-specific searching (Hendry, to appear).

Traditional OPACs allow users to browse and search using subject classifications. Allen (1995) describes two digital library interfaces based on two hierarchical classifications, the Dewey Decimal System and the ACM Computer Reviews classification. These interfaces show search results against the classification hierarchy and integrate several other features. HIBROWSE, an OPAC system, exploits faceted hierarchies to provide visual query specification and to organize results (Pollitt, 1997). Flamenco (Figure 5) provides interfaces to specialized collections (art, architecture, and tobacco documents), using faceted hierarchies to produce menus of choices for navigational searching (Hearst et al., 2002). The Envision digital library of computer science literature displayed search results using a matrix of icons, allowing searchers to easily manipulate the visualization (Nowell, France, Hix, Heath, & Fox, 1996). Citiviz (Figure 6) displays search results using a hyperbolic tree (Lamping & Rao, 1996) and a scatterplot (Perugini et al., 2004). The Technical Report Visualizer prototype (Ginsburg, 2004) allows users to browse a digital library by one of two user-selectable hierarchical classifications, also displayed as hyperbolic

trees and coordinated with a detailed document list. Categorized overviews are used in the Punchstock image search interface (punchstock.com, Figure 7) and the search interface for the North Carolina State University (NCSU) library catalog (www.lib.ncsu.edu/catalog/, Figure 8).

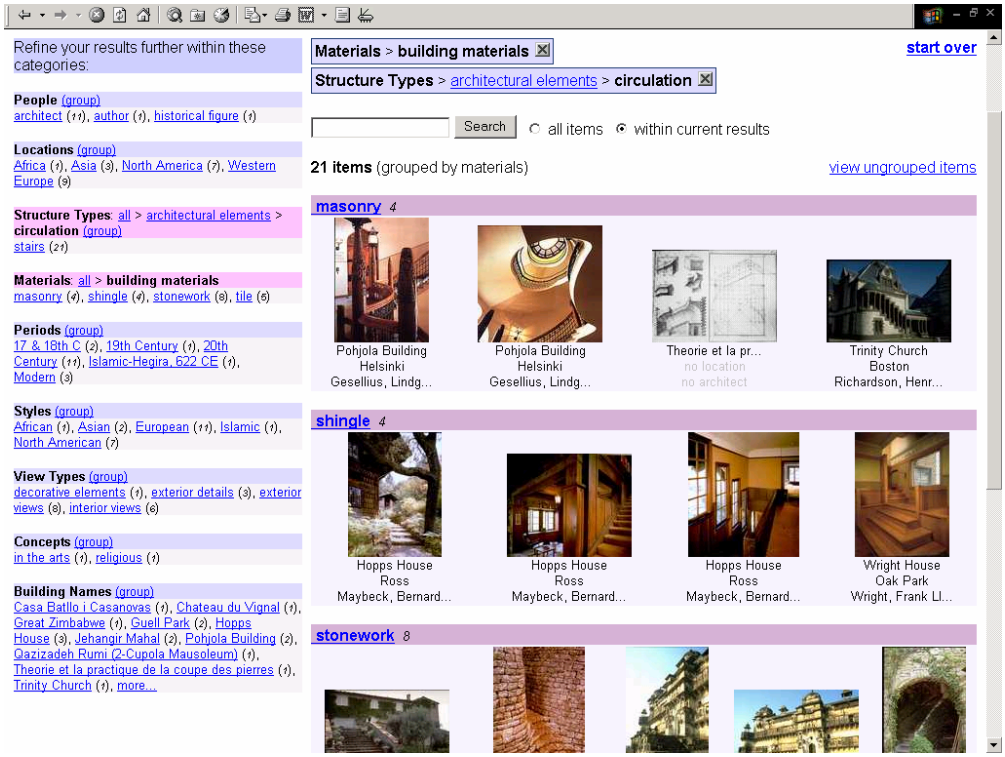

Figure 5. The Flamenco interface permits users to navigate by selecting from multiple facets. In this example, the displayed images have been filtered by specifying values for two facets (Materials and Structure Types). The matching images are grouped by subcategories of the Materials facet's selected Building Materials category.

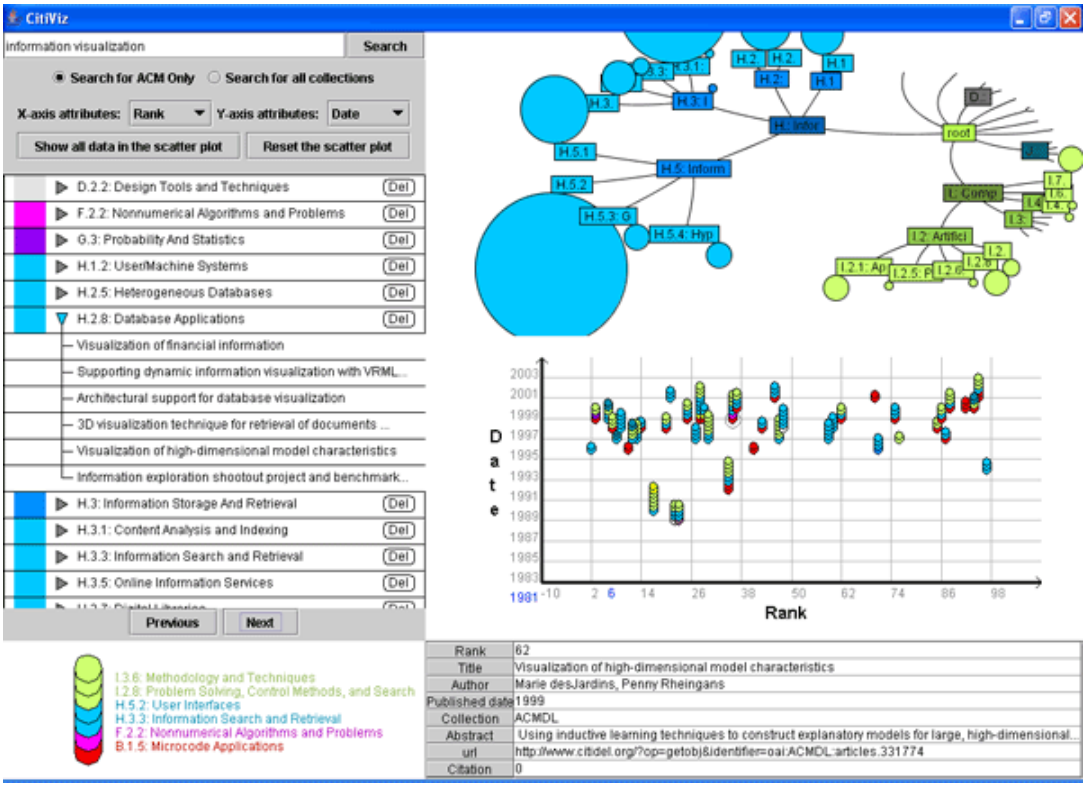

Figure 6. The CitiViz search interface visualizes search results using scatterplots, hyperbolic trees, and stacked discs. The hyperbolic tree, stacked disks, and textual list on the left are all based on the ACM Computing Classification System.

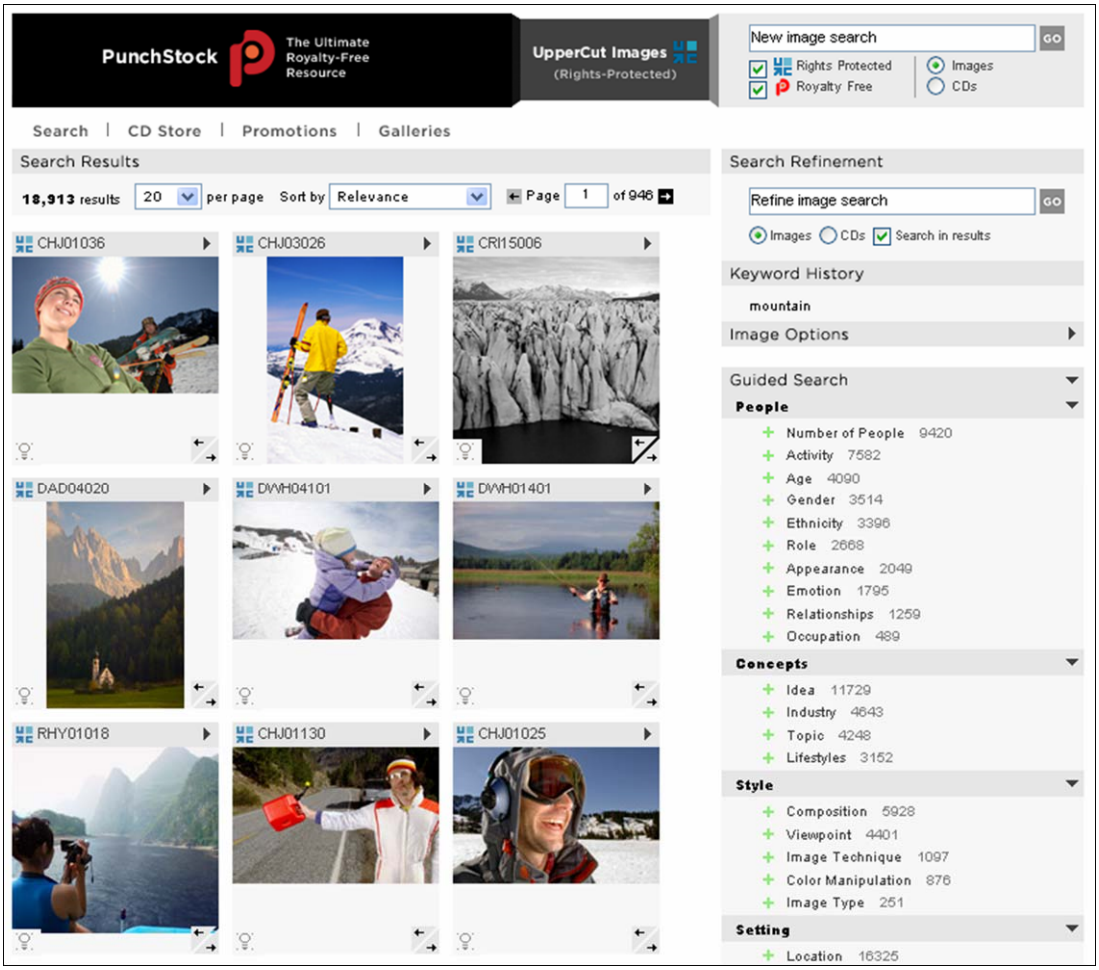

Figure 7. The PunchStock photo search interface provides categorized overviews of photo search results.

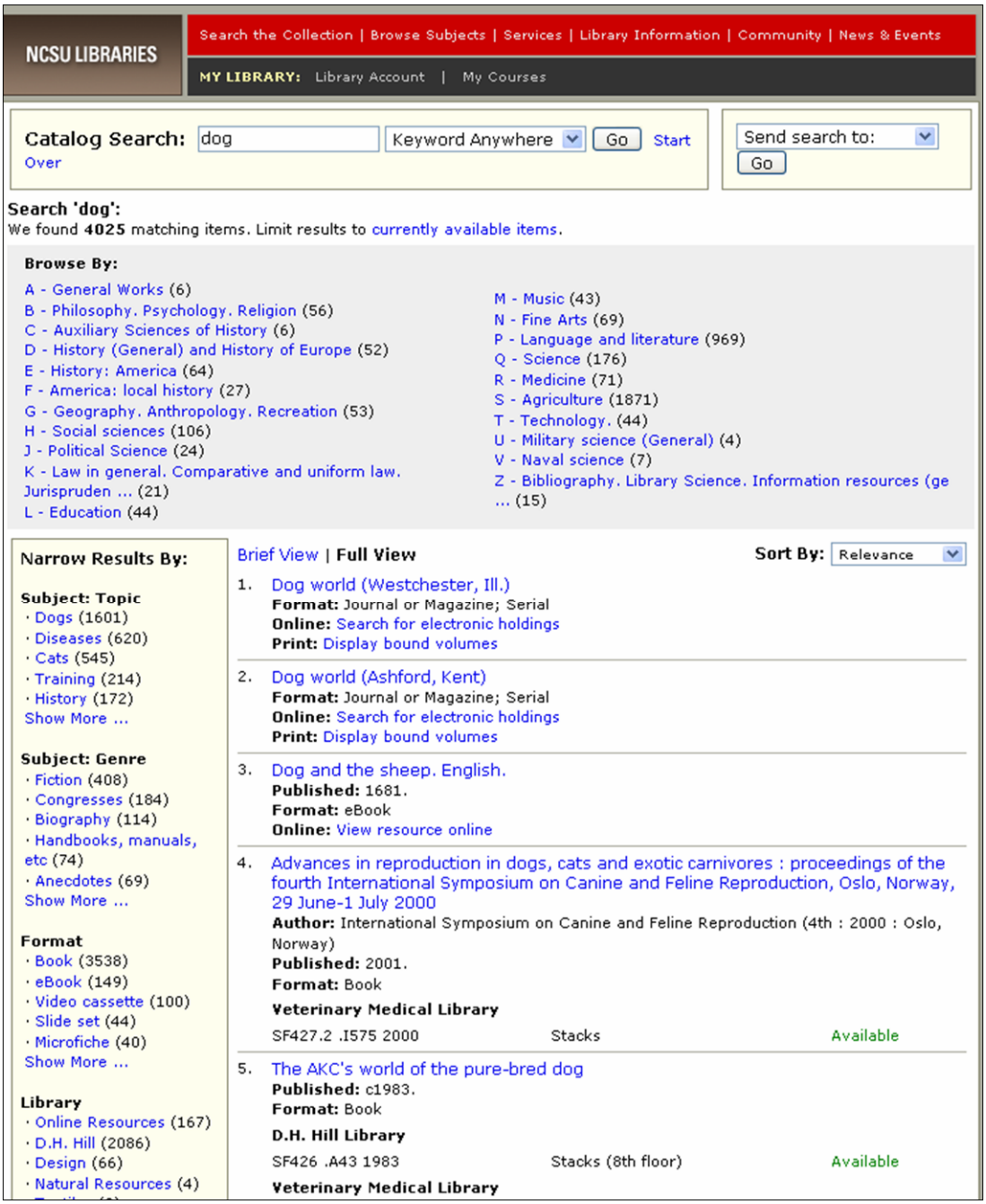

Figure 8. The NCSU library catalog provides categorized overviews of search results using subject headings, format, and library location.

2.2 Using categories for information retrieval

The field of Library and Information Science (LIS) has an established history of research in classification systems and their use. The emphasis within LIS on human

information behavior and information seeking has traditionally informed the development of classifications for libraries, archives, and museums. Faceted classification (Vickery, 1960), which is of particular interest in this dissertation, has influence beyond the LIS world, with human-computer interaction (HCI) researchers adopting its methods to support exploration and retrieval in large digital document collections.

For exploratory searchers, categories drawn from classifications, taxonomies, ontologies, 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.

In this dissertation, the interest is not in how classifiers work (e.g., machine learning), but simply that they provide a way to identify category membership for search results.

2.2.1 Studies of categorized overviews for web search

Meaningful and stable categories have been found beneficial for the organization of web search results in the limited 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 organizes intranet search results by an automatically generated web site overview (Figure 9). Preliminary evaluations were mixed, but promising, particularly for what users considered “hard-to-find information” (Chen, Hearst, Hong, & Lin, 1999). The WebTOC system (Figure 10) provides a table of contents visualization that supports search within a web site, although no evaluation of its search capability has been reported (Nation, Plaisant, Marchionini, & Komlodi, 1997). WebTOC displays an expandable/collapsible outliner (similar to a tree widget), with embedded colored histograms showing quantitative variables such as size or number of documents under the branch.

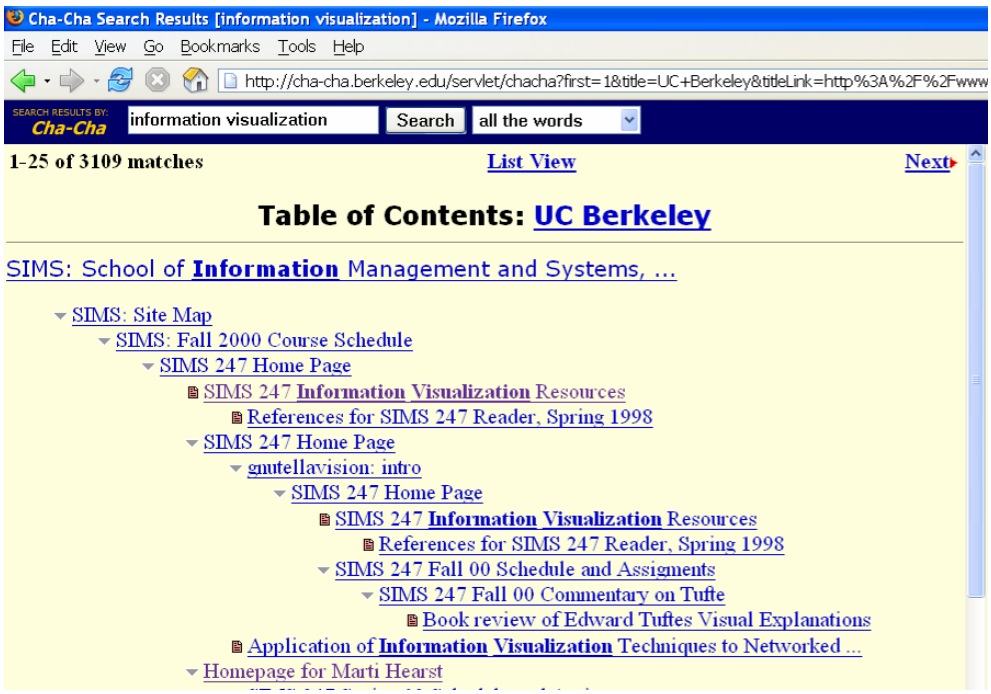

Figure 9. The Cha-Cha system organizes intranet search results by an automatically generated web site overview.

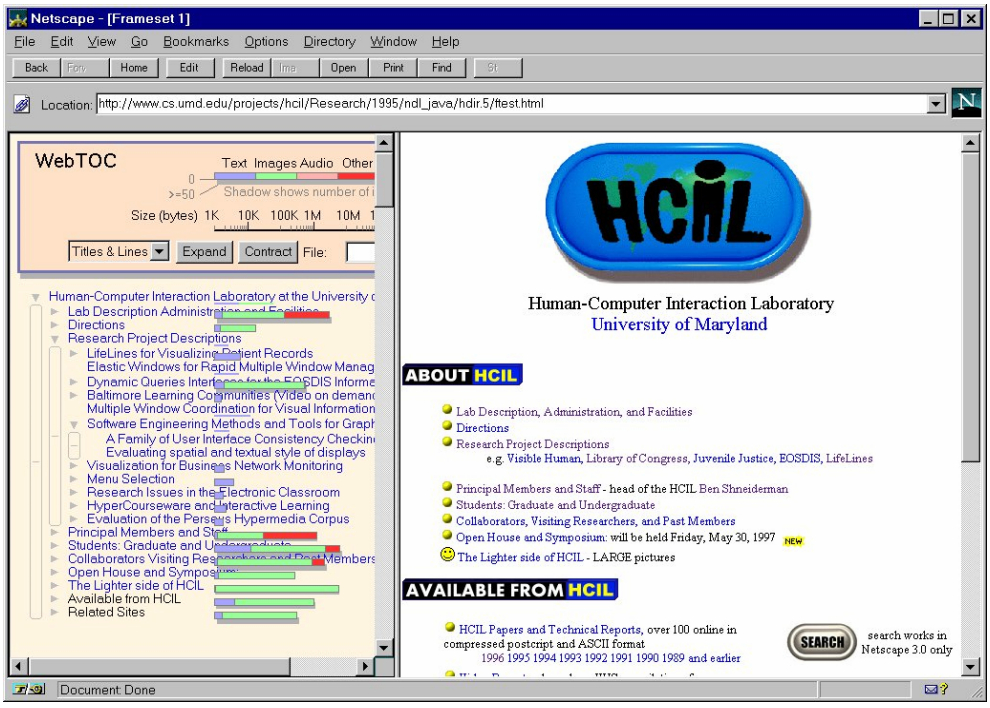

Figure 10. The WebTOC system provides a table of contents visualization that supports search within a web site.

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 a 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). Clusty (www.clusty.com) uses this technique to produce an expandable overview of labeled clusters (Figure 11). 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). As Rivadeneira and Bederson observed, web search results lack "1)… a natural spatial layout of the data; and 2)… good small representations," which makes designing effective visual representations of search results challenging. Using visual structures built around meaningful classifications may ameliorate this problem, as illustrated by promising interfaces like WebTOC.

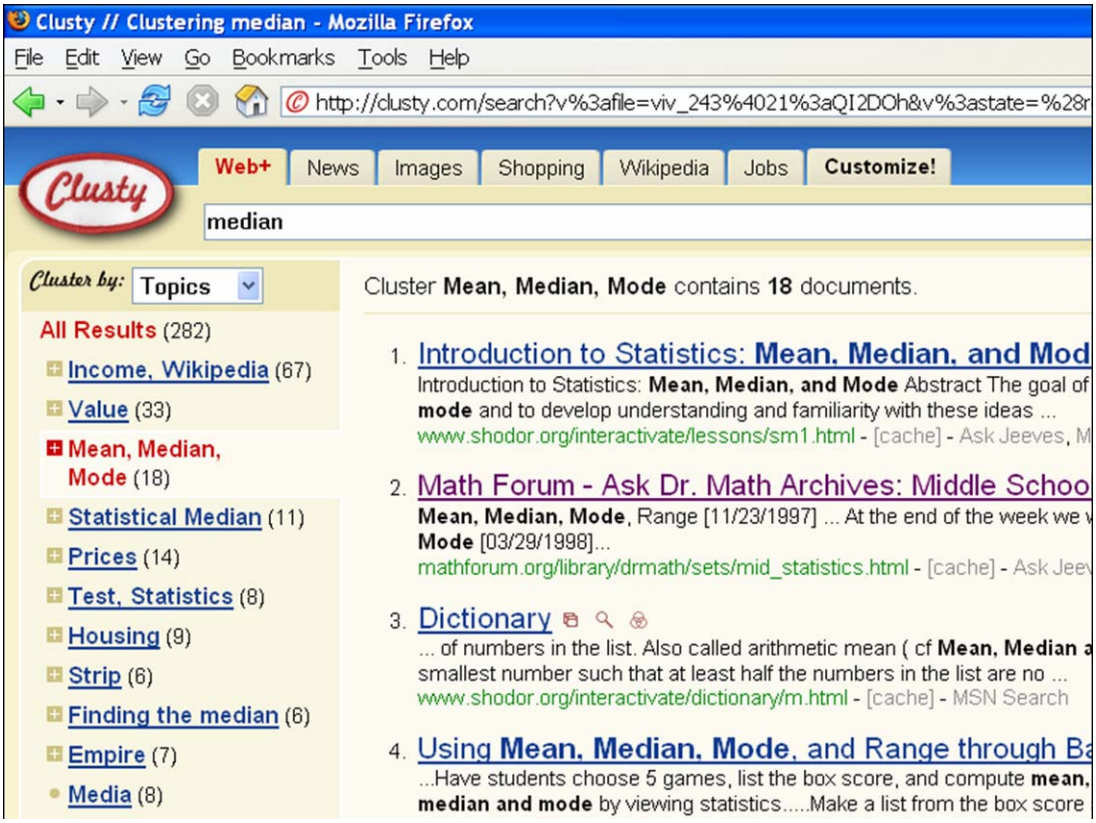

Figure 11. The Clusty metasearch engine uses automated clustering to produce an expandable overview of labeled clusters.

2.2.2 Other studies of categorized overviews for 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, Swearingen, Li, & Hearst, 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 (Figure 12) 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, allowing users to drill down for details (Shneiderman, Feldman, Rose, & Grau, 2000). The List and Matrix Browsers provide similar functionality, again using linear and grid-based displays (Kunz, 2003). Informal evaluations of these two interfaces have been promising, although no extensive studies of the techniques have been published.

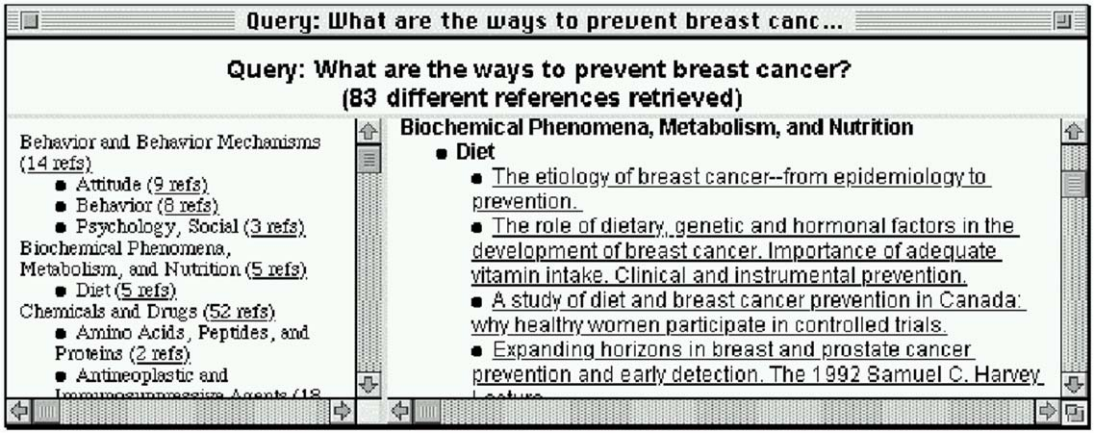

Figure 12. The Dyna-Cat system organized medical search results by a taxonomy of question types.

2.3 Visualizing and interacting with search results

The most common presentation of search results is the textual list, typically showing document titles and a few other pieces of information such as author, URL, a snippet of text (possibly with matching query terms highlighted). The results can be ordered by a computed relevance rank or by other attributes such as date, author, organization, etc. Drori and Alon (2003) compared four textual lists based on permutations of two variables (document category and lines from the document) in a 2x2 arrangement. Results were presented with and without categories, and with either the first lines of the document or the first lines relevant to the query. They found that the interface with categories and query-relevant lines from each document produced the fastest performance and was preferred by subjects. Dumais, Cutrell and Chen (2001) studied the effect of grouping results by a two-level category hierarchy and found that grouping by a well-defined classification speeds user retrieval of documents. Northern Light (www.northernlight.com), a commercial search engine, provides such a capability by grouping results in their Custom Search Folders. Exalead (exalead.com) organizes search results according to categories in the Open Directory Project. Other categories, such as organization charts, and geographic and temporal hierarchies, can also be used to organize search results.

The success of search result visualization has been mixed. Several web search (or metasearch) engines, including Grokker (www.grokker.com), Kartoo (www.kartoo.com), and FirstStop WebSearch (www.firststopwebsearch.com) incorporate visualization. Grokker clusters documents into a hierarchy and produces

an Euler diagram, a colored circle for each top-level cluster with sub-clusters nested recursively (Figure 13). Users explore the results by “drilling down” into clusters using a 2-D zooming metaphor. It also provides several dynamic query controls for filtering results. Unfortunately, this interface has been found to compare poorly with textual alternatives (Rivadeneira & Bederson, 2003). The authors found that the textual interfaces were significantly preferred. Kartoo, a metasearch engine, generates a thematic map from the top dozen search results for a query, laying out small icons representing results onto the map. When the pointer is placed over a document icon, arcs are displayed from that document to each relevant theme on the map. When the pointer is placed over a theme on the map, arcs are displayed to the related documents. This Flash-based alternative to search results is eye-catching (they offer a similar HTML-based version, too), but its utility is not clear. FirstStop WebSearch optionally displays collections of thumbnails instead of textual lists as part of a desktop-based search appliance.

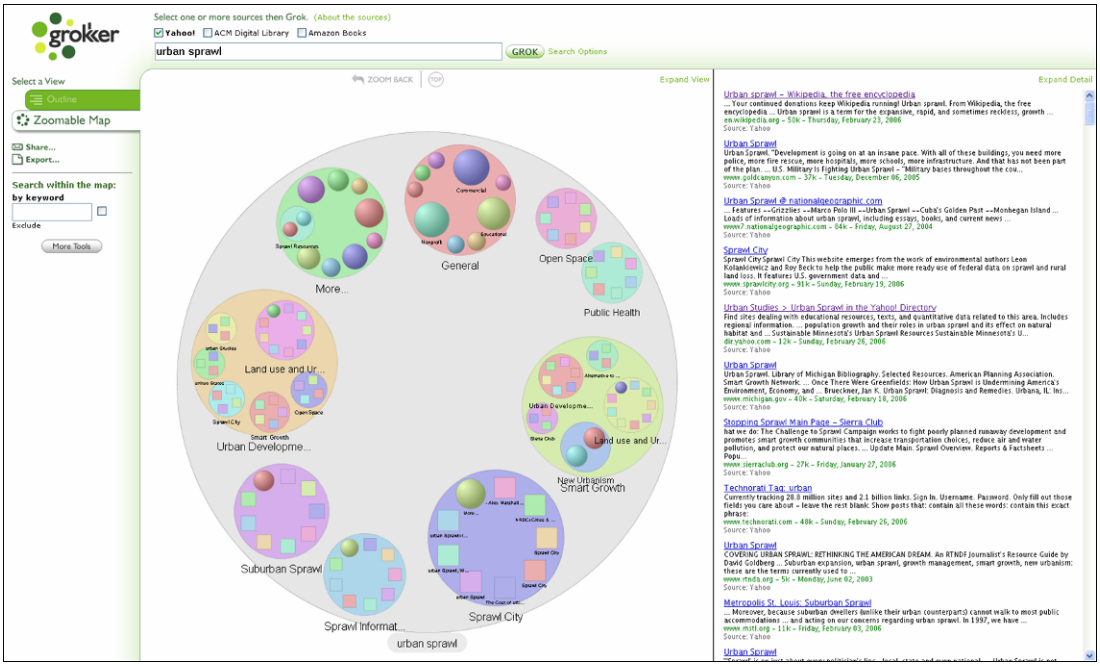

Figure 13. Grokker clusters documents into a hierarchy and produces an Euler diagram, a colored circle for each top-level cluster with sub-clusters nested recursively.

Figure 14. Kartoo generates a thematic map from the top dozen search results for a query, laying out small icons representing results onto the map.

The WebTOC and GRiDL prototypes display search results using hierarchical categories, allowing users to drill down for details (Nation, Plaisant, Marchionini, & Komlodi, 1997; Shneiderman, Feldman, Rose, & Grau, 2000). WebTOC displays an expandable/collapsable tree browser/outliner, with embedded colored histograms showing the number of documents under the branch and their sizes. GRiDL uses a grid to display two categorical attributes of a collection of documents. Each row/column of the grid represents a value for one of the categorical attributes. For each cell, if there are fewer than about 50 documents with that combination of values, each document is represented as a colored dot, where colors indicate a third categorical variable. If there are too many documents to fit into the cell, a histogram shows the distribution of documents across the third variable. More recently, outliner and matrix displays have been used to show search results, categorized into an ontology-based classification (Kunz & Botsch, 2002). SuperTable (Klein, Müller, Reiterer, & Eibl, 2002) integrates several information visualization techniques, including a scatterplot, TileBars (Hearst, 1995), and a bargraph, using linking and brushing to coordinate multiple tiled windows. Informal evaluations of these interfaces have been promising, but no extensive studies of the techniques have been published.

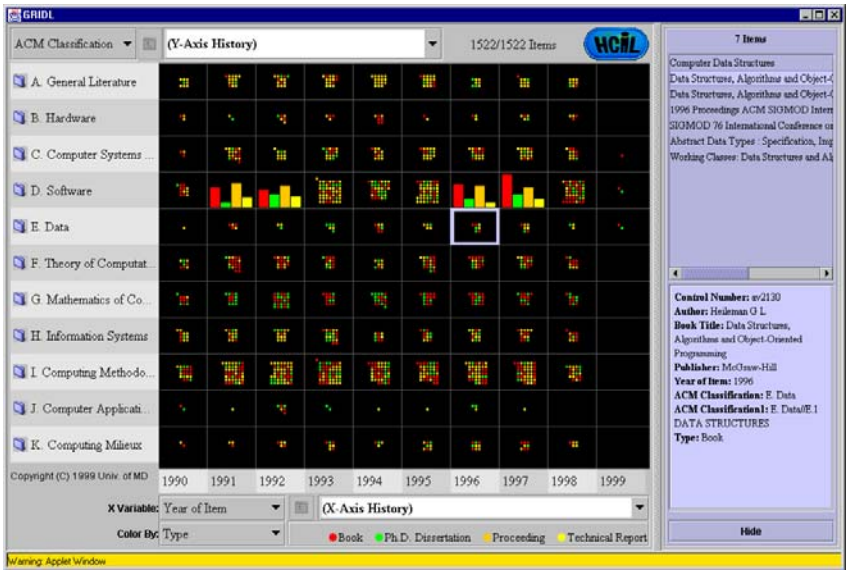

Figure 15. This GRiDL example shows search results organized by the ACM classification and date.

Evaluations often indicate that interface effectiveness is dependent on the specific information-seeking task. Risden, Czerwinski et al. (2000) compared a standard collapsible tree browser, a 2D textual layout (similar to Yahoo!) and a 3D interface for tasks that involved finding or creating categories of content in a web site. The 3D interface produced significantly faster performance when finding existing categories, but not when adding new categories. The authors speculate that the accessibility of context information in the 3D interface (not available in the other interfaces) may have been more beneficial for the finding task than the creation task. Sebrechts, Vasilakis et al. (1999) compared text, 2D, and 3D visualizations of clustered search results, finding that overall, the text was fastest and 3D was slowest, although for experienced users 3D was faster. They also found reliable differences in response time by the interaction of task type and interface, concluding that the match between

visualization features and tasks was more important than the dimensionality of the visualization. A comparison of information retrieval systems from TREC-6 found similar results (Swan & Allen, 1998). Kleiboemer, Lazear et al. (1996) found graphical displays to be more difficult than text, and Chen, Houston et al. (1998) suggest that the simple labels provided by Yahoo! were more useful for navigating a document space than a Kohonen map. Becks, Seeling and Minkenberg (2002) found document maps to be successful for tasks requiring detailed structural analysis of document inter-relationships, but also noted that users wanted to see more text, tightly coupled to the display, or another expressive arrangement of clusters.

2.4 Summary

Using categories to organize and explore general web search results is a promising but unproven technique (Hearst, 2006). 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. With the growing use of categorized overviews for search results, there is a need for design principles for more open, exploratory search interfaces that are based on a firm theoretical and empirical foundation. This dissertation addresses these issues.

Chapter 3: Early designs and formative studies

This chapter describes early user interface designs for the SERVICE system, and reports on two formative studies conducted with categorized overviews that used United States (US) government agencies and departments as meaningful and stable categories. The purpose of the studies was to illuminate searchers' use of categorized overviews to explore and understand search results. This would help to refine the emerging principles and analysis (both described in Chapter 4). The research goals motivating these studies include:

1. Identifying search tasks and sub-tasks that benefit from categorized overviews
2. Understanding how the visual presentation of the overview affects its utility
3. Understanding how the categories used for the overview affect its utility and the user's search experience

Study 1 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. Study 2 investigated the effect of two different kinds of categories. One search interface used the government organizational hierarchy and the other used Vivisimo's automated clustering. The information seeking tasks used in these two studies were motivated by work with government agencies through the GovStats project (Ceaparu & Shneiderman, 2004; Hert, 2002; Kules & Shneiderman, 2003). In this domain, web sites such as FirstGov (www.firstgov.gov), FedStats

(www.fedstats.gov), Science.gov (www.science.gov), and other specialized search engines provide some help for searchers. FirstGov has recently launched a search tool that incorporates Vivisimo's automated clustering technology to provide clustered overviews of search results, but to my knowledge no search engines currently provide overviews of search results categorized by government agency. Studies have found that queries for governmental information comprise 1.5%-3.0% of all queries to general web search engines (Jansen, Spink, & Pedersen, 2005; Spink & Jansen, 2004), suggesting that this would be a useful niche to study.

This chapter first presents early designs and the prototypes used for the two studies. The study designs and results are presented in sections 3.3 and 3.4, followed by a discussion of the findings and limitations of both studies in section 3.5.

3.1 Early designs

Early designs helped to define the design principles. They explored graphical approaches to display overviews of search results. Treemap displays used the leaf nodes (boxes) to represent items (individual web pages) in a thematic hierarchy (Figure 16, Figure 17), and government agencies (Figure 18, Figure 19). These displays effectively showed the distribution of results across categories and highlighted unusually placed results. An alternate mock-up (Figure 20) showed search results as red markers on vertical bars that represented categories. One bar was displayed for each category. The placement of the marker indicated the rank of the results within the entire list, with the highly ranked documents at the top of the bar. The vertical bars could be expanded horizontally to display the title and snippet for

the top 2-3 results within that category. Up to two categories could be expanded at a time. This design showed the distribution of results across categories along with the rank of each result, and embedded the text of the top 5-10 results in the overview. It allowed comparison of results between two categories.

The visual overviews were promising during informal reviews with professional colleagues and fellow students. The larger-than-usual number of results, the meaningful categories (thematic and government agency-based), and the color-coding were appreciated for their ability to provide a visual overview of the search results. The reviews also highlighted the importance of retaining the title, snippet, and URL in a textual list of results, simultaneously visible on the screen. Users wanted to read the text for details while they looked at the overview.

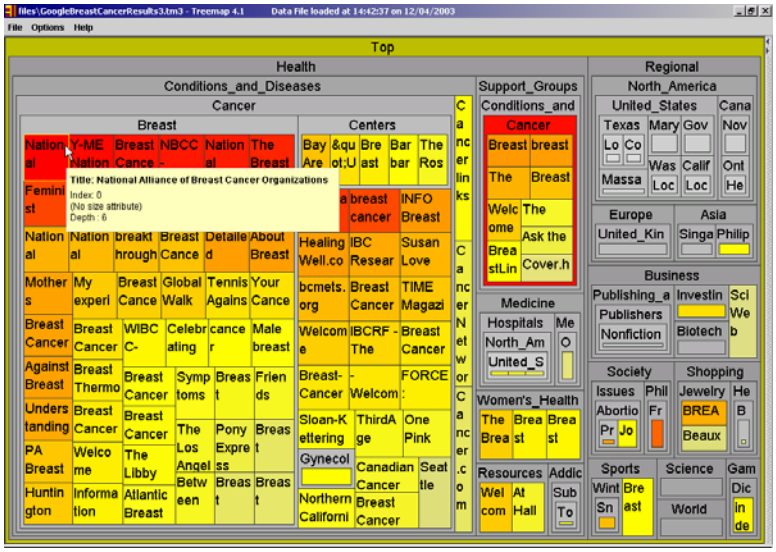

Figure 16. This treemap shows 157 search results for the query "breast cancer" encoded as leaf nodes in a broad and deep thematic hierarchy. The leaf nodes have constant size, so it is easy to see that most results fall under the Health top-level category. The bright red nodes (which appear as dark gray when rendered as a gray-scale image) are highly ranked, while the orange and yellow nodes are ranked lower. This makes it easy to see that there is at least one moderately ranked page in the Society category.

Figure 17. Zooming into the Society category provides previews of the three web pages falling in that category.

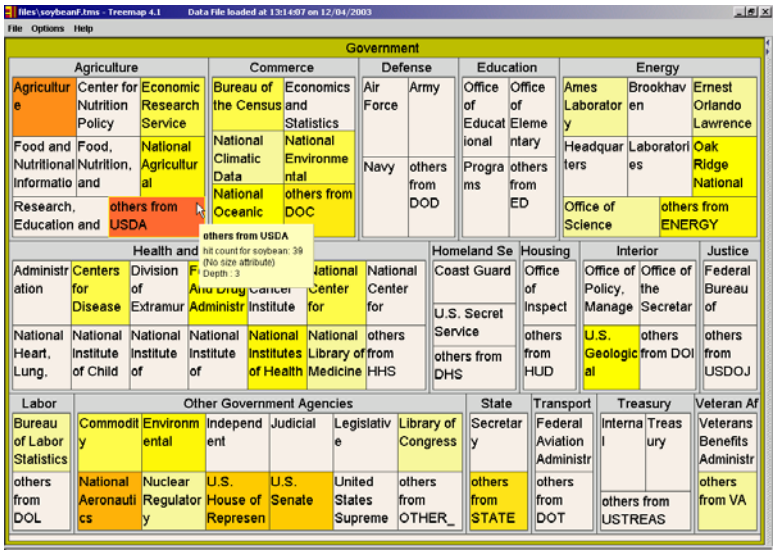

Figure 18. The top 200 search results for the query "soybeans" in government agency web sites is shown as a treemap. Each node represents an agency. The color coding shows that most results are from the Department of Agriculture, but the National Aeronautics and Space Administration (NASA), the House of Representatives, and the Senate all yielded many results, too. Leaf node size is constant.

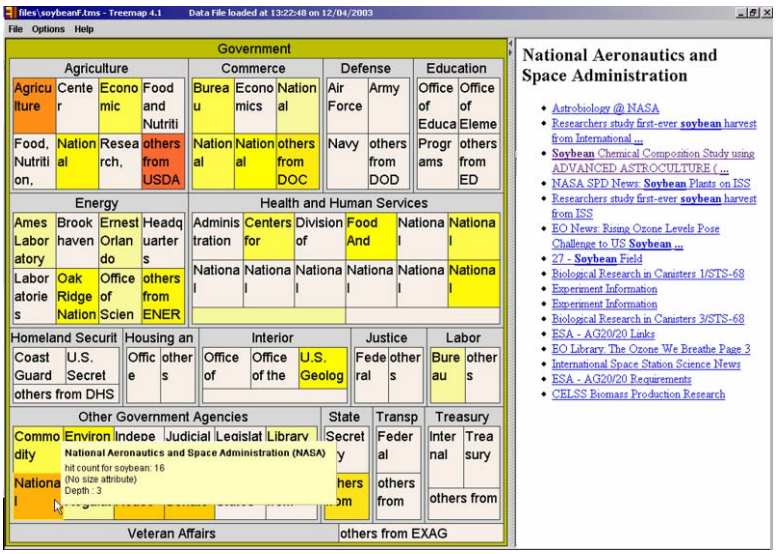

Figure 19. Clicking on the NASA node displays a text list of the search results from that agency.

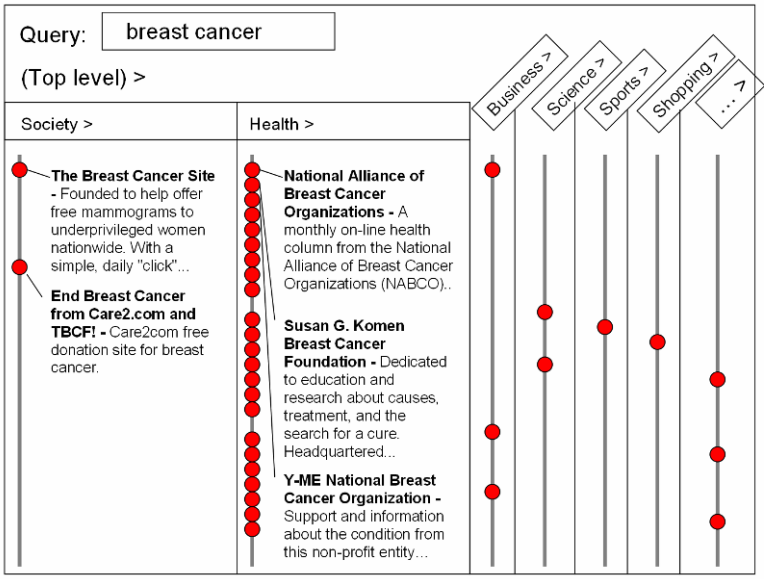

Figure 20. In this mock-up, the top 40 search results from the query "breast cancer" are organized by thematic categories and represented as red markers on vertical bars for each category. Two of the categories (Society and Health) are expanded horizontally to show the top results in those categories. The other categories are collapsed, showing just the bars and markers to indicate the number of results and their ranks within the entire list of results.

3.2 Formative study prototypes

These two early studies organized a pre-computed set of search results from government web sites into a two-level hierarchy of departments and agencies. The U.S. federal government organizational hierarchy was used 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.

Two forms of the categorized overview were prototyped: an expandable outliner and a treemap. Based on feedback on the initial designs, the overview was paired with a Google-style ranked list of search results. This provided the title, snippet, and URL in a form suitable for efficient skimming and scanning. The overview was tightly coupled with the list so that clicking on a node in the overview filtered the list results to show results from that category. Both 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.

Figure 21. 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 22. 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 21.

3.3 Study 1: Expandable outliner vs. treemap vs. control

3.3.1 Research questions

This study investigated the first two research goals listed above:

1. Identifying search tasks and sub-tasks that benefit from categorized overviews
2. Understanding how the visual presentation of the overview affects its utility

It used a constant categorization, thus the effect of different categorizations was not examined. For the visual presentation of results, an overview+detail approach was consistent with the initial principles. Three common exploratory search tasks were identified:

- 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 addressed three research questions:

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

3.3.2 Experimental conditions

The study compared presentations of search results with and without categorized overviews using pre-specified queries and a fixed set of search results. The U.S. federal government organizational hierarchy served 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 the organizational hierarchy is strictly a tree and not a hierarchy as defined by Kwasnik (1999) because it does not implement the *is-a* relationship or inheritance, 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 23) 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 (Figure 21), or a treemap (Figure 22). Both allowed 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. The expandable outliner additionally allowed participants to display or hide the counts of results in parentheses after each category, although this

was not used in the experiment. Both quantitative and qualitative data were collected. Preliminary results were reported in Kules & Shneiderman (2004).

Figure 23. The control condition mimics a typical set of Google search results, adding the government department and agency.

3.3.3 Hypotheses

In addition to collecting qualitative data, this study tested three hypotheses, based on the initial design principles for exploratory search interfaces:

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.

3.3.4 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. It was also desirable 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 organizational and cultural 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 enabled the use of 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, I chose to expose participants to three diverse scenarios and collect a wider range of data, rather than a tailored scenario advocated by Borlund.

The scenario content was motivated by work on the challenges of finding government information and publications (Ceaparu & Shneiderman, 2004; Kules & Shneiderman, 2003; Marchionini, Plaisant, & Komlodi, 1998). The GovStats project's 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 unusual results in Task C were interpreted by participants, with respect to individual results or the entire set of results. The tasks were time-limited to permit completion of the session within approximately one hour.

3.3.5 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 think-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 make a good start on each task.

3.3.6 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 formative

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.

3.3.7 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. Participants made individual determinations of what was unusual. After the sessions, the perspectives and unusual items identified were reviewed, along with the comments of participants and the observer's notes.

3.3.7.1 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 measured 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 in the mean total scores, $f(2, 15) = 6.74$, $p = 0.008$. 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 a meaningful categorical grouping would benefit users for this task.

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)

3.3.7.2 Perspectives found

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

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 Outlier group provided only two perspectives for the Breast Cancer scenario.

Position of perspectives found – For each scenario, I determined the positions (rank) of the pages from which each identified perspective was drawn and computed the median value (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, $f(2, 146) = 17.10$, $p << 0.01$. 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, I computed the mean number of categories selected during the task (Table 4). Note that no top-level categories were selected within the treemap. I 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. Random clicking could have had a similar effect.

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)

3.3.7.3 Unusual results task

The number of participants who found something unusual for each condition and scenario was counted (

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, the number of times participants identified unusual items was counted. The full tables for each scenario are in Appendix B. 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, 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. Since this was not anticipated, I reviewed 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. However, nine of the 12 overview participants at some time found this surprising. From participant comments, it appears that the display of agencies with zero results and the color coding contributed to the searchers making such observations.

3.3.7.4 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 semantic differentials 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, higher values indicate higher satisfaction ratings. Question 4 was originally written with a value of “9” meaning the most difficult and is reversed for presentation here. The values have been adjusted to reflect this reversal.

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. Signifiant differences are shown in bold.

				ANOVA		
	Control	Expandable Outliner	Treemap	df	F	sig
1. Under-standable	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. Easy	**4.50 (0.55)**	**7.67 (2.34)**	**7.00 (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. Clearly, users with an overview were more satisfied.

3.3.7.5 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 commented on using 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 satisficed 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 noticed that an agency had no results, and he expected it to, he would look at the uncategorized results. Since the result set had 200 results total, the ability to filter out the 130 results that were categorized into known agencies would allow him to focus on the remaining 70 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 [uncategorized] 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.

The results from this study are discussed in section 3.5, in conjunction with the results from the second study.

3.4 Study 2: Automated clustering vs. government hierarchy

3.4.1 Research questions

The second study focused on the first and third research goals listed at the beginning of this chapter:

1. Identifying search tasks and sub-tasks that benefit from categorized overviews
3. Understanding how the categories used for the overview affect its utility and the user's search experience

It varied the categories and used a single display style, an expandable outliner, for both conditions.

The emerging principles (described in section 4.2) asserted that overviews should be organized by meaningful, stable classifications, but the overviews built with dynamically generated categories used by clustering search engines (e.g. Vivisimo) have been found helpful, even though participants sometimes fail to understand the clusters or their labels. This motivated investigation of how automatically clustered overviews supported user examination of search results. For this study, two new tasks were identified: idea generation and resource finding. These more complex exploratory search tasks were refinements of the tasks used in study 1, and allowed me to explore different search tasks (research goal 1). This study addressed three specific research questions with a combination of observation and questionnaires:

1. What differences can we observe in how participants examine search results with respect to domain and classification knowledge when they use overviews based on dynamic categories (automated clustering) vs. overviews based on stable categories (government hierarchy)?
2. What differences can we observe in how participants examine search results with respect to the type of search task when they use overviews based on dynamic vs. overviews based on stable categories?
3. What differences do participants perceive in their search processes and outcomes when they use overviews based on dynamic categories vs. overviews based on stable categories?

3.4.2 Experimental Conditions

A within-subject experimental design (N=12) with qualitative observation was used to address these questions. Two experimental conditions were used by each participant: Condition 1 used the Vivisimo search engine (Figure 24), as an example of an interface using dynamic categories to provide an overview. Vivisimo uses a form of automated clustering that generates hierarchies of concisely labeled clusters. The clusters are formed and labeled by finding common words and phrases in the titles and snippets. 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, however, because

a) the basic layout of results and interaction styles were consistent,

b) the study did not seek specific quantitative measures that would be affected by these differences, and

c) the 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 me to collect data from each scenario eight times over the course of the study.

3.4.3 Scenario and task design

As 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 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 3.4.6.1):

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 24. The Vivisimo search engine was used for the clustered hierarchy condition.

3.4.4 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. Due to the small sample size and formative nature of the study, statistical significance was not analyzed. 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. The instructions were clarified so that participants would avoid Vivisimo's sponsored links and the "Find in clusters" feature, which was not available in the government hierarchy interface.

3.4.5 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 to see if their experience would differ from others.

3.4.6 Results

3.4.6.1 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). Differences between the two conditions were slight, not more than one point on the nine-point scale, and variance was high.

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	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 respondent found good resources		0.75 (1.54)
Q10. Confidence that respondent 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, I 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.

3.4.6.2 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 the study did not quantitatively measure this. Problems were 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 was expanded.

The following section discusses the results from this study in conjunction with the results from the first study.

3.5 Discussion of studies 1 and 2

These two studies began to answer the research goals posed at the beginning of this chapter and suggested additional insights. They showed that categorized overviews of the top 200 search results could be useful for the selected tasks. They also showed benefits and drawbacks of the dynamic categories. They corroborated several of the emerging principles (section 4.2) and entailed revisions to others, as discussed in the following sub-sections.

3.5.1 Benefits of categorized overviews

As expected, study 1 confirmed that the categorized 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 the premise that the categorized overview interfaces are seen as simple, understandable and easy to learn (i.e., hypothesis 3 of study 1 was not supported). For the treemap interface, this conclusion is qualified by noting that participants were provided brief training in the use of the treemap.

During the perspectives task in study 1 ("Find 3 web pages providing different aspects of or perspectives on this topic"), participants found their perspectives significantly deeper in the ranked list of results. This result is consistent with results reported in Käki (2005), that searchers viewed pages deeper in the results. It provides quantitative evidence that the categorized overviews also helped searchers *find* relevant and useful pages deeper in the results. 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. I 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. This can be attributed to the use of the meaningful and comprehensive hierarchy, which allowed users to make inferences and draw conclusions. In all of the experimental sessions for study 1, 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 explanation is supported by the participant comments and suggests that color coding might be more useful in the expandable outliner if used more extensively.

3.5.2 Effect of visual presentation of overviews

The appeal of both the expandable outliner and treemap presentation of overviews was confirmed by the lack of statistically significant differences between the expandable outliner and the treemap in study 1. 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.

3.5.3 Effect of categories 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 distinct nature of the categories probably contributed to differences in which tasks each was preferred for. Some participants appreciated the dynamically generated hierarchy for the ideas task. Its statistically based clustering yielded labels that they found suggestive of topic ideas. The labels of the dynamic categories were drawn from the titles and snippets in the results, and may have been more suggestive of themes. Some participants felt strongly that the government hierarchy helped them explore and understand the results more effectively. The labels in the government hierarchy indicated the provenance, or source, of web pages. The inclusion rules were more transparent and predictable to users for the government hierarchy than for the Vivisimo hierarchy, permitting more reliable inferences. Based on the results of study 2, one emerging design principle (originally “Organize results by meaningful, stable classifications,” in section 4.2.2) was revised 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 another design principle (Support multiple visual presentations and classifications), suggesting that the faceted category approach (Yee, Swearingen, Li, & Hearst, 2003) could be beneficial for organizing web search results. Participant comments suggested that there may also be value in personally-created or customizable taxonomies.

3.5.4 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 confidence in a third principle (Arrange text for scanning/skimming).

3.5.5 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 result 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.

3.5.6 Limitations of these studies

These studies were formative in nature, and the results must be interpreted within the context of the specified tasks and domain. They employed a small sample of subjects, who were presented with pre-defined scenarios, queries and tasks. The presentation of the categorized overview and results in study 2 was not strictly equivalent. 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 they 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), the immediate scope of the findings was limited, in return for gaining a deeper understanding of how searchers used categorized overviews within that domain.

3.5.7 Summary of studies 1 and 2

The results of these two formative studies suggested answers to the 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, valuable for different tasks. 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. The studies were used to refine the emerging design principles. They raised the question of which tasks are best supported by stable categories versus dynamic categories.

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 collection of a rich set of observations about how searchers use categorized overviews of search results.

Chapter 4: Analysis, principles, and design of the SERVICE system

This chapter presents the three main contributions of this dissertation: an analysis of categorized overviews (section 4.1), design principles for exploratory search interfaces (section 4.2), and the architecture and design process of the SERVICE system (sections 4.3 - 4.6). Although presented linearly here, they evolved in an interwoven, iterative manner. The results of the two early studies, described in Chapter 3, informed the design of the SERVICE system. They also helped refine the emerging analysis and design principles. The design process was informed by the analysis and design principles, and in turn these were challenged and refined by the design process. Each of the three was influenced by the process of developing and refining the other two. The third and final study, described in Chapter 5, helped validate elements of the analysis and principles, and suggested limitations that continued the iterative process of refinement.

4.1 Analysis of categorized overview use

The purpose of this analysis is to explain how categorized overviews can change the way searchers comprehend and interact with their search results. This helps to justify the design principles and ground the SERVICE interface design in a principled theoretical base. This analysis is applicable to the form of categorized overviews studied here, specifically the use of the categorized overview presented simultaneously with a list of search results. It is focused on one activity in the search process (examining search results) and one form of interface (categorized overview).

It is presented as one step in understanding how exploratory searchers conduct their searches, with the hope that it will be useful as a framework for more ambitious theoretical analysis.

This section first presents a process model of exploratory search, and then identifies functional capabilities that categorized overviews provide and actions that they permit searchers to take. It describes how searchers can reason about search results using categorized overviews and tactics that they may adopt to take advantage of the overviews.

4.1.1 Process model of exploratory search

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 motivated and initiated by a higher-level work task (Byström & Hansen, 2002; Järvelin & Ingwersen, 2004). Work tasks are situated in the work organization and reflect organizational culture and social norms, as well as organizational resources and constraints. The work task is similar to Sutcliffe and Ennis' goal or information need, or Marchionini's recognition and acceptance of an information problem, but the work task specifically situates these in an organizational context. In the context of the work task, a second level of context is defined, in which information-seeking tasks are identified. These tasks vary as the work task progresses. The third level of context is the information retrieval context, wherein searchers identify sources, issue queries, and examine results.

The process model proposed here (Figure 25) combines the Marchionini model with the three-level Byström & Hansen model used in the formative studies (described in section 3.3.4). The model defines five activities: recognize an information need (to satisfy a work task), define an information-seeking problem (to satisfy the information need), formulate query, examine results, and view documents. It places activities in the context of the three levels of information-seeking and work tasks. It shows how search activities are sequenced within the iterative search process. Each higher-level activity can involve multiple subsidiary activities.

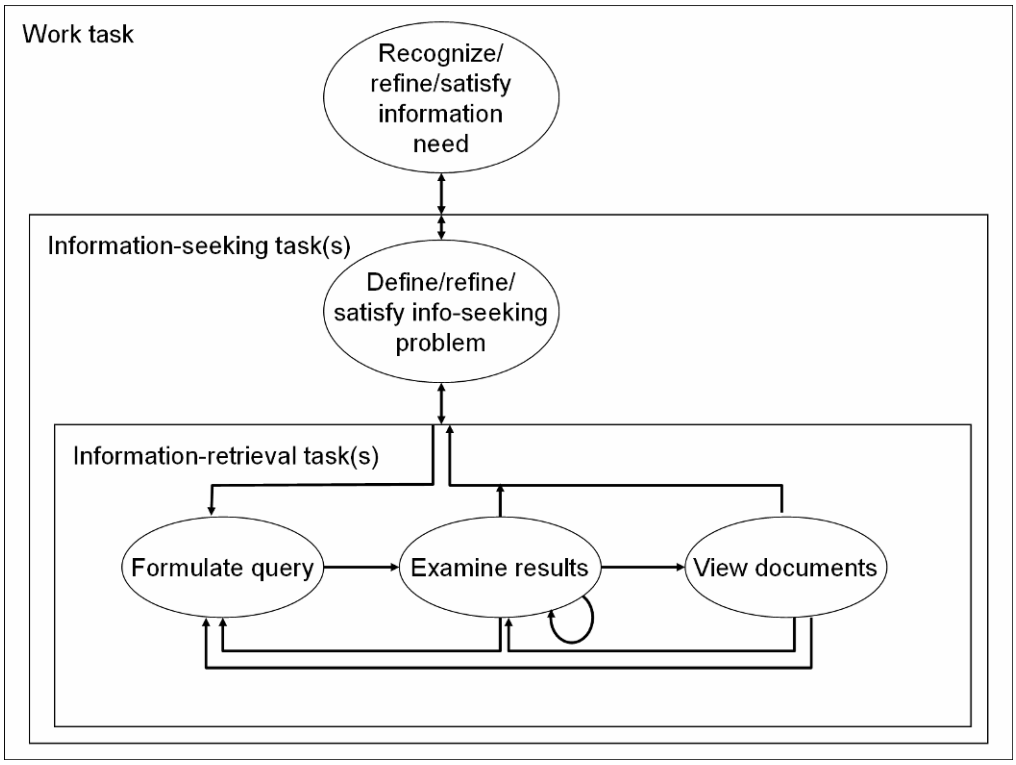

Figure 25. Process model of search in the context of work and information-seeking tasks.

The process is initiated when a searcher recognizes an information need and decides to try to satisfy it (Byström & Hansen, 2002; Marchionini, 1995). This need may arise because the searcher perceives a gap or anomaly in knowledge needed to satisfy an externally imposed work task (Belkin, 1980). To satisfy the information need, the searcher undertakes one or more information-seeking tasks, which can be structured as a linear sequence or hierarchical decomposition of tasks. For example, the paper writing process could be modeled as a series of stages (Kuhlthau, 1991), or a medical search could be modeled using a hierarchical decomposition of goals (Bhavnani & Bates, 2002). Each of these tasks requires selecting a source, and then engaging in one or more information retrieval tasks. Within each information retrieval task, the searcher formulates queries, examines results, and selects individual documents to view. As a result of examining search results and viewing documents, the searcher gathers information to help satisfy the immediate information-seeking problem and eventually the higher level information need. This model collapses Marchionini's source selection stage into the information-seeking problem. It also combines query formulation and execution. Reflection is inherent in each activity, and each activity except query formulation can return to a previous or higher level activity.

The strategies and tactics that searchers use are affected by the capabilities provided by the search interface (Bates, 1990; Golovchinsky, 1997). Strategies are high level plans for the whole search, and tactics are individual actions or sequences of actions (often called moves) taken to further the search (Bates, 1979; Marchionini, 1995). Searchers can take numerous actions while examining search results (Bates, 1990;

Fidel, 1985; Garcia & Sicilia, 2003; Marchionini, 1995; Shneiderman & Plaisant, 2004; Wildemuth, 2004). Specific actions supported by a web search interface can be discerned by analyzing the structure of text and hyperlinks on a search result page. For a typical search result page showing a ranked list of results, this yields the set of actions listed in Table 10. Each action involves cognitive and physical effort and can result in visual changes in the interface or changes in task, domain, or category knowledge (cognitive changes). The visual changes enable cognitive changes by making information visible on the screen. The cognitive changes are necessary to make progress on the information problem and are reflected in transitions between activities. For example, while examining results, searchers may scan a screen of results, causing them to identify additional query terms, causing a transition to the formulate query activity. In this analysis, actions that require visual scanning and/or moving the mouse without clicking are classified as low effort because they involve little physical effort, and they do not result in major changes to the display, thus minimizing cognitive effort. Actions such as selecting a result to view or scrolling the screen require a moderate amount of cognitive or physical effort because they require clicking and reorienting as the visual presentation changes. Issuing a new query requires a high amount of effort because of the cognitive effort required to formulate the query and the need to reorient when the new set of search results is displayed.

Table 10. Actions available to searchers when evaluating a typical search result list.

Action	Effort	Visual changes	Cognitive changes
Scan one screen of results list	Low physical; low-medium cognitive (depends on type of scan)	None	• Identify page to view • Assess results overall • Identify additional query terms • Refine information need • Refine information problem • Extract useful information
Scroll screen	Medium	Shift visible subset of search results	None
Select next or previous page of results	Medium	Shift visible subset of search results	None
Select a result to view specific web page	Medium	Bring web page into view	None
Reformulate query	High	Generate new set of search results	None
View specific web page	Variable	Variable	• Identify additional query terms • Refine information need • Refine information problem • Extract useful information

Adding a categorized overview to search results changes the information that is available and the actions searchers can take with low or moderate physical and cognitive effort. The categorized overviews used in this research add the following design elements:

- The overview presentation – a visual or graphical representation of the categories represented by the search results
- Hyperlinks to narrow and broaden the displayed set of search results
- When the pointer is placed over a category, the corresponding search results in that category are highlighted and a pop-up window is displaying with a list of non-empty sub-categories
- When the pointer is placed over a search result, the corresponding categories of which that result is a member are highlighted

Table 11 summarizes the actions afforded by these design elements, the effort required, and their visual and cognitive effects.

Table 11. Additional actions available to searchers when evaluating search results with categorized overviews.

Action	Effort	Visual changes	Cognitive changes
Scan categorized overview	Low physical; low-medium cognitive (depends on type of scan)	None	• Identify category to consider • Assess results overall • Identify additional query terms • Refine information need • Refine information problem • Extract useful information • Assess match between categories and information need
Select category to narrow or broaden results	Medium	Filter visible results, limiting to members of selected category	None
Move pointer over result	Low	Highlight category membership	• Identify categories to consider • Assess results overall • Identify additional query terms • Refine information need • Refine information problem
Move pointer over category	Low	Highlight results in category (currently visible results only) and display subcategories	• Identify page to view • Assess results overall • Identify additional query terms • Refine information need • Refine information problem

4.1.2 Action: Scan categorized overview

Scanning an overview is a lightweight physical action if the complete overview is visible on the screen (i.e., no scrolling is needed) and the elements are arranged in a consistent manner, using linear lists, columns, or matrices (Teitelbaum & Granda, 1983). The cognitive effort will be low when the categories and their structure are familiar, but may be higher when first encountered or for unfamiliar knowledge domains. This impacts the knowledge that searchers can draw on to reason about the results and make inferences or predictions about meaning, authority, validity, relevance, and overall utility (Marchionini, 1995). Anderson (1990) argues that categories are ideally suited to supporting prediction. The category labels indicate statistical and conceptual relationships among members of a category, as well as distinguishing relationships between members of different categories (Markman & Ross, 2003). They limit the information people need to consider when making inferences (Markman & Ross, 2003), thus permitting reduced cognitive effort. This helps searchers to efficiently predict the utility of subsets of pages (i.e., the pages in a selected category) within the search results. For example, in the "median" scenario described in Chapter 1, the task was to find an age-appropriate description of the term "median" for a ten year-old. In that context, web pages in the Kids and Teens category would be very likely to be useful. Category information should help searchers assess their search results overall, assess the match between the categories and their information need, and identify subsets of the results to consider exploring. The category labels can be thought of as suggesting alternative "patches" of data within the results (Bates, 1989). The categories can also be considered to increase the

"information scent" (Pirolli & Card, 1995; Pirolli & Card, 1999) of pages that fall below the first screen of results.

Exploratory searchers may value novelty in their search results. Unusual results or patterns of results may be important. When searchers value novelty, categories that were expected (or not expected) to contain results can surprise searchers when they do not (or do). This can cause searchers to reflect on their queries, information needs, or information problems. It may also prompt additional questions. This was particularly notable in the first study, when users spontaneously commented on the absence of any "breast cancer" search results from the Department of Education. In the context of the third study, it also prompted users to think of additional story ideas, which they pursued by selecting the category.

Relationships used to predict relevance or utility may be based on belief or logic (Marchionini, 1995). They can be based on searcher experience, or even bias or prejudice. As one subject admitted during study 1, he used his opinions about the Department of Housing and Urban Development (HUD) and his affinity for the concept of independence to guide his information seeking (quote on page 64). Of course incorrect knowledge can lead to incorrect predictions of relevance or utility, and thus to poor choices. This was observed in the first formative study when, for example, a participant incorrectly thought that the National Aeronautics and Space Administration was an agency under the Commerce Department and filtered using the wrong category.

Stable categories enable a searcher to develop familiarity and reuse category knowledge. When a searcher first encounters a category, he or she is not merely evaluating the results with respect to the information need. The searcher is also assessing whether the results are consistent with their expectation of the category. The assessment affects the category knowledge, confidence, and intentions for future use. Cognitive load is higher for the first encounter, but less upon subsequent encounters. The searcher's understanding of specific categories grows with use. A poor first impression, such as when a category selection yields mystifying results, can discourage use, as was occasionally observed in the empirical studies.

How people interpret category labels, the meanings they infer, and ultimately how they use categorized overviews, depends on personal knowledge of a subject, past experience, and the immediate context of use (Jacob, 2004). This is true even for well-defined categories, like the US government agencies used in the formative studies. With hierarchically organized categories, like the Open Directory Project, the interpretation of a category label can be affected by its parent category and any child categories. This was notable in the third study, which truncated categories at the third level. This had the effect of removing valuable contextual information from the category label, and resulted in confusion about the contents of the category.

4.1.3 Action: Narrow or broaden by category

Selecting a category to narrow results restricts the displayed results to those that are members of the specified category. After being narrowed by a category, results may

be broadened, removing these restrictions. This action can be considered as a form of query reformulation (Golovchinsky, 1997) or view navigation (Furnas, 1997). Study participants' comments indicated both perspectives. This action requires moderate physical effort because the user must move the pointer and click on a link. It requires moderate cognitive effort because the user must reorient to the changed list of results.

4.1.4 Action: Move pointer over result

In the SERVICE web search prototype (described in section 4.6), moving the pointer over a result provides additional details about that result by highlighting any categories displayed in the overview that it is a member of. This action requires low effort. Because the overview might only be displaying an upper level of the category, this does not necessarily provide complete category information. For example, if a result were a member of the thematic category /Arts/Television/Networks/Cable/BBC, but the display was only showing top-level thematic categories (e.g., Arts, Business, Computers), the Arts label would be highlighted. The search interface could also open a pop-up window near the result with the complete category information, although this might be large and distracting when a result is a member of many categories.

4.1.5 Action: Move pointer over category

In the SERVICE web search prototype, moving the pointer over a category in the overview has two effects. First, it highlights any results visible on the screen that are members of the category. This can provide examples of the members of that category. Second, it opens a pop-up window with a list of the non-empty subcategories of that

category, providing a preview of the effect of clicking. This action requires low effort.

4.1.6 Tactics

The actions enabled by categorized overviews can lead to altered search tactics because they change the information available and the range of possible interactions. This allows searchers to draw on new tactics and revise old ones while reducing effort and/or improving outcomes. For example, studies have found that most searchers do not examine more than the first page of search results (Jansen, Spink, & Saracevic, 2000), suggesting the often observed tactic for evaluating search results. With the typical list, the searcher may scan 10-20 results, assessing their predicted utility for the task. With the addition of a categorized overview, the searcher can also scan the overview, using the categories to help predict the utility of the results that fall within those categories, as part of a single cognitive action. The categorized overview can typically show 20-30 categories and slightly reduces the number of results that can be displayed on a screen, typically by less than one result. This increases the amount of information that searchers can acquire within a limited time without appreciably raising their cognitive effort and with no additional physical effort (beyond eye movement). The use of these tactics does depend on searchers having an appropriate mental model of the categorized overview. Seven tactics evidenced in study 3 are shown in Table 12.

Table 12. Tactics enabled by categorized overviews.

Tactic	Description	Benefit
Broad queries	Type broader queries in the search box, with few terms, then narrow results using the categorized overview.	Reduced cognitive effort to generate the query.
Organize examination by overview	Use the categorized overview to determine the order in which result subsets are examined.	Helps monitor search to keep it on track and efficient.
Overview as backup	Examine the top portion of the list first. If not satisfied, examine the overview to identify subsets to examine.	May help when relevant documents are not at top of list.
Preview before narrowing	Examine the subcategory information before narrowing results to that category.	Avoids low relevance results. Improves confidence in expected results of action.
Assess result set	Scan categorized overview to determine what categories are represented and how results are distributed across categories.	Helps provide an overall understanding of the results of the query. May help assess the overall quality of the results and by implication the query.
Probe using categorized overview	Select specific categories and examine the results to assess subsets of the results.	Reduces effort compared to typing multiple queries.
Ignore	Ignore the categorized overview.	Avoids or simplifies decisions about actions to take.

4.1.7 Other impacts of categorized overviews

One question raised by the changes in search tactic described above is whether the visibility of the category labels biases the way that searchers assess their results. This could be a particular concern when there is limited metadata for categorizing search results because there may be a non-trivial number of pages that remain uncategorized.

If searchers are biased toward pages that have been categorized, perhaps at the expense of uncategorized pages, they could overlook valuable information.

Satisficing is a well-known behavior in information seeking, as searchers deal with time constraints and information overload (Simon, 1979). Information seekers will not spend an unlimited amount of time and effort on a search. They stop when they achieve an acceptable level of achievement, which can often be quite low. Searchers seek to minimize effort or maximize expected value using the available information. Categorized overviews provide more choices and lower cognitive cost for those choices involving category selection. For example, in study 3, with thematic categories, one subject made extensive use of the News category, because the task involved generating ideas for news articles, and this appeared to simplify his task, even though it also removed many potentially useful results from consideration.

Categorized overviews do add visual and cognitive complexity. They add visual complexity because the overview typically contains 20-30 category labels. They increase the number of possible actions and therefore the number of decisions that must be made. This can lead to excessive cognitive effort for some searchers. For some searchers, these issues can overshadow the benefits. During one experimental session during the third study, the subject asked if he could turn off the overview. However, the same study confirmed that most subjects found that the added complexity was a reasonable trade-off.

4.1.8 Implications

Organizing search results by meaningful categories allows the category knowledge to be used when viewing results. Tightly coupling category labels to the result list allows searchers to efficiently narrow /refine results using the categories. Supporting multiple kinds of categories permits searchers to draw on multiple forms of category knowledge. Retaining the result list and arranging it for efficient scanning and skimming is essential to supporting efficient assessments of the surrogates. Interfaces that lack a result list prevent searchers from efficiently assessing results. Attempts to use purely graphical displays are inappropriate for many tasks, according to this analysis, because users need to evaluate concepts and semantics, and these are best represented by text, by language. Meaningful category labels, however, can compactly encode important concepts because even short labels can convey meaning accurately and effectively (although this sometimes requires learning the meaning of the labels). Categorized overviews do add to the visual complexity of the display, increase the number of decisions that searchers must make, and may bias searchers away from uncategorized but valuable search results.

The analysis leaves open the possibility that quantitative concepts, such as document counts, that influence relevance prediction or otherwise indicate utility or novelty, can be usefully encoded with graphical elements, provided they do not affect the primacy of the text. For example, the color-coded bars employed by WebTOC (Nation, Plaisant, Marchionini, & Komlodi, 1997) can provide ancillary value by showing which categories are highly populated, but with the design shown in Figure 26, longer

labels are obscured by the bars. The challenges of first generation search visualization tools can be analyzed from this perspective, too. Visualizations like Grokker (www.grokker.com) and Kartoo (www.kartoo.com) privileged graphical displays, relegating text to a secondary role. Moreover, the visualizations were not based on meaningful underlying categories.

Figure 26. Long labels are obscured by the bar charts in this WebTOC display.

4.2 Design principles for exploratory search interfaces

User interface design principles capture important constraints, capabilities, features, tradeoffs, human preferences, domain knowledge, and human and machine processing limits encompassed by a design space. They can document best practices, useful heuristic strategies, and design patterns. Principles represent an integration of theoretical knowledge and empirical study, distilled to provide practical guidance to interface designers. They evolve, informed by theory, study, and reflection. The design principles proposed here integrate knowledge from human-computer interaction, information visualization, and information science with the analysis in

section 4.1, the results of the three studies (described in Chapters 3 and 5), and practical experience developing search interfaces. The principles are:

- Provide overviews of large sets of results
- Organize overviews around meaningful categories
- Clarify and visualize category structure
- Tightly couple category labels to result list
- Ensure that the full category information is available
- Support multiple types of categories and visual presentations
- Use separate facets for each type of category
- Arrange text for scanning/skimming
- Visually encode quantitative attributes on a stable visual structure

This set of design principles is based on the premise that consistent, comprehensible visual displays built on meaningful and stable classifications will better support user understanding of search results. As users explore search results, they are grappling with multiple simultaneous information problems: Their conceptualizations of the high-level information needs are imperfect and evolving; their understandings of the relevant concepts and terminology are limited, and their understandings of the presentation and interactions available in the interface are incomplete (Marchionini, 1995). Helping searchers to incrementally solve these problems allows them to fluently transition between the information seeking activities shown in Figure 25 and enables them to make more effective progress toward their high-level objectives.

4.2.1 Provide overviews of large sets of results

During an exploratory search, users may not have clearly formed information needs, the needs may be evolving, or they may not know the terminology and concepts of the search domain. In contrast to known-item search, fact retrieval navigational search, where the smallest possible number of highly relevant documents is desirable, during exploratory search, there may be hundreds or thousands of potentially relevant results. The visual information-seeking mantra prescribes, "Overview first…" (Ahlberg, 1993), and this is as appropriate for displays of search results as it is for other forms of information visualization. This is not a new idea – Table 13 lists several web search interfaces that display large number of search results – but it is important to reiterate that not all searches can be satisfied with a high-precision result set. The ideal number will certainly depend on many factors, including (but not limited to) the task domain, topic, the quality and quantity of documents, and search engine capabilities. The fact that many of the pages viewed in the three studies were ranked in the range of 50^{th}-100^{th} suggests that at least 100 results will be required to form the basis of a useful overview of Web search results.

Table 13. Seven web search interfaces that represent large result sets in the initial results. The default value and user-selectable range are shown where it was reported or could be determined.

	Number of results displayed	
Search interface	**Default**	**Range**
Vivisimo	200	100-500
Findex	150	Unknown
Google	10	10-100
Grokker	160	Unknown
Grouper	50	10-200
SWISH	100	N/A
Yahoo!	10	10-100

4.2.2 Organize overviews around meaningful categories

Gaining an overview of search results involves a number of cognitive subtasks, including interpretation of the results within the context of the searcher's internal mental model of the knowledge domain. Using meaningful, stable categories to organize results can place each result in a known context. Soergel (1999) has observed that classifications, taxonomies, and ontologies provide semantic roadmaps to fields of knowledge, improve communication and learning, and support information retrieval, among other benefits. The categories help searchers understand what concepts, ideas, and relationships are relevant in a domain, as well as suggesting query refinements. Categories based on document format, language, or Domain Name Service (DNS) domain can be useful. Numeric attributes such as date or size can be grouped into meaningful categories. For example, the Last Time Visited classifier

described in section Last Time Visited Classifier categorizes web search results into the categories: Today, Yesterday, Within a Week, Before Last Week, and Never Visited. Even abstract or computed attributes such as a journal impact factor (Garfield, 2005) can form the basis of meaningful, albeit controversial or limited, categories. Kwasnik (1999) argued that classifications support reflection, discovery, and knowledge creation.

The analysis in section 4.1 suggests that stable categories will allow searchers to reuse category knowledge on subsequent searches. This principle was originally formulated as, "Organize results by meaningful and stable classifications," emphasizing the importance of the stability imposed by traditional classification schemes. Dynamic categories, such as those generated by automated clustering techniques change with each query. Thus the learning benefits of stable categories may accrue less. In study 2, however, participants commented on the benefits of both stable and dynamic categories for different exploratory search tasks. Consequently, it was revised to reflect the complementary value of both stable and dynamic categories.

4.2.3 Visualize and clarify category structure

If the categories are drawn from a classification, taxonomy, or ontology, the structure should be made visible. This simple rule can be overlooked by implementers, who use that information for sophisticated query modification or relevance ranking schemes but then neglect to present it to the end user. The structure provides context for individual category labels, shows relationships between concepts, allows users to

focus on the portions of the concept space that are of most interest. The visual presentation must be disciplined to avoid overwhelming or disorienting searchers.

Practitioners should review at least the top two levels of a hierarchy, considering whether they need to be adjusted to provide the clearest overview. Parent-child (or broader-narrower) relationships that are clear when encountered while browsing a thesaurus or directory of web pages are not always clear when used in the context of a categorized overview of search results. The structure of the hierarchy may need to be changed in these cases. The importance of this emerged from the third study (described in Chapter 5). Some participants were puzzled because Television was a subcategory of Arts. Both of these categories were drawn from the Open Directory. The relationship between the two is clear when they are browsed within the Open Directory but not when used to organize search results, which lack the context provided on the home page.

4.2.4 Tightly couple category labels to result list

Tightly coupling the category labels displayed in the overview with the result list enables searchers to rapidly explore relationships between the two. Most commonly, the category labels can be clicked to narrow or broaden the result list. When this capability is implemented, it is important to provide clear feedback indicating which categories are currently applied. In all three studies, participants appeared to occasionally forget or overlook the fact that they were viewing a subset of their original query.

One benefit of tight coupling between the categories and results is that it allows searchers to very quickly see examples. Within a category, example results help to clarify the meaning of the categories and often provide indications of relevance, quality, etc. Even within well-known classifications, some category labels may be ambiguous or unfamiliar. A few examples can often clarify this. Dumais, Cutrell, & Chen (2001) noted that individual page titles helped disambiguate category names in their study of search results. This principle was initially formulated as "Provide examples of documents for each category." It was replaced with the current version because tightly coupling an overview with the result list provides a mechanism for users to quickly view a few examples or the complete set of all matching documents.

Brushing and linking techniques tightly couple multiple views of data in an information visualization, so that an action in one view (brushing) is linked to an action in another view. This can be applied to search results (Klein, Reiterer, Müller, & Limbach, 2003) to synchronize two views of the results, an overview and a detailed list. This can support richer interactions between category information and individual results. For example, pausing the pointer over a result in the list can highlight (in the overview) all categories containing that result. Brushing must be carefully used, though. During the evolution of the SERVICE prototypes, I experimented with a variation of this technique. In one version, pausing the pointer over a category label had the effect of immediately hiding all results that were not from that category. This was a very quick way to see results in a category, but was very disruptive. The screen would flash excessively as users moved the pointer over the categorized overview,

and the rearrangement of the list required users to visually reorient themselves with each change. The final design highlights the currently visible results that are members of a category when the pointer is placed over the category.

4.2.5 Ensure that full category information is available

When using deep hierarchies, designers should ensure that full category information (the complete label or descriptor) is available to searchers. The category labels in the overview indicate which categories results are in, but this may be limited to the top few levels because of the limited display space. During all three studies, but particularly during study 3, participants wondered aloud what specific category results were in. They were occasionally confused because only the top two levels of the category were visible in the overview. For example, the category /Arts/Television/Networks/Cable/BBC was truncated to /Arts/Television in the overview. Providing the full category label could clarify this. Displaying category labels in each result can be helpful (Drori & Alon, 2003). However, when this was implemented in the SERVICE system, the individual results became too large because results often appeared in multiple categories. Therefore, it was disabled prior to study 3. During development, we also experimented briefly with opening a pop-up window when the pointer moved over the result, but this was found to be visually distracting, because of the large size of the pop-up window. A small hyperlink in each result may be an appropriate design compromise, although this was not implemented or evaluated.

4.2.6 Support multiple types of categories and visual presentations

No single type of category is effective for all users, tasks, and domains. In her comparison of categories and clustering for organizing search results, Hearst (1999) noted that neither categories nor automatically constructed clusters will always align with users' interests. Libraries provide subject, author, and title indexes and archives provide multiple finding aids for their holdings. GRiDL, SuperTable (Klein, Müller, Reiterer, & Eibl, 2002), and Vivisimo's new Clusty.com search engine are examples of search result interfaces that permit users to reorganize results using alternate sets of categories. During the studies, several participants noted that they would like to be able to select or define their own categories and re-arrange them for their own purposes. Likewise, no single presentation style is ideal for all situations and tasks (Risden, Czerwinski, Munzner, & Cook, 2000; Sebrechts, Vasilakis, Miller, Cugini, & Laskowski, 1999; Shneiderman & Plaisant, 2004; Swan & Allen, 1998). Exploratory searchers should be allowed to select a task-appropriate form of data display (Shneiderman, Byrd, & Croft, 1997). Alternatively, if that level of control and the corresponding increase in complexity is not appropriate for the intended users, designers should have a variety of categories and presentation styles to choose from, so they can choose appropriate categories and visual presentation styles. It may be useful to provide functionality that enables a knowledgeable proxy for the user (e.g., a "power user") to customize the overview and share it with others. Supporting multiple classifications and multiple visual presentations will enable users to view and explore search results from the perspectives most appropriate to their needs.

4.2.7 Use separate facets for each type of category

When a rich set of categories encodes multiple types of relationships, presenting them as separate facets can clarify meanings and relationships that might otherwise be ambiguous. For example, categories for *is-a*, *is-about*, and *part-of* relationships should be presented separately. Faceted classifications organize a domain into orthogonal sets of categories, which are ideally homogeneous, mutually exclusive, and represent a single characteristic of division (Vickery, 1960). They have been used to organize catalogs, classifications, and thesauri (Soergel, 1974; Vickery, 1960), information spaces on the Web (Louie, Maddox, & Washington, 2003), and search interfaces (Yee, Swearingen, Li, & Hearst, 2003). Facets are flexible and extensible; they do not require comprehensive knowledge or impose a rigid ordering, and they allow the indexed entities to be viewed from a variety of perspectives (Kwasnik, 1999). The importance of this principle was clarified during the development of the SERVICE system. During informal user tests, searchers experienced confusion when categories with different meanings were used in the same facet. Separating geographic categories from topical categories in the final interface helped reduce this problem in the third study (described in Chapter 5). Other instances of categories that should have been separated out remained problematic. Therefore, hierarchies used in a categorized overview should be analyzed to determine whether they should be restructured into separate facets. The informal analysis performed during development yielded a noticeable improvement, suggesting that even a lightweight faceted analysis focused on the upper levels of a hierarchy could be beneficial.

4.2.8 Arrange text for scanning/skimming

At a perceptual level, users of search results attempt to rapidly ingest large amounts of text. In the formative studies, I observed searchers scanning titles and snippets of text to quickly select specific pages to view. They skimmed the pages and returned to the list to repeat this cycle. It could be argued that this is simply a result of the textual presentation format, but it also reflects more fundamentally that the source documents are inherently textual and are not easily presented graphically. Considered from an information visualization and perceptual processing perspective, text may be one of the most compact representations available for the broad range of information to be displayed as the result of a search. The graphical marks that humans recognize as letters are rapidly processed into words and concepts, allowing such diverse concepts as “war in Iraq,” “hot coffee and muffin,” and “search result visualization” to be represented in just a few pixels. This reflects a fundamental distinction between the strengths of human and machine capabilities. As humans we have an extensive, nuanced understanding of language that allows us to take advantage of a rich set of cues, including morphology, syntax, lexicon, context, and pragmatics that are only approximated by the algorithms implemented in machines.

Three important attributes of web search results identified by Drori (2003) – title, line in context (a snippet of text containing one or more query terms), and keywords – are free-text and not easily represented visually. The fourth important attribute identified by Drori, category, can be drawn from a controlled vocabulary and often structured hierarchically in thematic groupings. Arranging these elements in a consistent manner

(e.g. linear lists, columns, or matrices) (Teitelbaum & Granda, 1983) and ensuring that they are visible (rather than requiring interaction such as moving the pointer over an item) will support fast scanning and skimming. Aula (2004) found that presenting snippets as bulleted lists was 20% faster than the standard textual display. Appropriate use of font weights, styles, sizes, and colors will also help (Tullis, 1988).

4.2.9 Visually encode quantitative attributes on a stable visual structure

Information visualization principles are grounded in our understanding of human perceptual and cognitive systems, particularly their structure, functions, strengths, and limitations. Visualization techniques such as size, color, or shape-coding engage the human perceptual and cognitive systems by encoding data into visual constructions (Card, Mackinlay, & Shneiderman, 1999). Quantitative attributes such as dates or document counts and nominal attributes with a small range of values such as document types can be visually encoded by position, color, shape or size. Compared with text, quantitative attributes may be effectively visualized in more flexible ways. The underlying structure (the visual substrate) upon which the quantitative attributes are displayed is not limited to a list or grid because the perceptual systems are effective at detecting visual patterns, outliers, etc.

Stable, consistent, and meaningful displays have been shown to promote success in user interfaces (Shneiderman & Plaisant, 2004; Tullis, 1988). Niemela & Saariluoma (2003) demonstrated the importance of both spatial layout and semantics (labels) in learning a visual display. Providing a stable visual structure for the overviews,

structured around meaningful categories, will allow searchers to focus on the task at hand rather than re-interpreting a changing presentation of the results.

4.2.10 Summary

These eight design principles for categorized overviews have been refined and validated by the design and evaluation of the SERVICE system. They complement and extend general human-computer interaction, web design, information architecture, and information visualization principles. They will be useful for search interface designers because they provide guidance for the appropriate integration of visual overviews with search result lists, and particularly for the textual surrogates embedded in result lists. They do not yet address a number of issues, including how much stability is needed in the visual structure versus how much variability can be tolerated, what the permissible trade-offs are, and how much context is needed when navigating search results. These principles represent a strong call for exposing structure – which is often used internally by search engines, but less often exposed at the user interface – without abandoning the tried and true value of text.

4.3 SERVICE requirements and architecture

The initial SERVICE platform (version 1.0) supported the formative studies (studies 1 and 2) by providing tools to generate prototype interfaces with categorized overviews of search results using a government hierarchy. These prototypes could be used to explore a pre-computed set of results. SERVICE 2.0 was designed to satisfy three objectives:

- Provide a platform for investigating categorized overview interfaces

- Implement an architecture that facilitates easy plug-in of web search result classifiers
- Provide working search interfaces and logging features for study 3 (described in Chapter 5)

The findings of the early studies were used with the analysis and emerging design principles to define a set of high level requirements and specific desirable features. The feature list was pared to the features most important for the final study. The requirements and feature list guided the design and development of SERVICE 2.0. The following sections describe the SERVICE 2.0 architecture, the Fast Feature Classifiers that were implemented, the AOL Music Search Prototype, and the search interface constructed for the study 3.

The SERVICE architecture is organized around three major subsystems, all built using Java technology: the user interface, the data model (which includes the search result classes and machine interfaces to two search engines), and the classifiers (Figure 27). It also includes a small subsystem for logging JavaScript events from the search result page. The general operation of the system is shown as a data flow in Figure 28. Queries are sent to the search engine, and the results are categorized using one or more classifiers. the overview is generated from the categorized results.

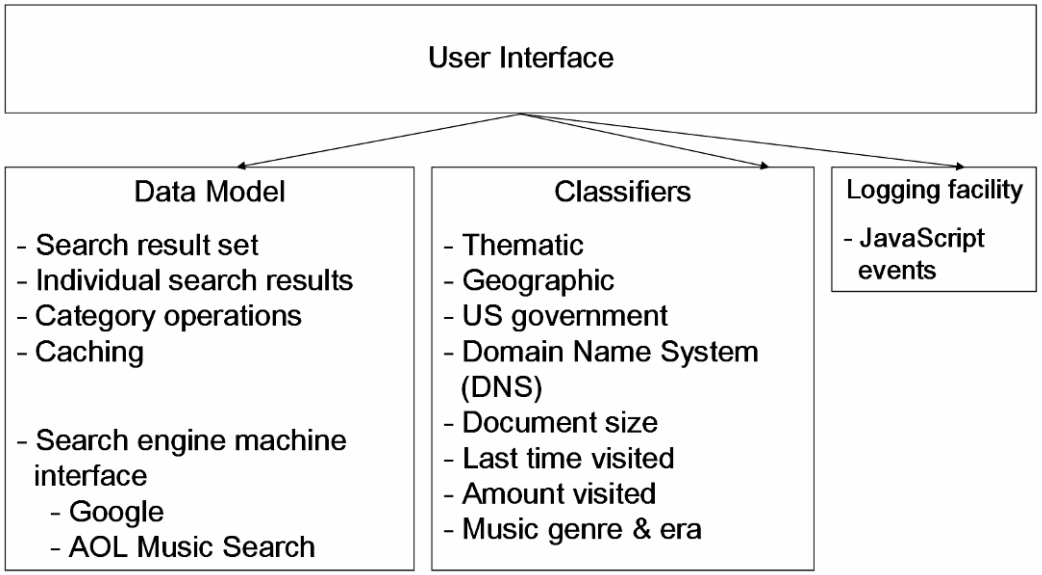

Figure 27. The SERVICE system consists of three major subsystems: the user interface, the data model (which includes machine interfaces to two search engines and the search result classes), and the classifiers. It also includes facilities to log JavaScript events from the search result page.

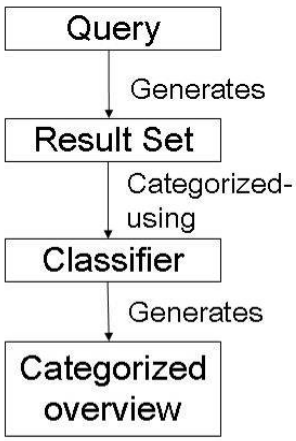

Figure 28. SERVICE operation is shown as a dataflow. Queries are sent to the search engine, which generates a result set. The results are categorized using one or more classifiers. The overview is created from the categorized search results.

The search result classes are used to create and manage search results, and include an interface to the search engines. They send queries to the search engine and parse the results to extract individual search result elements and group them into a set of search results. There are methods to cache results to and retrieve them from a database, optionally using a user ID. By default, when processing a query, the cache is first checked to see if it the query can be satisfied locally.

The classifiers are Java classes that that implement a common *Classifier* interface to categorize search results into meaningful and stable categories. A SERVICE classifier at minimum implements methods to:

- Categorize a single search result
- Categorize a set of search results
- Return the name of the classifier

A SERVICE classifier is any class that provides these minimal services. They do not necessarily implement machine learning or automated classification methods, although these could be integrated using the SERVICE architecture. An important design criterion for the classifiers was that they rapidly categorize results, using only data available in the search results (Zamir & Etzioni, 1998). This motivated the development of a set of *Fast Feature* classifiers, described below. SERVICE 2.0 supports nine classifiers, allowing the search interface to categorize search results into thematic categories and a US government organizational hierarchy, as well as others.

The user interface (UI) is the third major component of the SERVICE system. To support future studies, an important design goal was that the system be easily used by a variety of users. Ideally, the system would be accessible from any standard web browser without requiring special configuration. Early versions of the UI were implemented as Java applications, but early in the development process this was changed to a web-based application, using JavaServer Pages (JSP). This allows Java and HTML code to be combined in a single file, which is useful for rapidly prototyping and refining the UI, even though it does tightly couple content generation and presentation. Using Java applets or building browser plug-ins would have supported a richer set of interactions and visualizations, but for the purposes of this research, a combination of JSP and JavaScript provided enough functionality with minimal end-user demands. The design process and prototype evolution are discussed below.

SERVICE 2.0 implements a client-side logging function to capture events on the search results pages. The interface used for the study logs new queries (via the onsubmit event), page loads (oninit), mouseovers (onmouseover), page scrolls (onscroll), and link selection (onclick). Event time and the user ID are captured along with an event type and optional event data. JavaScript functions manage a set of log buffers, which are filled by calls from event handlers on the search result page. As the buffers fill, they are asynchronously sent to a log service. This is currently done by encoding the log contents as a URL and using that as the source for a JavaScript image object. This causes the JavaScript engine to send a request using that URL,

ostensibly to retrieve an image file. A more elegant approach would use the XMLHttpRequest object. The log service parses the URL request to recover the individual events. It timestamps entries upon receipt, so that any large differences in the clocks between the client and server can be accounted for. It does not account for differences due to network transport. For the study, the client and server were both hosted on the same machine, so the clock differences were not an issue, but future studies will involve remote clients. An important limitation of the JavaScript-based logging function is that it only logs events on search result pages. These pages are generated by SERVICE, so the logging code can be included. Other pages are not instrumented, and therefore do not generate log events. Since the primary interest of study 3 was on the search result page, this was acceptable, but a proxy server was installed to log all non-local pages.

4.4 Fast Feature classifiers

The need to rapidly categorize search results into meaningful and stable categories motivated development of a set of nine *Fast Feature* classifiers (Kules, Kustanowitz, & Shneiderman, to appear).[1] These classifiers use information available in the search results, typically the title, snippet, and URL, with valuable knowledge from external digital resources. The need to augment search results with additional metadata is indicative of the growing challenge facing digital libraries and archives caused by semi-structured and unstructured documents. Traditional digital libraries maintain

[1] Jack Kustanowitz contributed to the initial design of the fast feature classifiers and implemented five classifiers, under my direction.

rich metadata for their holdings, but as their holdings expand to include heterogeneous collections of semi-structured information, the available metadata dwindles, and human-generated metadata is expensive to create. External sources of digital knowledge can be integrated to provide valuable metadata, in this case, by supplying meaningful category information.

Figure 29 elaborates on Figure 28, showing a general data flow for the process of categorizing search results. Classifiers can be characterized along three dimensions: Lean/rich, online/offline and fast-feature/full-feature (Kules, Kustanowitz, & Shneiderman, to appear). Lean/rich captures the scope, breadth, and depth of the categories used. Online/offline refers to whether the categorization process requires extensive offline setup or configuration (e.g., training a statistical text classifier). Fast-feature/full-feature indicates whether the classifier can rapidly categorize search results at search time. These three dimensions are used as a framework to characterize the SERVICE classifiers.

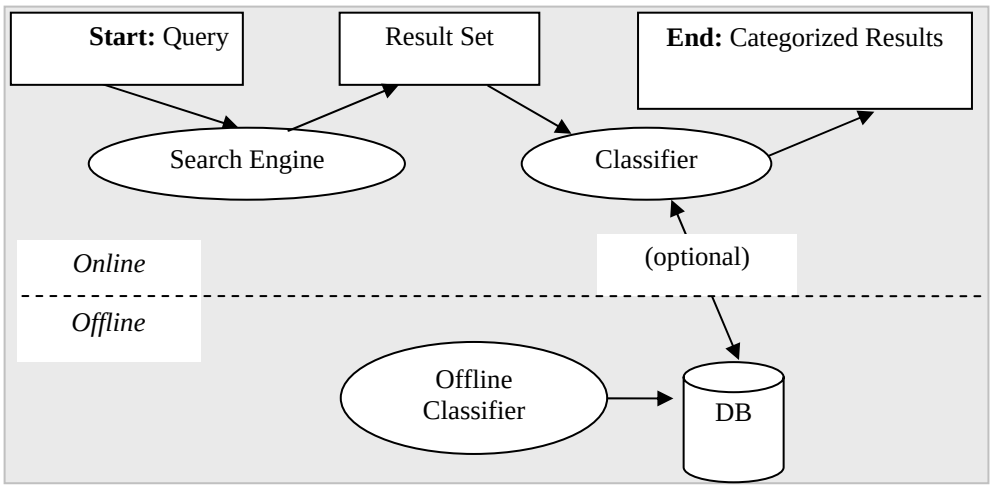

Figure 29. Components used to categorize web search results. A set of search results returned from a search engine is categorized by a classifier. The classifier may optionally reference previously acquired information or knowledge, such as a database of rules or training data.

Lean categories are simple, readily understandable categories with modest breadth and depth. In the context of the web, they can be constructed from document attributes such as file formats (DOC, PDF, PPT, etc.), DNS top-level domains (COM, GOV, ORG, etc.), and meaningful date or size ranges. As an example of the utility of lean categories, Matsuda & Fukushima (1999) found that using the document type (e.g., product catalog, online shop, call for papers, home page, bulletin board) in searches improved precision of the results.

Rich categories are extensive classifications, taxonomies, ontologies, or other knowledge structures, often professionally developed, that provide "semantic roadmaps" of an area of knowledge that can be useful for searchers (Soergel, 1999). Examples of rich classifications include the ACM Computing Classification System, West Publishing's Key Numbers classification of legal topics, Library of Congress

Subject Headings, and the US Government organizational hierarchy. Web directories like Yahoo! and ODP organize web sites into thematic hierarchies. They are of interest here because they cover a small but important portion of the web with high quality. Taxonomies such as MeSH also have been used to organize search results in specialized (non-web) search applications (Hearst & Karadi, 1997; Pratt, Hearst, & Fagan, 1999).

Categorization can be done either completely online (at query time), or it may require prior processing (offline). Online categorization can be done when the search results are generated if the mapping of page to the hierarchy is trivial (for example, grouping by the DNS domain suffix such as .GOV, .COM, .EDU, etc.), or if it comes "for free" with the result set (search engines may provide one or more topical categories for each result), or if it is a function of the result set (such as grouping by document size, where the size ranges depend on the result set). Online categorization can be done from a database, either local or remote (such as querying the Open Directory Project (ODP) web directory (dmoz.org) if the topical category is not provided with the query result set).

Offline categorization is required if no database exists to map search results to the desired categories. In that case, an agent such as a web crawler looks at URLs (fast-feature) or actual web pages (full-feature), potentially creates a hierarchy or reads an existing one, and places that page into the appropriate place in the hierarchy, storing the resulting mapping in a database. Run-time activity is then simply looking up the

URL in question in the database and returning the appropriate mapping. Web page classifiers may require offline training to learn statistical models of the categories.

A search-result categorization technique is referred to as *fast-feature* if it requires only information provided in the search result set, and therefore does not require the full text of each link destination. In contrast, a *full-feature* technique is one that requires the full text of the link destination (or possibly other documents, e.g., if it uses structural information such as hyperlinks). Typically information returned includes URL, date, size, and perhaps summary and/or topical category. Thus, for example, a technique such as a text match on the URL would be considered fast-feature, but one that does textual analysis of the body of the HTML page pointed to by the link would not. Table 14 summarizes how these distinctions may divide up the space that describes how search results are analyzed.

Table 14: Techniques for Search Result Categorization. SERVICE implements a set of online, fast-feature classifiers, in the black border

	Online (at query time)	**Offline** (requires prior setup or background processing)
Full-feature	Accessing each web page in a search result and doing extensive analysis (not addressed here; often impractical due to performance)	Extensive text processing, manual, link analysis, machine learning (Work done by information retrieval and classification researchers)
Fast-feature	Uses only features in result set, such as title, snippet, URL, domain, size, ODP, pre-existing database map	Web crawler for URL directory hierarchy parsing, search engine mining (query probing)

Full-feature online techniques would consist of reading a list of links returned from a search engine, and then at runtime, downloading each destination, performing some analysis on each page, and then doing some kind of categorization. This is not easily scalable to large result sets, because it requires *N* network calls for *N* results and is largely dependent on remote sites for correct functionality. While it might be feasible on a set of pages with reliable links and guaranteed fast network performance, or when pages are available on the local machine (e.g. a search engine that caches indexed pages), it is not practical in general.

Much research has been done on *full-feature offline* techniques by information retrieval classification researchers. In general, these require downloading and analyzing the full contents of each page, whether it is using link data to automatically build site maps as in MAPA (Durand & Kahn, 1998), or machine-based learning techniques that can categorize pages based on statistical analysis of word counts. Manual categorization, in which page designers categorize their respective pages can also be seen as a full-feature technique, as it also requires knowledge of the page contents.

The following subsections discuss two kinds of fast-feature classifiers. Six online lean techniques are briefly considered before focusing on the three online rich techniques. The fast-feature techniques draw on meaningful relationships between a feature in the search result and some external database or other knowledge structure. If the relationship exists, that is evidence of membership in the category. The converse,

however, is not true. If no relationship exists, that can either mean that the page is not a member of the category, or that the external database is incomplete. When no relationship exists, an assignment could be made using traditional classification techniques. This might result in more pages being categorized, but it could also result in incorrectly categorized pages. The techniques described here are conservative; they do not assign pages to categories without an explicit relationship in the external database.

When analyzing these techniques, an important characteristic is what proportion of search results can be categorized. To assess the potential utility of these methods, examples of each kind of classifier were implemented, and the percentage of search results that each categorized (which will be referred to as *coverage*) was measured or analytically assessed. Each classifier was targeted to a specific domain, so five representative queries were constructed for each target domain. For each query, the top 100 search results were retrieved from the Google search engine, and the number of results categorized by the classifier was measured. Additional analysis was performed on the ODP classifier, because I intended to use it in study 3.

4.4.1 Online Lean Techniques

A fast-feature online categorization technique is one that does not require the offline creation of a database, and also does not require the full text of the link destinations. The lean techniques often draw on surface features of the URL, such as the top-level domain to classify documents into simple categories. Table 15 contains a sample of lean classifiers. This list is far from complete, but it illustrates the breadth of

classifications available using only the data returned from the search engine and any freely available, pre-existing databases. The following three sections describe online lean fast-feature classifiers.

Table 15. Online lean classifiers can provide simple categories to help users locate relevant information. The three classifiers that have been implemented in SERVICE 2.0 are highlighted in bold.

Name	Description
Top-level DNS Domain	This classifier extracts the final part of the hostname, which typically indicates either a country code (e.g., us, jp, uk, de, etc.), or one of the defined top-level domains (.COM, .EDU, .ORG, .GOV, etc.). This provides a simple way to provide a flat (non-hierarchical) categorization. A search for "chip manufacturers", for example, could be usefully organized according to country code.
Last Time Visited	The web browser history can be used to categorize documents by how recently they were visited (e.g., today, yesterday, this week, this month, never).
Document Format	The file format of the document (e.g., HTML, PDF, PS), can often be determined from the suffix of the filename in the URL or from a format indicator in the search results.
Document Language	The document language can be inferred from the title and snippet using dictionary lookup, yielding a flat categorization.
Document Size	This classifier groups results into similar size classes. Size categorization may be useful for image search.
Document Indexing Date	Search engines sometimes provide the date the document was indexed (or "crawled") in search results. This can be used to categorize documents by how recently they were indexed, using values similar to the previous example.

4.4.2 Top-Level DNS Domain Classifier

The domain classifier is one of the simplest of the classifiers implemented in SERVICE. It places URLs into a flat set of about 110 categories based on the domain suffix (.COM, .EDU, .GOV, etc.) or the appropriate country code. A lookup table

maps the country code to country name, so that the categorization text can use the actual country name. For example, the following two URLs are categorized as follows:

- www.whitehouse.gov/ -> GOV
- http://www.corriere.it/ -> Italy

(Country names can also be determined for non-country code-based URLs (Periakaruppan & Nemeth, 1999; Watters & Amoudi, 2003), which would provide a mechanism for categorizing URLs into a geographic hierarchy.)

A user interface showing this categorization would allow quick navigation to all educational institution web sites, for example. Because the domain is available in almost every search result, this has the desirable property of nearly 100% coverage, that is, almost no results are left "uncategorized." Country codes may not be immediately recognizable to searchers, and at least one country (Tuvalu) has used its top-level domain (.tv) to host television websites, which could pose some challenges to searchers.

4.4.3 Last Time Visited Classifier

Categorizing search results by when they were last seen can be useful in certain situations. Although searchers attempt to re-access previously found documents via search engines, they have trouble remembering the specific query and/or navigation sequence that they originally used (Aula, Jhaveri, & Käki, 2005; Wen, 2003). Integrating these categories into a search interface could help searchers more readily find previously visited pages. Alternatively, these pages could be excluded from

search results if the searcher wished to find new material. Personal browse histories maintained by a web browser can be used to indicate whether a web page or its web site has been visited and if so, when it was last visited. The SERVICE classifier categorizes web pages into five categories: Today, Yesterday, Within a Week, Before Last Week, and Never Visited. This classifier depends on the existence of a complete browse history, which introduces the issues of privacy and data storage size. The initial implementation works with the Firefox web browser (www.mozilla.com/firefox). It uses an external script to read the web browser history file, which is only updated when the browser exits, so sites visited in the current session are not immediately visible. If a complete browse history is available, this technique will provide 100% coverage, because any page not in the history can accurately be placed in the Never Visited category. If the browse history is limited, however, the Never Visited category cannot be used, because the absence of a page in the history file could either mean the page was never seen, or that it was seen but subsequently removed from the history.

4.4.4 Document Size Classifier

The Document Size Classifier uses page size information when it is available in the search results. When search engines return size information for pages, a dynamic categorization of sizes can be determined automatically, and this classifier can thus also run online. This could be useful when searching for images or multimedia documents. Categorization may be done uniformly using a fixed set of ranges (which may yield many categories with 0 results), or by online defining ranges that contain matches within the result set. Our implementation defines a constant number of

groups, divides the range of page sizes by the number of groups, and then places the results into one of those groups. This is useful for visualizing a uniform distribution of page sizes. An alternative implementation could choose categories of fixed intervals, such as 100k-200k, 200k-300k, etc., even if the categories were not a uniform size. This would be useful for seeing, for example, that no results were between 100k and 3MB for a given query. If both of these implementations were published and adhered to the common interface, a searcher could choose which size classifier to use based on the desired visualization or search. This classifier will trivially yield 100% coverage.

4.4.5 Online Rich Techniques

Rich categories are appealing to users because the descriptive terms facilitate understanding. The fast-feature, rich techniques typically use a pre-existing database to map a URL to one or more categories. Table 16 identifies several rich classifiers. This illustrates the breadth of classifiers available. The following sections describe online, rich, fast-feature classifiers.

Table 16. Online rich classifiers can provide meaningful and stable categories that add context to the search results.

Name	Description
US Government	This classifier uses a pre-existing database that maps URLs to a government hierarchy. For example www.whitehouse.gov/president maps to the second-level category Executive/Executive_Office_of_the_President.
Open Directory Project (ODP)	This classifier uses the Open Directory Project category information that is returned with the query results to build its hierarchy. The ODP is a human-edited web directory (www.dmoz.org).
Musical Genre	This classifier parses search results from the AOL Music search engine to categorize songs according to a two-level musical genre. (A similar classifier categorizes songs by period.)

4.4.6 U.S. Government Classifier

The government classifier uses an existing database that maps government web pages into a government hierarchy, for example mapping *http://www.af.mil/* to the hierarchy node /Executive/Executive_Agencies/Department_of_Defense/ Department_of_the_Air_Force. Since the lookup is done locally, this can be done online at query-time. On its own, this classifier has coverage that is limited to the list of URLs in the database. However, any URL that is an extension of this base URL is also associated with the Air Force. The coverage can therefore be extended by using prefix matching, i.e., *any* URL beginning with *www.af.mil/* would be mapped to this node, unless a more detailed match was found. Five representative queries were constructed by selecting the most commonly asked questions reported by the First.Gov web site (http://answers.firstgov.gov/ cgi-bin/gsa_ict.cfg/php/enduser/ std_alp.php), removing obviously navigational questions, as described in Broder (2002), and creating short queries from keywords in the questions. The results are

shown in Table 17. For both the “new passport” and “foreign embassy” queries, many of the uncategorized pages were from the domain “usembassy.gov”, whereas the database had “usembassy.state.gov”. This slight difference illustrates the sensitivity of this approach to URL variations, and suggests that additional heuristics or an index of synonymous URLs could be developed to make it more robust, a technique used by search engines.

Table 17. Percent of the top 100 results categorized by the US Government classifier for five representative queries.

Query	% Categorized
new passport site:gov	39
start business site:gov	58
gasoline prices site:gov	100
foreign embassy site:gov	43
obtain grant site:gov	72

4.4.7 Open Directory Project Classifier

Web directories such as Yahoo! (www.yahoo.com), LookSmart (www.looksmart.com), and the Open Directory Project (www.dmoz.org) catalog a small but important fraction of the Web. They provide an overview of general Web content and enable information seekers to find information by browsing a familiar subject hierarchy. As of April, 2006, its 72,000 volunteer editors had indexed 5.3 million web sites in 590,000 categories (16 top level categories). The Open Directory Project classifier uses Open Directory Project information to place search results into categories within the ODP hierarchy. Even though web directories cover only a small fraction of the web, popularity follows a power law (Cunha, Bestavros, & Crovella,

1995). That is, a few sites receive much use. I conjectured that the highest ranking pages in search results would often be cataloged in the ODP. To categorize a search result into the ODP hierarchy, the web site is looked up in the ODP using prefix matching as in the US Government classifier. Since web sites can be cataloged in multiple categories, this yields a list of categories for the result. For example, a web page from the web site of the University of Maryland Human-Computer Interaction Lab would be categorized into the following three ODP categories:

- /Computers/Human-Computer_Interaction/Academic
- /Computers/Computer_Science/Academic_Departments/North_America/United_States/Maryland
- /Reference/Education/Colleges_and_Universities/North_America/United_States/Maryland/University_of_Maryland/College_Park/Departments_and_Programs

The classifier used a web service provided by Alexa.com. The Alexa service only categorized a single web page per HTTP request, so a cache was implemented to minimize processing time for search results when they have been previously encountered.

Five queries representative of general web search were selected from the most common searches reported by AskJeeves search engine (Ask.com, 2005), after removing navigational queries. In addition, the five government queries described above were also evaluated (Table 5).

Table 18. Percent of the top 100 results categorized by the Open Directory Project classifier for five representative queries in each of two domains: general web search and government web search.

Query	% Categorized
General web search	
music lyrics	76
Games	83
Maps	90
real estate	82
Poems	76
Government web search	
new passport site:gov	69
start business site:gov	73
gasoline prices site:gov	90
foreign embassy site:gov	68
obtain grant site:gov	88

The preliminary tests were promising, and I wished to measure coverage for a more extensive set of searches. Coverage rates for the ODP were particularly interesting, because the study 3 would use these categories for general web search. The TREC 2004 Robust Topics provided a set of 250 queries created as realistic, but difficult topics for information retrieval. For each of the 250 topics, the contents of the Title field were submitted to a Google search and the top 350 results were collected. This yielded 86,900 results. Because of the quantity of results, it was not practical to use the Alexa service to categorize them. The ODP data was imported it into a MySQL database and processed using PHP scripts. Each result was then checked to see if it could be categorized in the ODP. The number of results categorized within the top 100, 250 and 350 results was measured (Table 19). The average coverage for the 246

queries successfully processed and categorized was 66.0%, 62.9% and 61.6% for the top 100, 250 and 350 results, respectively.

Table 19. Coverage for the top 100, 250 and 350 search results from 246 queries based on the TREC 2004 Robust Topics.

	Range	Mean (SD)	% Categorized
Top 100	36-87	66.0 (7.68)	66.0
Top 250	87-194	157.2 (16.00)	62.9
Top 350	110-257	215.6 (21.11)	61.6

This work is related to work by Chirita, Nejdl, Paiu, & Kohlschütter (2005) in its use of ODP data to organize search results. They used the ODP data to re-rank Google search results, boosting the rank of preferred categories, which were selected in advance by the searchers. They found that the top 5 re-ranked results were judged better than the original top 5, which illustrates the value that a large-scale knowledge resource can provide. The ODP is used differently in SERVICE, to expose the structure of the search results to searchers in the form of an overview, allowing searchers to choose categories at search time and avoiding the need to pre-specify categories of interest. The measured coverage results were higher for the SERVICE tests than in theirs, and we can consider two possible causes for this. They elicited specific types of queries (ambiguous, partially ambiguous, and unambiguous) from their test participants, who were research colleagues, whereas the SERVICE tests used a set of TREC topics. It is possible that their queries were focused more narrowly to yield the desired level of ambiguity. It is also possible that the prefix

matching strategy allowed the SERVICE classifier to categorize a larger fraction of pages. The results of study 3 lend support to the prefix matching approach.

4.4.8 Multi-threading the ODP Classifier

During development of the search UI, two additional requirements were identified. As mentioned above, the Alexa Web Service categorizer processes only one URL per call. Each call typically required 1-2 seconds to send the HTTP request, and then receive and parse the XML response. To categorize 100 search results was found to take close to two minutes. At that time, the ODP data had not yet been downloaded to a local database, so to reduce the categorization time, I implemented a multi-threading option. This allowed multiple search results to be simultaneously processed. Alexa, however, would occasionally return an error, with likelihood of the error increasing with the number of simultaneously outstanding requests. This required retrying the request, which could further add to the load. A backoff strategy was needed to slow the request rate when this happened. By evaluating different combinations of values for the number of threads and the backoff times, I was able to reduce the typical time to process a set of 100 results to 20-30 seconds, with an acceptable error rate, approximately one or two per result set.

4.4.9 Extracting multiple facets from the ODP hierarchy

The second additional requirement was due to the need to extract multiple facets from the (single) ODP hierarchy. Specifically, the Geographic facet is extracted from the top-level Region category. This was desirable because the geographic categories implied a different relationship than the other categories. When a web site was

categorized under the other categories that generally (although far from always) meant that the web site was *about* the concept represented by the category. However, web sites were categorized under the Region category when the organization that published the site was *located in* a specific region. This qualitative difference in the relationship between a category and its member web sites warranted a visually separate facet in the UI. To accommodate this need, the classifier was extended, adding a new constructor that accepted a top-level category value as the root of a category hierarchy.

4.5 AOL Music prototype

The AOL Music prototype demonstrated categorized overviews of music search results, integrating an external database with AOL Music Search results to generate the overviews. It began to explore the use of multiple facets by displaying two types of categories in the overview: genre and era. The genre facet consisted of 11 top-level genres of music (e.g., blues, classical, country, etc.), combined with an optional second-level, drawn from an uncontrolled vocabulary. Thus a song could be categorized into the top-level category, Rock, and a second-level category, Pop-Folk. The era facet was composed of decades from 1910 to 2000.

This design is similar to guided search designs such as Tower Records music search (towerrecords.com). Whereas that system draws data from a single database, the AOL music prototype demonstrates the integration of web-based search results with an external data resource. It also illustrates the utility of the SERVICE 2.0 architecture, since the prototype was built in less than a day. The category information for both

facets was extracted from the freedb.org CD database, which contains entries for 1.9 million CD albums. Two new classifiers were constructed, and the web-based interface was adapted to display the song, artist, and album for each search result. Additionally, a search engine interface was constructed to send the user-specified song query to AOL's music search engine and parse the HTML results. A query typically can be processed, categorized, and results displayed within 5-10 seconds, depending on the speed of the AOL Music search engine.

Figure 30. A search for songs with the words "road" and "travel" in the title yields 124 results. The results are presented with two categorized overviews: by genre and by date. Here, the results have been filtered (by clicking) to show just the 21 Country songs.

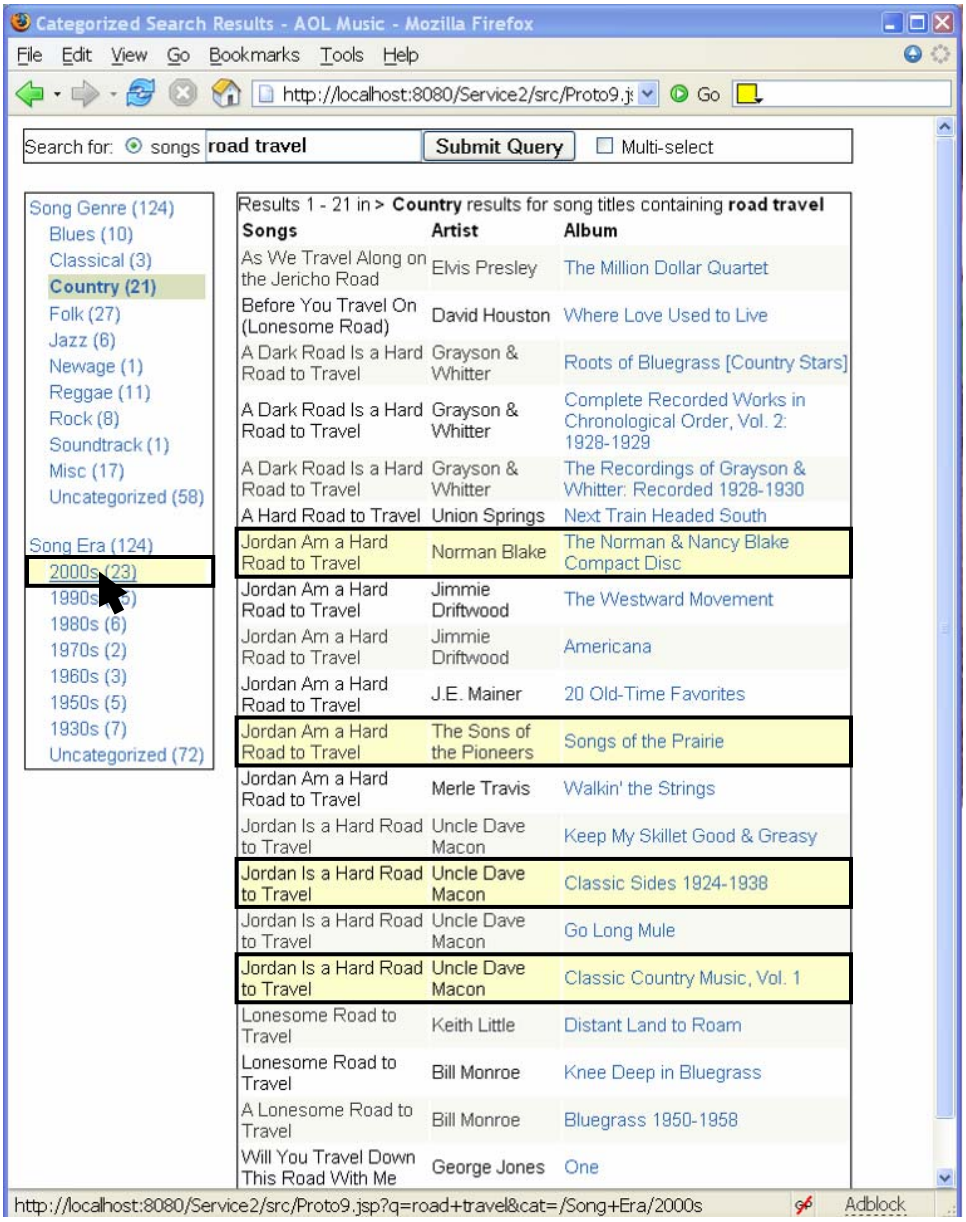

Figure 31. Brushing the pointer over a category highlights the results that fall in that category. In this screenshot, the pointer has been placed over the "2000s" category, showing albums released in the 2000s highlighted with yellow (shown boxed for clarity in these figures).

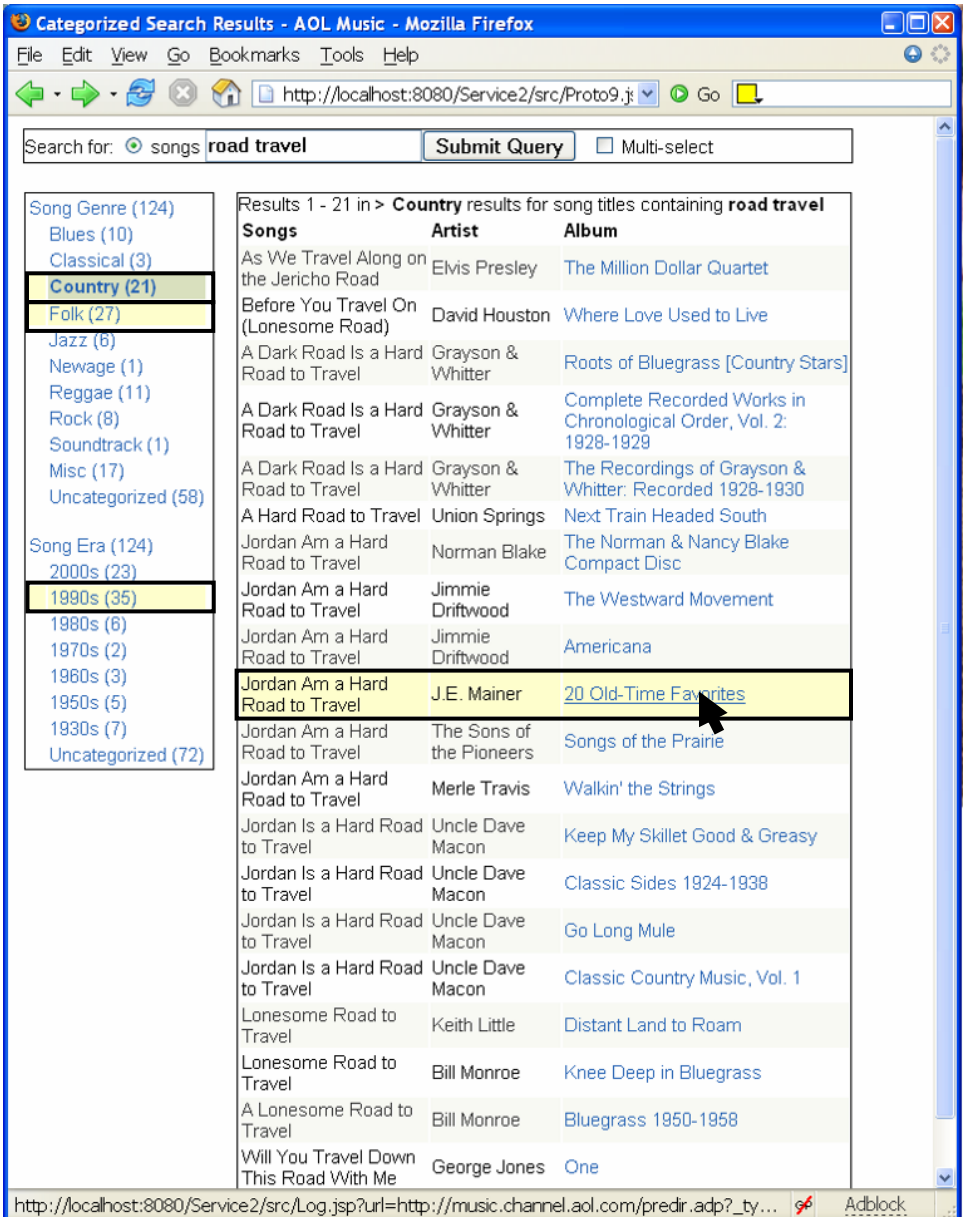

Figure 32. Brushing the pointer over an album title highlights all the categories for that album. Here we see that J.E. Mainer's "20 Old-Time favorites" is in both the Country and Folk categories, and that it was released in the 1990s.

4.6 General web search interface

The SERVICE requirements document and feature list guided development of the search interface for study 3. The SERVICE architecture facilitated implementation of alternate user interface designs and categorization schemes. User interface designs were informally reviewed with HCIL and professional colleagues through the evolution of the interface designs to the design used in the third study.

Since the study 3 would investigate categorized overviews in the context of general web search, it was important to select sets of categories that were appropriate for that domain. The evolving SERVICE designs explored multiple presentations based on the Open Directory Project classifier. As a formal classification, web directories have several limitations (Taylor, 1999); however, they provide a rich hierarchy appropriate for categorizing general web search results. As reported in section 4.4.7, a substantial portion of typical web search results have been cataloged within the ODP, which made it practical to use for the study. The evolving SERVICE designs explored multiple presentation based on the ODP categories.

The first implementation of SERVICE 2.0 used a Java application to send queries to the search engine, parse and categorize the results, and cache them in a local database (Figure 33). The application opened an external web browser to display the results. The URL passed to the browser pointed to a local JSP script and encoded any selected category filters. The JSP script extracted the results from the database and formatted them to display the (possibly filtered) list of results to mimic Google. The

overview allowed users to select (via a drop-down list at the top) from multiple category sets (ODP categories, US government, and DNS domain). One category set was visible at a time, displayed using an expandable outliner. Because only a single facet was displayed, the entire height of the screen could be used, and multiple branches or sub-categories could be expanded at one time.

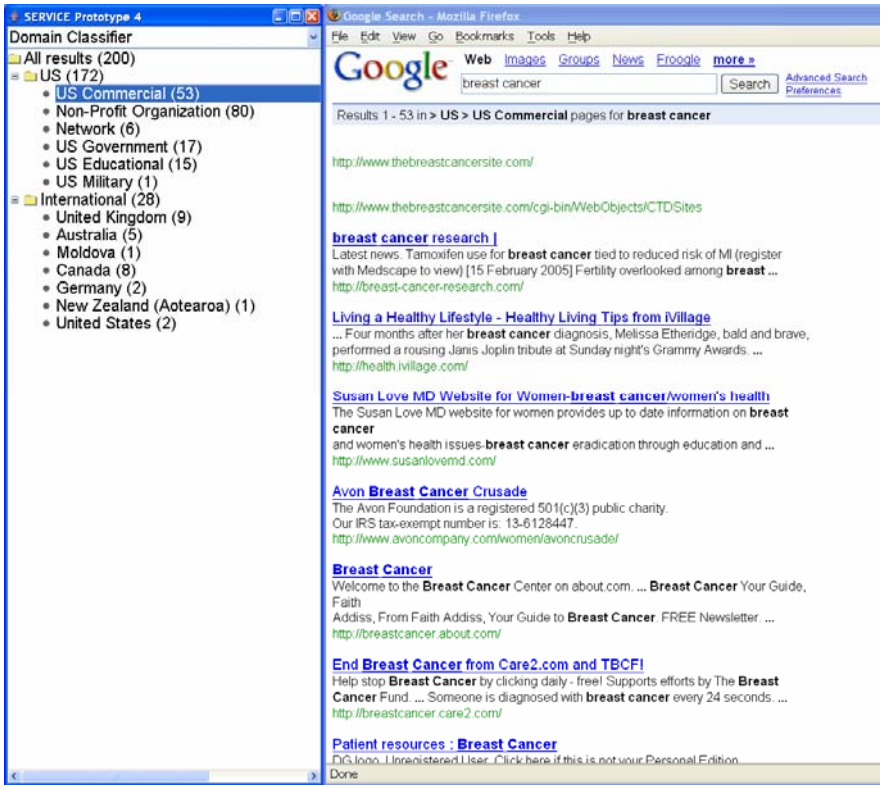

Figure 33. This SERVICE search interface allowed users to select one set of categories at a time, which were displayed with an expandable outliner. This screenshot shows search results with a categorized overview based on the DNS domain. The US and international categories have been expanded. The results have been filtered to display just the 53 US commercial (.COM) sites. A drop-down list at the top of the overview allows users to select alternate category sets.

This design had several drawbacks. It required searchers to manage two windows. Tiling the windows was preferred, but the one of the windows could inadvertently be moved, minimized, obscured, or closed. When searchers clicked on a category in the overview, the results page would load in the browser window, but if that window was obscured, searchers would not see the results, because the browser did not receive the focus and get moved to the top of the visible stack of windows. The design also exacerbated existing usability issues with the browser Back button (Cockburn & Jones, 1996; Kaasten & Greenberg, 2001; Milic-Frayling et al., 2004), because the state of the overview could become inconsistent with the browser window. As users navigated with the Back and Forward buttons, the overview was not tightly coupled to the browser and would not be updated. Finally, the use of a Java application, although acceptable in the confines of a controlled experimental study, would present installation challenges for end-users.

This led to two important decisions about the evolving SERVICE system. Integrating the categorized overview and the list into a single JSP page would present a cleaner, more consistent interface to searchers. And displaying multiple facets simultaneously would provide alternate perspectives within the overview, hopefully providing a more complete overview of the results. These changes were first instantiated in the AOL music search prototype, followed by the next web search design, which displayed four facets simultaneously in the overview (ODP categories, DNS domain, US government, and document size). Display and selection of categories within each facet was sequential, however. This meant that only one branch at a time could be

explored within a facet, and only one category at a time could be selected within a facet. Trade-offs inherent in the display and navigation of hierarchies are well-recognized (Hochheiser & Shneiderman, 1999; Larson & Czerwinski, 1998; Miller, 1981; Norman, 1991; Zaphiris & Mtei, 1997). For the constrained space available to the categorized overviews these were reasonable choices.

Subsequent designs explored variations in the structure of the categorized facets. For example, one design promoted all 16 top-level categories in the ODP to separate facets (Figure 34). This forced the overview to extend beyond the first screen and required excessive scrolling. Another design promoted just one ODP top-level category, Reference, to a separate facet, retaining the others in a Topic facet (Figure 35). This was done based on the observation that the Reference category implied a different relationship with its member sites than the other ODP categories. Reference indicated a *kind-of* web site, whereas other categories were thematic, indicating what the web site was *about.* This separation seemed to help searchers better comprehend the search results as they filtered and explored. The ODP Regional category was also promoted to a facet for a similar reason. Although the official description of the category in the ODP states that, “The Regional category contains English language sites about geographical regions of the world,” in practice, web sites are apparently categorized there because the publishing organization is located in a geographic area, thus encoding a *location-of-publisher* meaning.

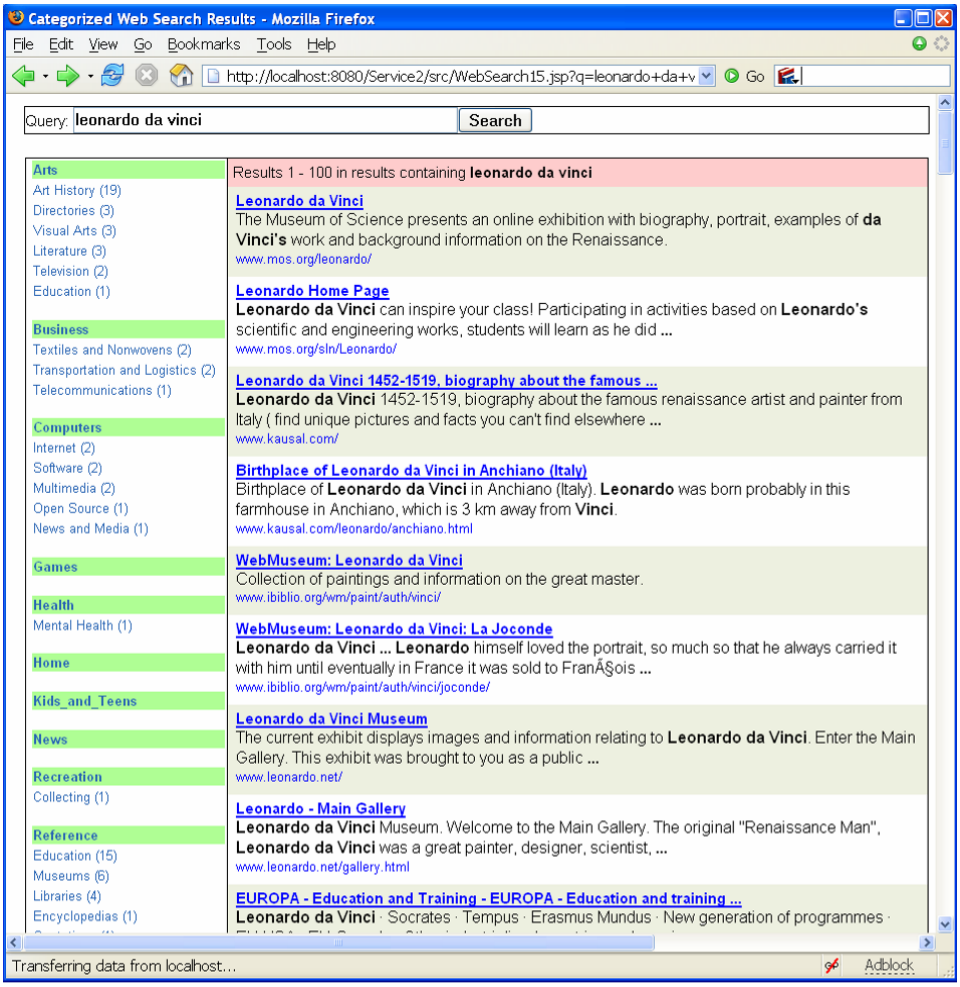

Figure 34. In this search interface, ODP top-level categories are shown as separate facets.

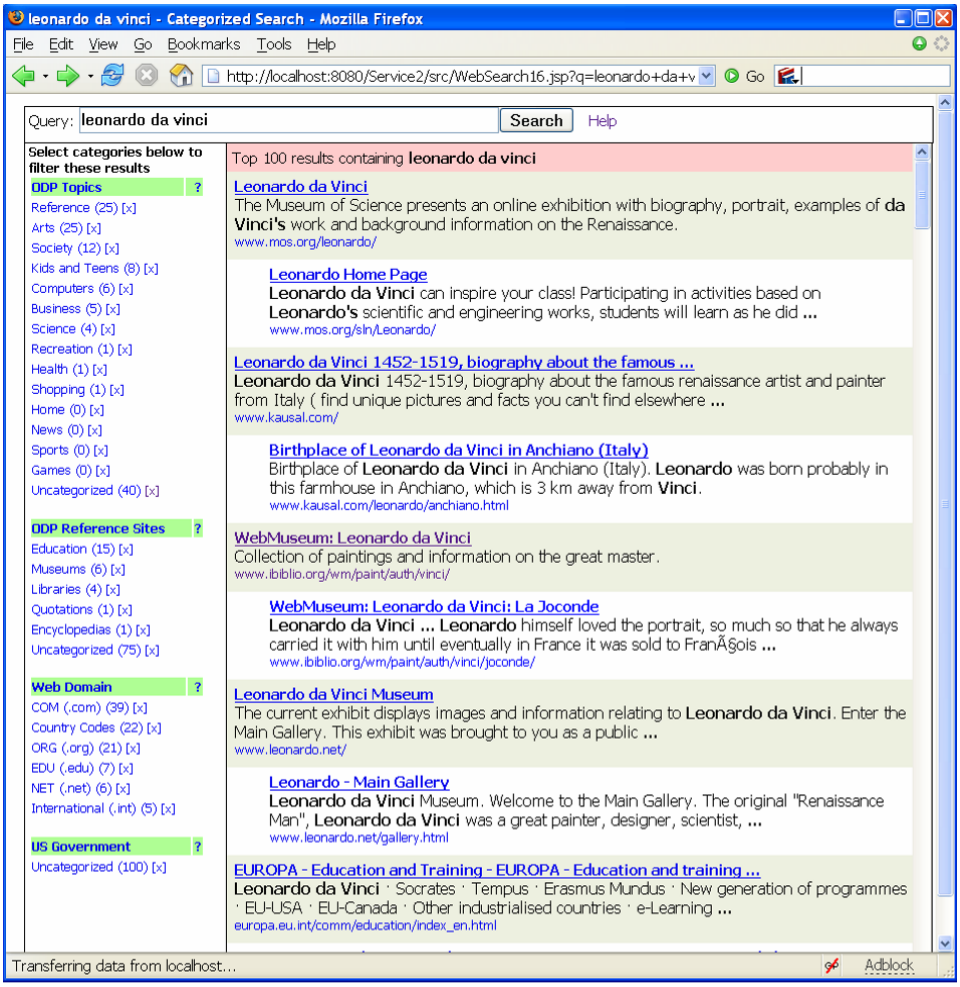

Figure 35. The search interface treats the ODP Reference category as a top-level facet. The remaining ODP categories are treated as another facet, in conjunction with the top-level DNS domain and the US government categories.

The Last-Time-Visited facet, seen in the “median” search example of Figure 1, was incorporated to explore the use of personally meaningful categories in the overview. It currently requires running external scripts to update the database for each use, so it has not been used extensively or evaluated. An approach to implementing a practical classifier is discussed as future work in Chapter 7.

There is a trade-off between the number of facets displayed and the need to constrain the overview to a single screen. Additional facets also bring additional visual and cognitive complexity to the overview. With rich sets of categories (which yield wide and deep hierarchies), as the user navigates into the second and third-level categories, the number of categories often expands substantially, and this can cause the overview to grow beyond a single screen. The final study design used three facets: ODP topics, geography (drawn from the ODP Regional category), and US government (Figure 36). Limiting the overview to three facets helped ensure that they did not extend beyond one screen and avoided "facet overload." As the study results reported in Chapter 5 show, this was an effective compromise for most searchers. An alternative, permitting searchers to customize facet and category display, is discussed as future work in Chapter 7.

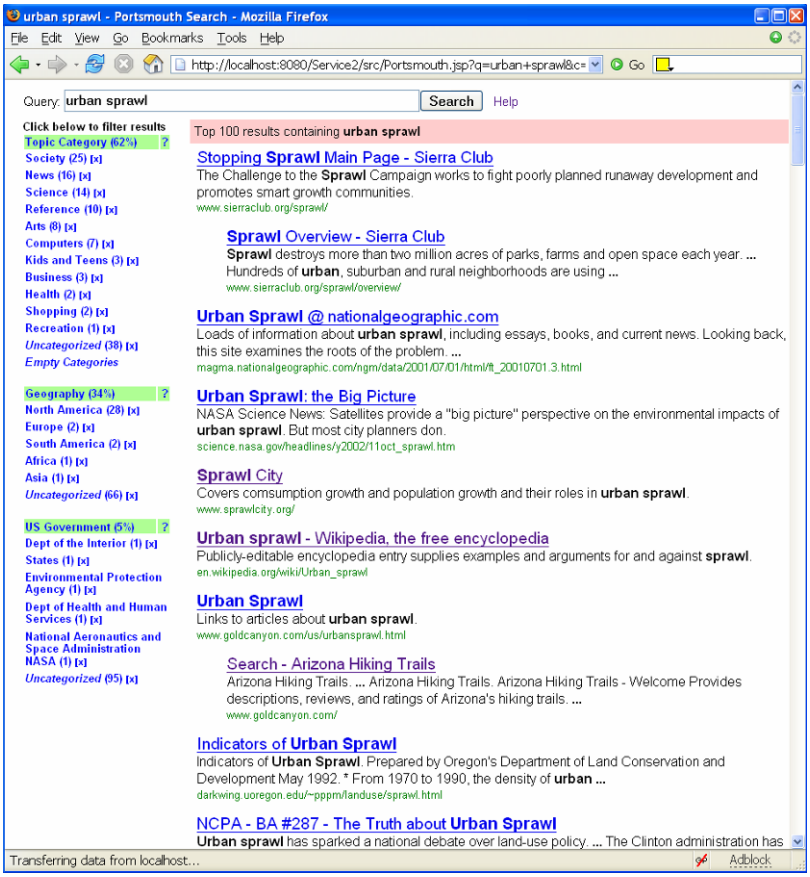

Figure 36. The search interface for the final study coupled the ranked result list with a categorized overview based on topical, geographical and US government classifications.

The design decisions made during this process are shown in Table 20. They illustrate the breadth of issues encountered when designing categorized overviews.

Table 20. Dimensions of the design space for categorized overviews.

Design dimension	Design choices	Support in SERVICE?	In study 3 interface?
Display of facets	• Design-time • User-controlled	✓	✓
Selection of facets	• One-at-a-time • Simultaneous	✓ ✓	 ✓
Display and selection of categories within a facet	• Sequential • Simultaneous	✓ ✓	✓
Display and selection of categories between facets	• Sequential • Simultaneous	✓ ✓	 ✓
Visible levels of hierarchy displayed	• Current level • Current + children • Current + grandchildren • Display children in pop-up	✓ ✓ ✓	✓ ✓
Overall depth of hierarchy	• Fixed • Unlimited	✓ ✓	✓ (3)
Display of "uncategorized" pseudo-category	• Displayed • Hidden	✓ ✓	✓
Display of empty categories	• Displayed • Hidden	✓ ✓	 ✓
Sort order of categories	• Alphabetic • Thematic • Numeric (largest first)	✓ ✓ ✓	 ✓
Actions / operations on overview	• Filter/narrow • Broaden • Exclude category • Hide category • Brushing and linking • Edit/restructure hierarchy	✓ ✓ ✓ ✓	✓ ✓ ✓ ✓

4.7 Summary of the SERVICE system

The SERVICE architecture and infrastructure support two working categorizing overview interfaces: AOL music search and general web search. These search

interfaces were developed in accord with the analysis and emerging design principles for exploratory search interfaces. They support multi-faceted exploration of large sets of search results, providing categorized overviews based on meaningful and stable categories. The general web search interface was evaluated in the third study, as reported in the next chapter, which helped to validate and refine the principles and analysis.

The SERVICE architecture defines a common Java interface to support easy plug-in of alternate category schemes. The technology is comprised of approximately 40 Java class files, which implement nine classifiers plus the two search interfaces. The two search interfaces use JavaServer Pages (JSP), hosted by an Apache Tomcat servlet container. The system runs on Windows and Linux, and uses the Java Database Connectivity (JDBC) API to integrate with MySQL and MS-Access databases. The system also implements a client-side logging facility that supports capture of any JavaScript events, including scrolling, mouse clicks and mouseovers, passing the timestamped events back to a Java-based logging tool. Four external data resources containing over 500 MB of data were processed to extract category information, using Java, Perl and PHP. The ideas embedded in the user interface will be useful to designers of other search interfaces, and it will be made available on the Categorized Search web page (http://www.cs.umd.edu/hcil/categorizedsearch/). The SERVICE system will be a flexible, extensible platform for additional research in categorizing search interfaces.

Chapter 5: Study 3: Categorized overviews using ODP and US government categories

Study 3 built on the results of the formative studies, scaling up from prototype to the working SERVICE 2.0 system. SERVICE enables support of general web search with the thematic classifications provided by the Open Directory Project (ODP). At least one commercial search engine (Exalead.com) has implemented categorized overviews of web search results using an adaptation of the ODP. However I am not aware of any studies of this approach.

The previous studies used a diverse set of participants and scenarios. That was desirable because of the formative nature of those studies. The queries and search results were fixed, and the search tasks were narrowly described. This third study used a narrowly tailored scenario – asking participants to generate newspaper article ideas for selected topics – that would be meaningful for a homogeneous group of study participants, recruited primarily from journalism students. This allowed participants to perform real web searches using the working SERVICE prototype and permitted data to be collected in a more realistic context, enhancing the external validity of the study. To minimize the impact on the study's internal validity, it was important to control the participant and scenario/task variables. This has been shown to be an effective way to balance the need for experimental control with realism when evaluating information retrieval systems (Borlund, 2000).

We first looked for a pool of subjects whose common background could be used as the basis for a simulated work task in an exploratory search scenario. Visiting Professor of Journalism Ira Chinoy was intrigued by the potential benefits of categorized overviews of search results for journalists and agreed to critique the study scenario and help recruit journalism students. He helped develop a narrowly tailored scenario that was meaningful for the homogeneous group of study participants.

5.1 Research questions

The research questions for this study were:

1. How do searchers think differently about their search tactics when categorized overviews are available to augment the result list?
2. What kinds of novel behaviors do searchers exhibit when categorized overviews are available?
3. How do the benefits of categorized overviews of search results for exploratory search observed in the first two studies compare with those observed in the domain of general web search? Specifically, how do searchers experience different topical perspectives or unusual/surprising results? Do they notice categories that are particularly well-covered by search results?
4. In what ways could the presence of categorized overviews affect the quality of the search outcome?
5. When categorized overviews are used, what differences can we identify for the above questions between broad and narrow topics?

Evaluating exploratory search interfaces is challenging. The nature of exploratory tasks can make it difficult to specify objective performance measures like time to completion, error rates, precision, or recall. Completing an exploratory task often involves developing and refining an information need that is specific to the individual. Mistakes and back-tracking are part of the process as searchers learn concepts and vocabulary. Documents that have great utility or novelty to one person may have little value to another, because of variations in domain knowledge, interests, and previously encountered information, so establishing ground truth for a measure of relevance is problematic. Evaluations have assessed and rated the quality of a task outcome to generate quantitative measures on a lesson plan creation task (Kabel, Hoog, Wielinga, & Anjewierden) or measured incidental learning that occurred during a search session (Pirolli, Schank, Hearst, & Diehl, 1996). Exploratory tasks have been decomposed or narrowed to constrain the task (Janecek & Pu, 2005). A combination of quantitative and qualitative evaluation methods have also been used (Yee, Swearingen, Li, & Hearst, 2003).

This study adopted the latter approach. Based on previous research (the formative studies and other studies), I expected to observe quantifiable and significant differences relative to the first three questions. They suggested hypotheses, described below, that were empirically tested. A qualitative approach extended the hypothesis tests by looking for phenomena not modeled by the research variables. For example, I expected that searchers would explore deeper in their result lists using the categorized overview, which was a testable hypothesis. I also anticipated that the interface would

prompt additional behavioral changes, but there was no *a priori* list; that would be developed from the data. Thus a combination of observation and semi-structured interview questions was used to investigate all five questions.

5.2 Experimental conditions

This study compared presentations of search results with and without categorized overviews. The categorized overviews were based on three facets: Topic, Geography and Government Agency. The topical facet (extracted from the ODP) classified web sites according to the 14 top-level categories shown in Table 21. The geographical facet was extracted from the ODP top-level category, Region. The US Government facet used a revised version of the hierarchy of the prior studies. The main difference was the addition of information to categorize state-level web sites. The topic and geography facets were chosen because they would apply to most search results and because they had been perceived favorably during the design and informal user testing. The government facet was included because participants in the earlier studies had commented on the credibility of government information. It was likely that the exploratory search scenario would induce subjects to look for government information because of its perceived credibility.

Web sites were categorized into the top 2 levels of each hierarchy. Unlike the previous studies, in which all sites were placed into leaf nodes, in this study sites could be cataloged into any level of the hierarchy. Based on observations during those studies, I did not expect users to find this problematic. This structure was consistent

with the organization of the ODP and simplified development of the prototype. The categorized overviews were thus comprised of three 2-level facets. Each facet included a top-level pseudo-category (called Uncategorized) in which pages not categorized within that facet were placed. This allowed searchers to narrow results to just those not categorized within the facet.

Table 21. Top level categories extracted from the ODP for the Topic facet.

Arts	Business	Computers
Games	Health	Home
Kids and Teens	News	Recreation
Reference	Science	Shopping
Society	Sports	

The study also attempted to investigate the effect of broad and narrow topics on the search process and outcomes. Because the topical facet was based on a general-purpose classification and limited to a depth of two levels, it forms a set of broad categories. I wished to investigate whether broader topics were more amenable to the categorized overview than narrower topics. Ultimately, the variability in the participants' perceptions of topics foiled a rigorous comparison, but it did provide illuminating qualitative results.

This study used a 2x2 within-subjects comparative design (N=24), with System (baseline or categorized overview) and Topic Type (broad or narrow) as the independent variables. The baseline condition presented search results as a typical ranked list, similar to Google (Figure 37). The experimental condition augmented the list with a categorized overview (Figure 38).

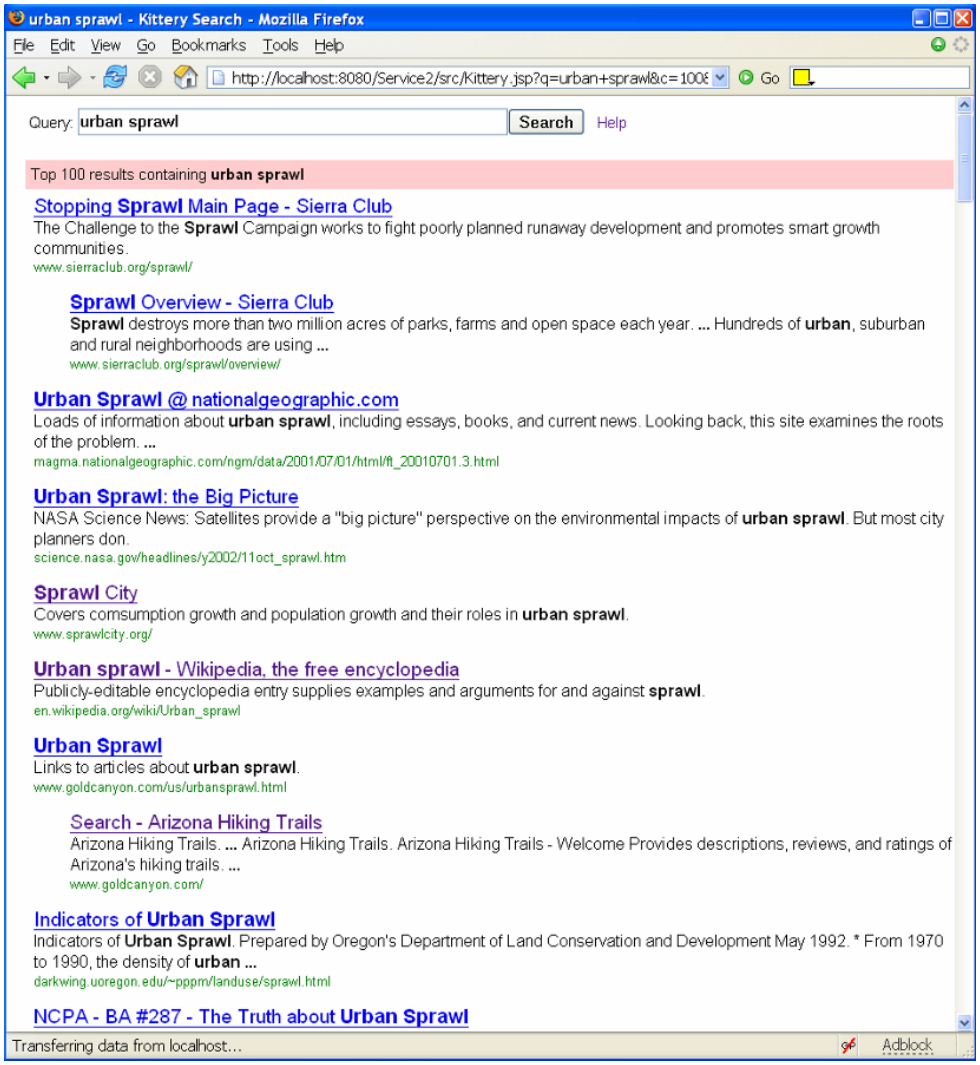

Figure 37. The baseline system (control condition) presented search results as a typical ranked list, similar to Google. It was referred to as the Kittery system in the study.

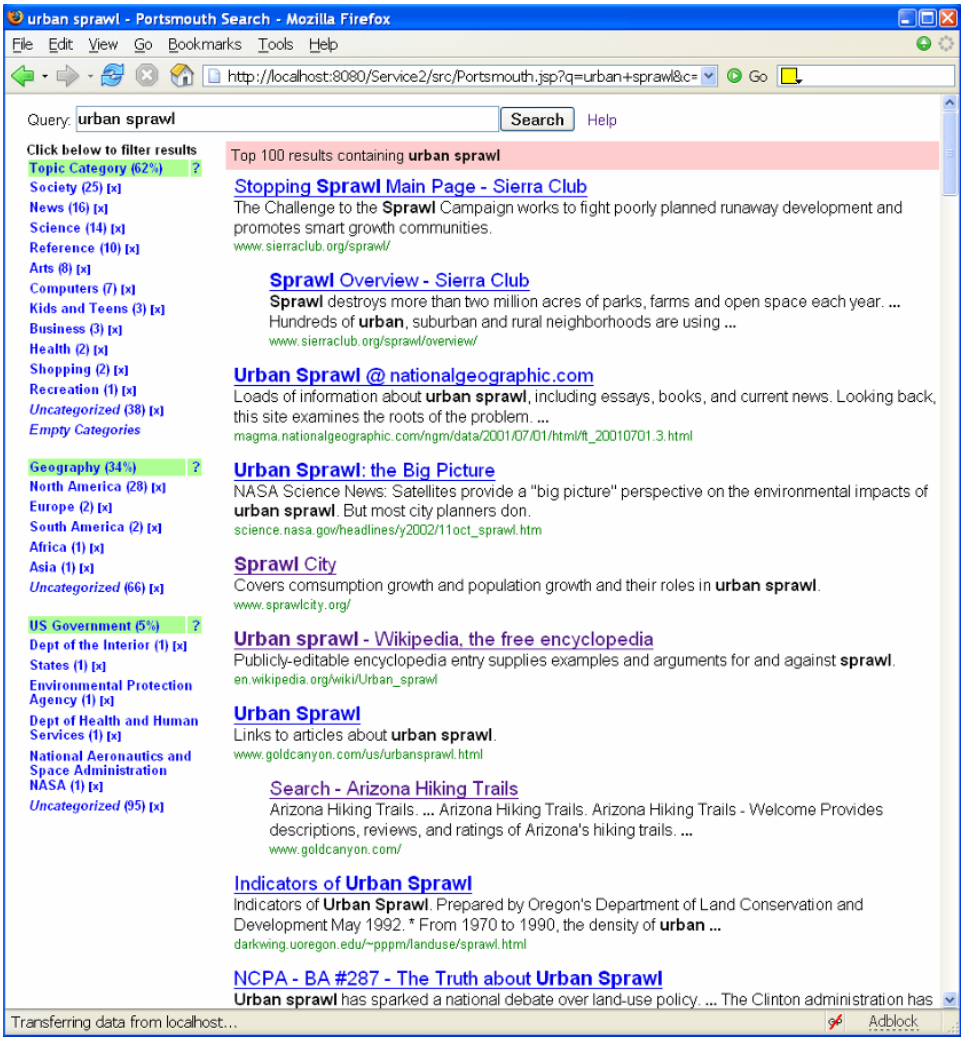

Figure 38. The experimental condition coupled the ranked result list with a categorized overview based on topical, geographical and US government classifications. This was referred to as the Portsmouth system in the study.

5.3 Scenario and task design

As in the formative studies, a high-level scenario was constructed around an exploratory search task. The task involved generating ideas for newspaper articles. Information seeking by journalists involves identification of an “angle” or perspective on the story. The angle is often structured as a working hypothesis that drives

development of the story. Information needs are often uncertain because of the fluidity of evolving plans, and the story angle can change – even at later stages of the information seeking process –in response to external events (such as breaking news) or internal needs (e.g., increasing or decreasing the desired story length) . Journalists work under tight deadlines, often with only hours between story assignment and filing. These characteristics guided design of the scenario and task. Professor Chinoy verified that the scenario and task were appropriate for the journalism students we would recruit as study participants. They were also verified as part of the exit interview. The scenario and task were described to participants as follows:

> *Imagine that you are a reporter for a national newspaper. Due to some recent events, your editor has just asked you to generate a list of ideas for a series of articles on [the topic, e.g. urban sprawl]. There's a meeting in an hour, so she doesn't need a lot of detail, but she wants a diverse list of 8-10 (or more) ideas for discussion. They should cover many different aspects of the topic, to appeal to a broad range of readers. Unusual or provocative ideas are good. You have about 10 minutes to conduct a short web search to find out what information is available and generate the ideas. Your results will be judged (by your imaginary editor) on the quality and diversity of ideas. For example, "public health impact" would be an okay idea. and "obesity as a public health impact of urban sprawl" would be even better, because it is a bit more specific. As you use the search engine to explore and generate article ideas, enter them in the Collector form and include the web page that inspired your*

idea. It is important that you enter the ideas, not notes like "a good page". Think of this list [point to the Collector] as a bullet list for the discussion.

Matched pairs of topics (broad and narrow) were developed using the following procedures (see Table 22). For the broad topics, an informal survey of the literature generated a list of potential topics. For each potential topic, a query was constructed from the topic terms. A Google search was conducted, and for those searches that produced at least 500,000 results, the top 100 results were categorized into the three facets and the percentage of categorized results within each facet was computed. Topics that had similar percentages between the three facets were used in various combinations during the early study design and the pilot testing, and a pair of topics that participants found similar was selected. A similar procedure was used to select the narrow topics, starting with 250 topics from the 2004 TREC Robust Topics, eliminating topics with specific geographic references. To further narrow the scope of the topic for participants, the TREC description field was adapted and included in the description that was provided to participants. This procedure did not ultimately have the intended effect of providing broad and narrow topics, and it is critiqued in section 5.11.1.

Table 22. Paired topics (broad and narrow) used for the study. This was the complete text read to the participants to describe the topic.

	Topic 1	**Topic 2**
Broad	Workplace allergies (WA)	The aging workforce (AW)
Narrow	Human smuggling - Human smugglers make money by smuggling, although the people being smuggled may or may not be willing participants. (HS)	International art crime - Includes theft, fraud or embezzlement in the international buying or selling of art objects. (IAC)

5.4 Hypotheses

The research questions entailed two kinds of hypotheses, process-oriented and outcome-oriented. Process-oriented hypotheses addressed questions related to how the interface affected the search process and attitudes. Outcome-oriented hypotheses addressed the question of how the interface affected the quality of the participants' generated ideas and overall progress toward the scenario goal.

5.4.1 Process-oriented hypotheses

The categorized overviews make more terms visible to the searcher, at the slight cost of reducing the number of individual results visible on the screen. The analysis suggests that this could induce searchers to examine results distributed more evenly throughout the list, in effect, more deeply within the list. Because relevant search results can be found well beyond the top 10 or 20 results, this behavior could be beneficial. Studies of clustered search results have observed this beneficial behavior. For narrow topics, the effect could be reduced either because of fewer results being categorized (i.e., more uncategorized results) or less cognitive overlap between the topic and searcher domain and category knowledge.

H1a. *Searchers will view (click on) results more deeply when using the categorized overview than when using the baseline.*

H1b. *When using the categorized overview, searchers will view (click on) results more deeply for broad topics than for narrow topics.*

If searchers view deeper pages, then they might also collect pages from more deeply within the list, too. Similarly, if searchers use the categorized overviews to filter results, this might collect more pages that have been categorized (into any category) instead of uncategorized pages. For narrow topics, the effect could be reduced because of less use of the categorized overview. Although viewing pages more deeply within the result list is considered beneficial, these two behaviors could indicate that the categorized overview biased searchers. In the context of the study scenario, the decision to collect a page depends on the searchers' assessment of the utility of the page, and specifically whether that page suggested an idea. Thus a higher-level cognitive factor could be involved.

H2a. *Searchers will collect results more deeply when using the categorized overviews.*

H2b. *When using the categorized overview, searchers will collect results more deeply for broad topics.*

H3a. *Searchers will collect a larger proportion of links from categorized facets (i.e. in ODP or government sites) when using the categorized overviews.*
H3b. *When using the categorized overview, searchers will collect a larger proportion of links from categorized facets (i.e. in ODP or government sites) for broad topics.*

If searchers are exploring each set of results more fully with the categorized overview, then they might issue fewer queries overall during the time allotted for searching (12 minutes). For narrow topics, the effect could be reduced because of less use of the categories.

H4a. *Searchers will issue fewer queries with the categorized overviews.*
H4b. *When using the categorized overviews, searchers will issue fewer queries for broad topics.*

The analysis suggests that the availability of categorized overviews will provide more information to the searcher and give them additional control over their results. This should lead them to rate the categorized overview interface higher than the baseline for organizing, exploring and gaining an overview of the results, finding useful pages and several measures of user satisfaction. The additional display and interaction elements, and the need to make more search decisions, could also cause users to perceive the categorized overview as more complex. For narrow topics, the effect could be reduced because of less use of the categorized overview.

H5a. *Searchers will find it easier to explore search results with the categorized overview.*

H5b. *When using the categorized overview, searchers will find it easier to explore search results for broad topics.*

H6a. *Searchers will agree more strongly that the system provided a good overview of information available about this topic on the Web when using the categorized overview.*

H6b. *When using the categorized overview, searchers will agree more strongly that the system provided a good overview of information available about this topic on the Web for broad topics.*

H7a. *Searchers will agree more strongly that the system organized the results well when using the categorized overview.*

H7b. *When using the categorized overview, searchers will agree more strongly that the system organized the results well for broad topics.*

H8a. S*earchers will agree more strongly that the system helped them assess results and decide what to do next when using the categorized overview.*

H8b. *When using the categorized overview, searchers will agree more strongly that the system helped them assess results and decide what to do next for broad topics.*

H9. *Searchers will rate the categorized overview easier to use than the baseline.*

H9b. *When using the categorized overview, searchers will rate the system easier to use for broad topics.*

H10a. *Searchers will rate the categorized overview more stimulating than the baseline.*

H10b. *When using the categorized overview, searchers will rate the system more stimulating for broad topics.*

H11a. *Searchers will rate the categorized overview more "wonderful" than the baseline.*

H11b. *When using the categorized overview, searchers will rate the system more "wonderful" for broad topics.*

H12a. *Searchers will rate the categorized overview more satisfying than the baseline.*

H12b. *When using the categorized overview, searchers will rate the system more satisfying than the baseline.*

H13a. *Searchers will rate the categorized overview more complex than the baseline.*

H13b. *When using the categorized overview, searchers will rate the system more complex than the baseline.*

5.4.2 Outcome-oriented hypotheses

Categorized overviews may enable searchers to make more connections between the search results displayed to them and their existing knowledge. Although there are more intervening variables between interface and outcomes than between interface and behavior, the categorized overviews might help searchers become more familiar with the topic, find more useful information, make more progress towards the scenario goal, and produce better article ideas. For narrow topics, the effect could be reduced either because of less use of the categorized overview or less cognitive overlap between the topic and searcher domain and category knowledge.

H13a. *Searchers will feel more familiar with the topic with the categorized overview.*
H13b. *With the categorized overview, searchers will feel more familiar with the topic for broad topics.*

H14a. *Searchers will find more useful information with the categorized overview.*
H14b. *With the categorized overview, searchers will find more useful information for broad topics.*

H15a. *Searchers will make more progress toward the scenario goal with the categorized overview.*
H15b. *With the categorized overview, searchers will make more progress toward the scenario goal for broad topics.*

H16a. *Searchers will produce higher quality article ideas with the categorized overview.*

H16b. *With the categorized overview, searchers will produce higher quality article ideas for broad topics.*

5.5 Participants

Twenty-four participants (5 male, 19 female) were recruited primarily from the University of Maryland's Philip Merrill College of Journalism and paid $30 for their participation. Campus colleagues agreed to distribute an email solicitation to mailing lists. Respondents were asked to complete an online questionnaire to collect basic demographic information, yielding a pool of 59 potential participants. This included journalism and non-journalism students. The journalism students were invited to sign up for experiment sessions via an online scheduling form, yielding 20 participants. Non-journalism students were then invited to sign up for the remaining sessions. They ranged in age from 18 to 27 years, with a median age of 20. Twenty-one were undergraduate students, one was a graduate student and two had graduate degrees. Participants reported being experienced and proficient at web searching. All reported at least three years of search experience, and all but two reported searching at least once per day, with two reporting searching 1-2 times per week. They all reported being successful in their searches "Most of the time" or "Always or almost always." When asked to rate their search skills on a 1-5 scale (1 = novice, 5 = expert), nine reported a 3, twelve reported a 4, and three reported a 5. All used the Google search engine and 14 reported using other search engines. All reported performing searches

for class research, 23 searched for entertainment or recreation, and 22 searched for news and information on events.

5.6 Materials

5.6.1 Interfaces

The search interfaces were assigned neutral names (Kittery for the baseline and Portsmouth for the experimental) and displayed alongside a small web application, the Collector form (Figure 39). The Collector form provided fields to capture ideas and the relevant URLs, and listed them in reverse chronological order so participants could refer to them during the session. The screen resolution was 1280x1024 pixels. Prior to search, the search window was set to 1024 pixels wide and the collector window to 256 pixels wide.

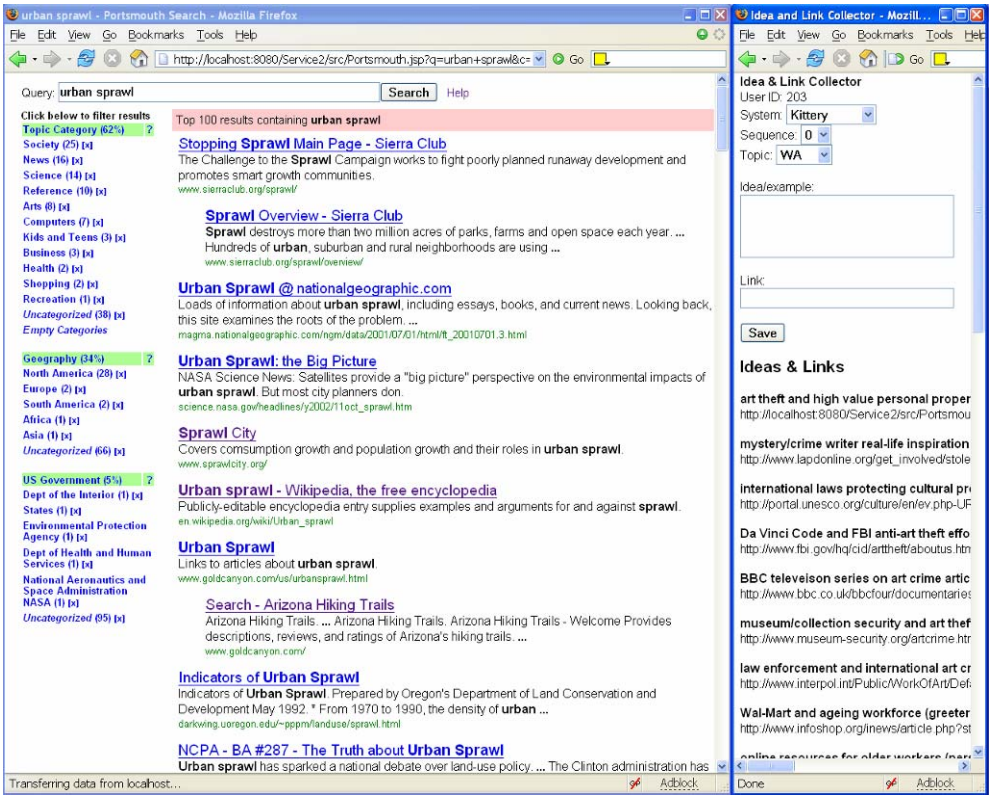

Figure 39. The interface used by participants was comprised of the system under test (left) and the Collector form (right).

5.6.2 Script and training videos

A written script provided participants with background information on the study, to describe the scenario and task and to introduce the training task. Three short (1-3 minute) training videos, produced using Camtasia Studio, introduced participants to the two interfaces and the Collector form.

5.6.3 Online questionnaires

Three online questionnaires were used during the experimental sessions (see Appendix D). An entry questionnaire collected participants' demographic and search experience data. A pre-search questionnaire captured knowledge of and interest in

each topic prior to the search. A post-search questionnaire repeated the pre-search questions and collected reactions to the topic, interface and search process. Paper print-outs of all forms were available in case of communication problems with the external server (but were never needed).

5.6.4 Paper forms

The informed consent form was approved by the University of Maryland Institutional Review Board (see Appendix C). A payment acknowledgement form was used to verify that subjects had received payment for their participation in the study. One paper checklist ensured completion of all parts of the experimental procedure in the correct order, and another checklist ensured that participants were exposed to the basic system features and task elements during the training task. The exit interview questions were read to the participants from a paper form.

5.6.5 System technology

Participants used an IBM T42p laptop with a 15 inch display, 1 GB of RAM, and a 1.8 GHz Intel Pentium M processor running Windows XP Professional. An external keyboard and mouse were attached, with an external pair of speakers for the training videos. Camtasia Studio 3 was used to capture screen video and audio, with a desktop microphone. The SERVICE 2 web search prototype was configured in two versions (one for each interface), both with logging enabled to capture category and result list clicks, as well as mouseover and scroll events. A Tomcat 5 server running on the laptop hosted the search application, interfacing to the search engine, managing the fast-feature classifiers, and generating the user interfaces. It also hosted the Collector

application. The applications connected via JDBC to a MS-Access database that was used to cache search results and to store the ideas and links. An Apache web server was configured as a proxy server to log all pages visited during the experiment sessions. This was desirable because the JavaScript based logs only capture web pages directly visited from the search results page. An open source web survey system, phpESP (phpesp.sourceforge.net), was adapted for the online questionnaires. This was hosted on a Redhat Linux server with a MySQL database to store questionnaire results. Participants used the Mozilla Firefox browser (v. 1.0.7) configured to use the proxy server for non-local HTTP requests. The laptop was connected to the Internet via the campus T3 connection.

5.7 Procedure

The experiment sessions were individually conducted in an office on the University of Maryland campus (Figure 40). As participants arrived, they were welcomed and provided a short introduction to the study, informed that they would be asked to perform four searches, answer several questionnaires, and that they would receive $30 at the end of the session. After an opportunity to ask questions, they signed two copies of the informed consent form. They were invited to adjust the chair, keyboard and mouse for their comfort and offered water and some candy snacks. After they signed the informed consent form, they completed the online entry questionnaire, providing demographic and search experience data, and viewed the training video appropriate to the first interface condition. Following the video, the scenario and task were described, and they used the system for a training task on the topic “urban sprawl.” They were encouraged to ask questions and think out loud, using a think-

aloud protocol (Ericsson & Simon, 1984). The training checklist ensured that they used the basic system features on their own or with prompting. When the checklist was completed, they were asked if they had any questions and if they were ready to continue.

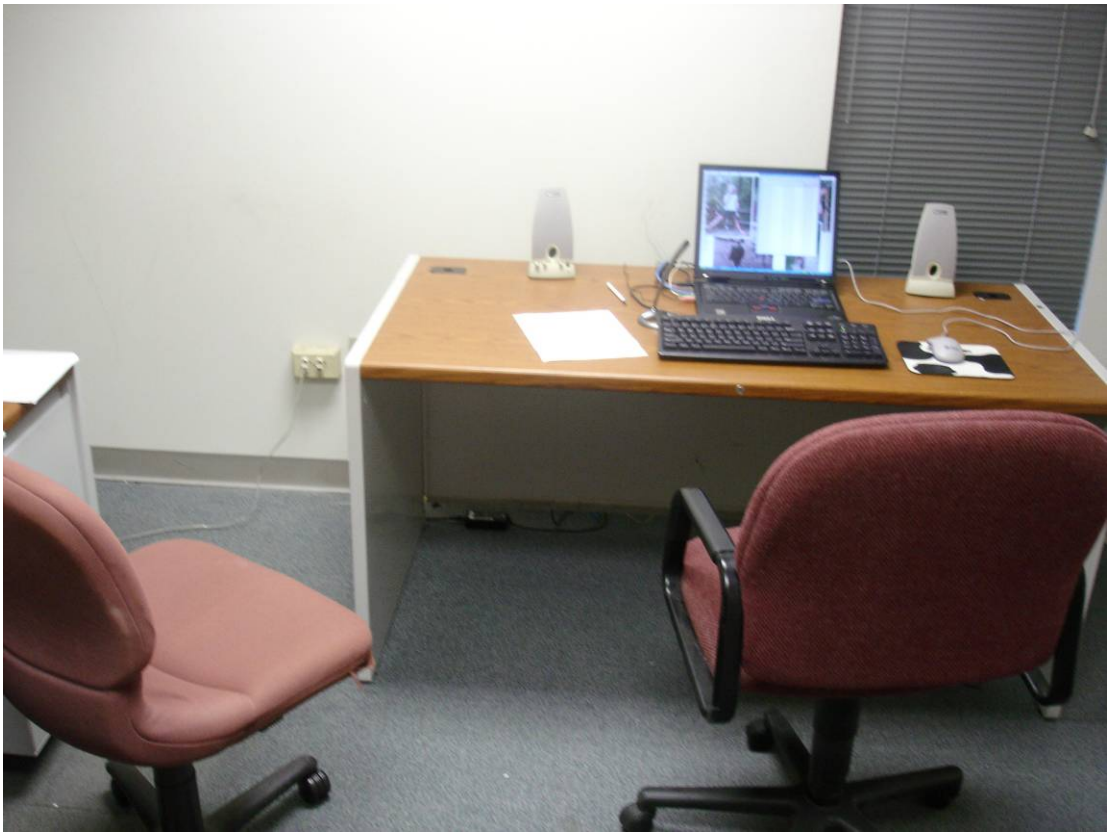

Figure 40. The experimental setup. Study participants sat in front of the computer, and the observer sat to their left.

They were then presented with the first topic. They completed the online pre-search questionnaire, performed the timed search and completed the post-search questionnaire. This was repeated this for the second topic. After a short break, they were shown the second interface and given time to become comfortable with it. The

remaining two searches were then completed as before. The session concluded with a semi-structured exit interview and payment of the $30.

The order of the training videos varied slightly depending on the interface presentation order. When the baseline interface was used first, the video for the collector form only was shown prior to the training task, and then the video for the experimental interface was shown after the break, immediately before they would use that interface. When they used the experimental interface first, they viewed the video for both the collector and the experimental interface prior to the training task. An alternative approach would have shown all videos and conducted all trainings at the beginning of the session. After discussion with colleagues, I decided that participants would be more likely to forget how to use the experimental interface if they weren't shown it immediately prior to use.

5.8 Pilot testing

Before conducting the study, I pilot tested portions of the materials and procedures with six participants, and then ran the entire final experimental protocol with six others. During the pilot testing, various pairings of broad and narrow topics were used to select the final pairing. I observed how participants responded to the topics, and after the session asked them to compare the pairs of topics, and used this feedback to select the final pairs of topics for each topic type. The training time was extended to permit participants to work until they felt comfortable with both the systems and the task, and the scenario and task descriptions were refined. The final pilot tests

confirmed that the session duration was about two hours and 15 minutes, including about 30 minutes for the semi-structured exit interview.

5.9 Analysis methodology

5.9.1 Quantitative analysis methodology

The quantitative data to be analyzed included the original location of items in the search result lists, counts, and subject preferences rated on an interval scale. In all cases, the null hypothesis was that there was no difference between the groups. A p-value of 0.05 was used to reject that hypothesis, yielding a 5% chance of incorrectly rejecting the null hypothesis (a false positive, or type I error). Marginally significant differences ($p < 0.10$) are also reported. Except where noted statistical tests were performed with the R Statistics package, version 2.2.0 (R Development Core Team, 2005). R is an open source implementation of the S language and environment which was developed at Bell Laboratories.

Search result location data

The original location of selected or collected items in the search result list was treated as an interval scale. A professional statistician confirmed that an ANOVA analysis would be appropriate for these data. For all significant ANOVA results, the normal Quantile-Quantile (Q-Q) Plots were examined to confirm that the residuals were distributed normally. Where the raw data did not follow a normal distribution, the raw data was transformed using a logarithmic transform, an accepted technique for handling non-normal distributions (Jaccard, 1983).

Collecting pages from categorized facets

When entering ideas into the idea collector, searchers simultaneously entered the URL of the page that prompted the idea, referred to here as the collected page. For each collected page a boolean variable (InAnyFacet) was computed indicating whether the page was found in any of the facets (topic, geographic, or government). Chi square analyses were used to test the relationship between the InAnyFacet variable and the System and Topic variables. For the System analysis, where there contingency table contains two rows and two columns, the Yates' continuity correction was applied. This is a commonly, albeit not universally, applied adjustment intended to provide a better estimate of the significance level (Jaccard, 1983). A factorial logistic regression analysis would also have been appropriate, and would have allowed me to additionally investigate the interaction between System and Topic variables. However, this was not a compelling interest for this study, and the Chi square tests are simpler to interpret and report.

Quality of generated ideas

The quality of the generated ideas was assessed using newsworthiness criteria suggested by Professor Chinoy. High quality ideas would pose a question or paradox, contain conflict and human interest elements, indicate the context of the idea, and reflect intangible elements such as “coolness”. Other factors included timeliness, potential impact, and proximity. Diverse ideas were preferred, and redundant ideas were ignored. Each idea was rated on a scale from 1 (poor) to 9 (excellent) by a single judge (the researcher). The ideas were assessed blind, without knowledge of the

system used or participant, using a MS-Access form. Two passes were made through the ideas for each topic. This allowed me to become familiar with all the ideas before assigning final quality rating.

To test the relationship between the Idea Quality variable and the System factor, a Wilcoxon rank sum test was used. This is the non-parametric equivalent of the independent samples t-test. It is appropriate here because the independent variable (System) is categorical with two levels, and the dependent variable (Idea Quality) is ordinal. To test the relationship between Idea Quality and Topic, a Kruskal-Wallis test was used. It is the non-parametric equivalent of a one-way ANOVA, and is used here because the independent variable is categorical with four levels. As the results below indicate, no statistically significant differences were detected for the System factor. If any had been detected, a second assessor would have rated the ideas, and the inter-rater agreement would have been checked to provide a more rigorous assessment.

Subjective measures

Subjective measures used Likert scales and semantic differentials on a 9 point scale. ANOVA statistics were used as above to identify significant differences.

5.9.2 Qualitative analysis methodology

The research questions were addressed qualitatively by direct observation, review of selected video and participant response to questions, and by a limited quantitative analysis of responses to three selected questions. Three forms of raw data were

available for this purpose. First, all sessions, including training, searches and interviews were recorded and participants were instructed to think out loud while they searched, which enabled me to flag interesting actions or comments in my observation notes and then review the sessions afterwards. This provided a total of about 100 minutes of audio and video per session. Second, immediately after each search, subjects were asked, "What are your thoughts at this point?" They were asked to respond verbally or in written form on the post-search questionnaire. This typically yielded a 1-2 minute reply or 3-5 written sentences. Third, the exit interview included 10 open-ended questions, usually lasting 20-35 minutes.

Participants were instructed to think out loud during their searches, with the following request:

> *Please think out loud as you take each action, for example, when you enter a query, click on something, or scroll a page. Briefly say why you did it and then tell me your reaction to the system's response. I'm also interested in what's good or bad, problems or insights, and anything confusing.*

During the training task, they were encouraged to think out loud, but due to the limited amount of time for each session, they were not specifically trained in the use of the think-aloud technique (Ericsson & Simon, 1984). Instead they were encouraged during the training tasks and prompted several times at the beginning of each search. Several of the participants responded well and provided useful concurrent reports. Others were more reticent during the search, but all subjects responded eagerly during the exit interview.

Three open-ended questions from the exit interview were chosen for a detailed quantitative analysis. These questions related directly to the research questions, and I expected that the responses would help identify the concepts and issues that were most salient to searchers as they reflected on their experience. The selected questions were:

- *Did the categorized overview change the way you searched? Can you describe an example?*
- *Can you describe an example where the categorized overview [helped; OR hindered, frustrated or mislead – whichever not indicated in previous question]?*
- *Did you notice any difference in how you used the categorized overview each time? Can you describe an example?*

Notice that in the second question, the object was to elicit feedback on whichever aspect (positive or negative) the participant did not mention when answering the first question.

Responses for each question were transcribed into an Access database table, and an inductive approach was used to develop and assign an initial code list. Although a qualitative data analysis tool such as NUD*IST or NVivo could have been used, the availability of Access and the limited nature of the analysis made Access preferable. Each response was reviewed, noting salient comments that appeared relevant to the research questions, and assigning a short label to sets of related comments. After

responses from 12 participants were transcribed and coded, the codes were reviewed and obvious duplicates were merged. These codes were divided into five groups:

- Behavior differences
- Cognitive and affective impacts
- Judgments of outcomes
- Facet usage
- Miscellany

The code also noted whether the comment reflected a positive or negative judgment by the participant (some comments were neutral or did not have a judgment element). Each response was then entered in the Access table. Before the remaining 12 responses were coded, the code list was again reviewed. A second full pass was conducted to review the initial code assignments and assign a small number of new codes. This yielded a set of 64 codes (see Appendix E for complete list). The five code groups were used to organize the subsequent analysis, and individual code values were used to prompt consideration of specific behaviors, judgments, etc. I reviewed my notes, participant responses to the interview questions and the session recordings to analyze each code.

Validity

This analysis represents a principled approach answering to the research questions, drawing on the naturalistic inquiry paradigm (Guba & Lincoln, 1982). It complements the quantitative analysis, which seeks to identify commonalities across search experiences, by illuminating differences in search experiences. As a step

toward validating the analysis, this section has been peer-reviewed by a colleague in the College of Information Studies, Katy Newton Lawley.[2]

The three interview questions required introspection and reflection. Introspection and reflection can allow the investigator to gain access to thoughts that are "mediated by knowledge structures or artefacts that we design and use," (Nielsen, Clemmensen, & Yssing, 2002) Categorized overviews are designed expressly to expose specific knowledge structures, thus this form of analysis is appropriate for examining responses to categorized overviews. Verbal reports, and retrospective reports in particular, are subject to known problems and limitations (Ericsson & Simon, 1984). Respondents may misremember a thought or action, or inadvertently use inferences instead of memory. The form of the verbal probe or even its emphasis can affect the information provided. Subjects were asked to report on aspects of their thoughts and actions that they did not necessarily attend to at the time of the interaction. Inevitably, respondents make judgments about past thoughts, decisions or actions that emphasize some and distort or overlook others. To minimize these problems, the questions were constructed to elicit specific examples and concrete details in conjunction with reflection/introspection.

[2] This section has benefited from Ms. Lawley's critique and advice, but any errors or deficiencies in this section are, of course, the responsibility of the author.

5.10 Results

These sophisticated users coping with challenging search tasks over a two hour period produced a wealth of data. The quantitative results provide a baseline for future studies while showing some differences in behavior and strong preferences. They do not show objective differences in outcomes. The qualitative data include thoughtful comments indicating many strengths and some weaknesses of the categorized overviews.

5.10.1 Quantitative results

5.10.1.1 Breadth of Topics

The original intent of the analysis was to consider the pairs of broad (aging workforce and workplace allergies) and narrow (human smuggling and international art crime) topics as equivalent. This would have permitted a within-subject analysis, but subject comments during the first dozen sessions raised questions about the validity of this matched pair assumption. For the latter half of the sessions, a question was added that specifically asked participants to rate the breadth of the topics. Analysis of their responses using a one-way ANOVA indicated no significant differences. This confirmed that participants did not perceive topic breadth consistently (Figure 41). In the analysis, the Topic factor was therefore treated as a between-subjects variable with four levels: aging workforce (AW), human smuggling (HS), international art crime (IAC), and workplace allergies (WA). Where significant differences by topic were detected, the Camtasia video and observation notes were reviewed to investigate the differences.

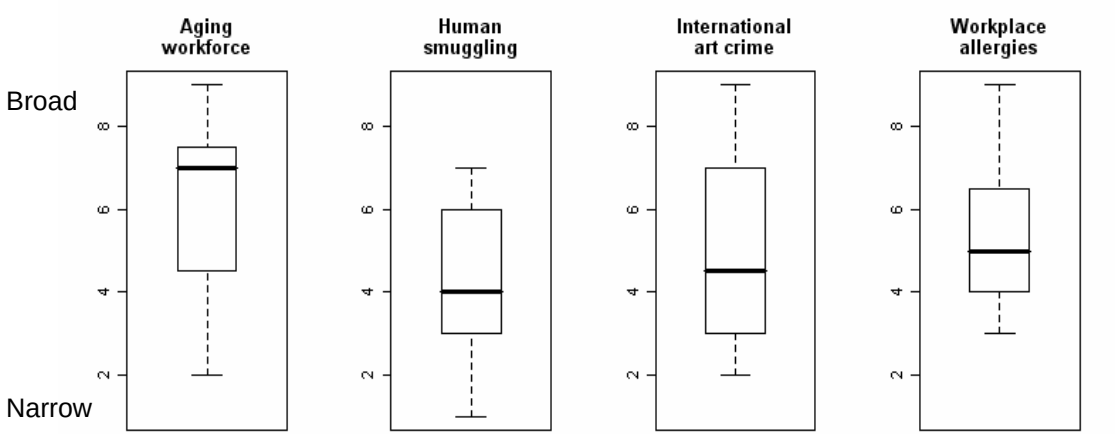

Figure 41. Subject assessment of topic breadth (N=12). Participants did not perceive the breadth of the topics significantly differently.

5.10.1.2 Original location of viewed (clicked-on) pages in search result list

Searchers viewed (clicked on) a total of 924 pages from the search results. A histogram of the raw data reveals that the results are highly skewed (Figure 42). The log-transformed values do appear to be normally distributed, so they are used here. The results of a 2 (system) x 4(topic) factorial analysis indicated a significant difference by system $F(1,919)=8.96$, $p<0.01$ and by topic $F(3,919)=5.73$, $p<0.01$, and a marginally significant interaction between system and topic $F(3,919)=2.19$, $p<0.10$. The Normal Quantile-Quantile (Q-Q) plot shows that the residuals are moderately skewed, but not enough to invalidate the ANOVA results (Figure 43). Searchers viewed pages at a mean (median) depth of 28.4 (18) when using the categorized overview, whereas they viewed pages at a mean depth of 22.3 (12) with the baseline. The plot in Figure 45 shows modest but noticeable differences in the distribution of viewed pages of views. With the categorized overview, searchers viewed results from a broader portion of the result list.

Tukey post-hoc tests on topic indicated that there were significant differences between two of the four topic pairs: IAC-AW and IAC-WA. The mean (median) depth of pages viewed for IAC was 20.2 (10), whereas the mean depths for AW and WA were 29.6 (19) and 27.0 (20), respectively. In general, searchers searched somewhat more deeply with the categorized overview than the baseline across all topics (Figure 46 and Figure 47). The exception was with HS, where the mean depth was the same between systems. The log data indicates that two of the subjects who used the baseline system for the HS topic viewed substantially more pages than the other subjects (22 and 18, versus an average of 9.4 for the other 22 subjects). Video of those two searches shows that both subjects scrolled far down the result list and selected several pages from deep in the list. One of these subjects indicated that she found the baseline interface easier to use.

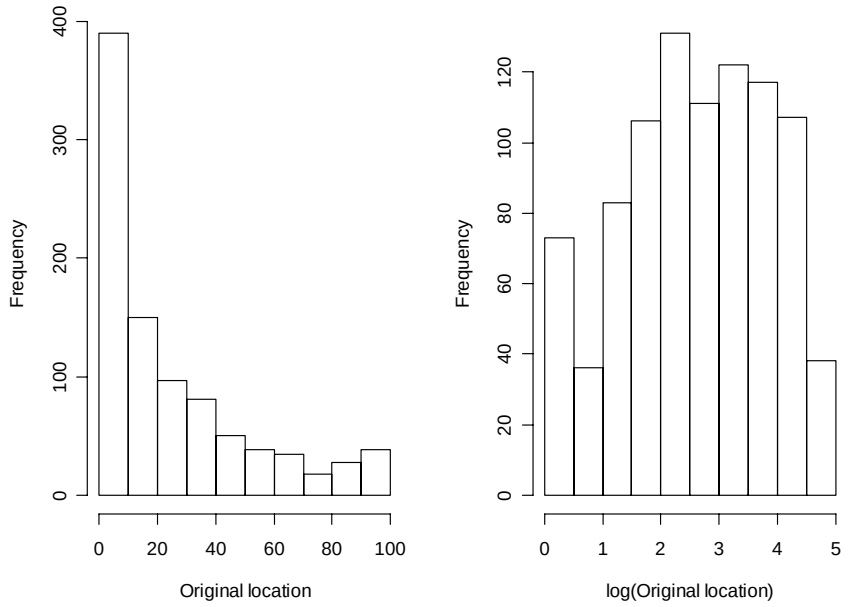

Figure 42. Histograms of a) original location of search result in list, and b) log(original location).

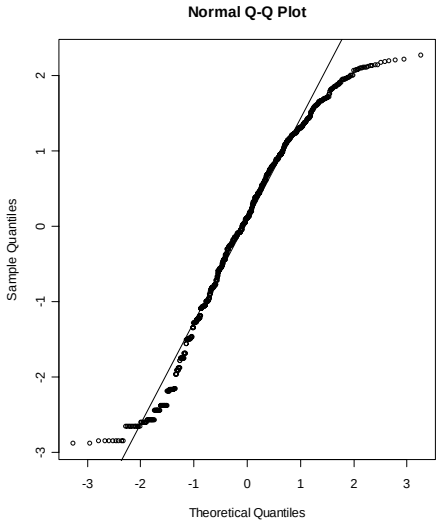

Figure 43. Normal Quantile-Quantile plot of the residuals for the log(original location) model. Residuals are moderately skewed, but not enough to invalidate the ANOVA results.

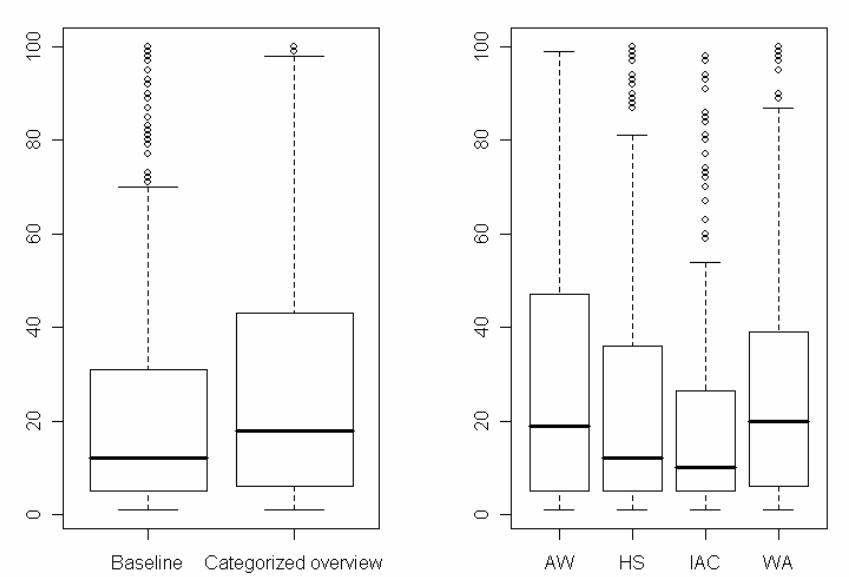

Figure 44. Original location of viewed pages in search results, a) by System[*], and b) by Topic[+] (N=924). (Note: For all boxplots, the bold line in the middle of the box indicates the median; the upper and lower boundaries of the box indicate the first and third quartiles, and the whiskers extend 1.5 times the interquartile range from the box. For all figures, statistically significant differences, p<0.05, are marked with an asterisk in the caption, and marginally significant differences, p<0.10, are marked with a plus sign.)

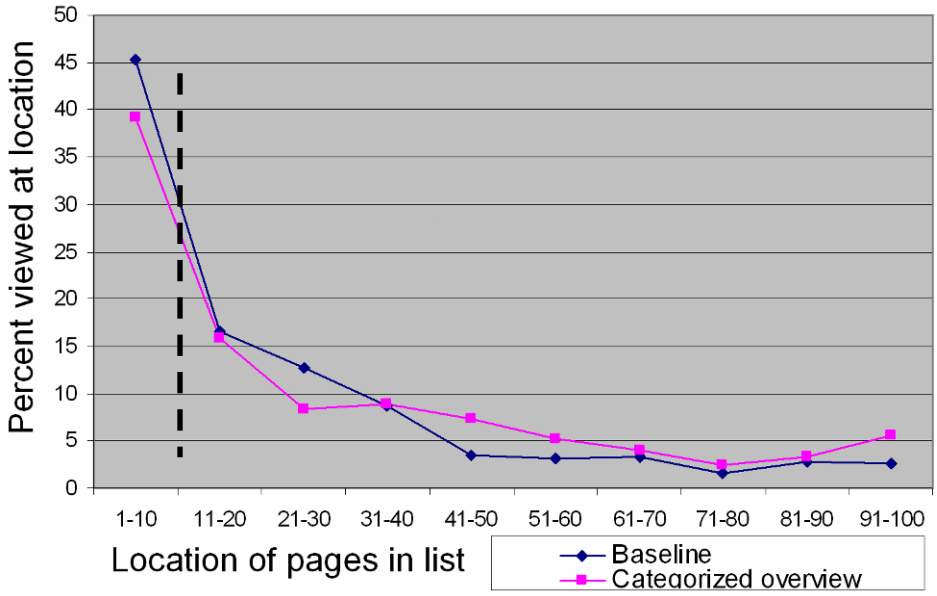

Figure 45. Percent of pages viewed by original location of page within search results, for each system. The interface displayed approximately 10 results per screen. The dashed line shows the initial screen break.

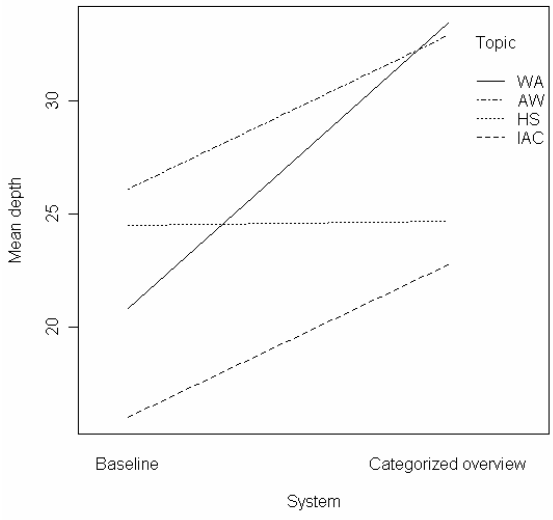

Figure 46. Interaction plot of mean depth of viewed pages for System and Topic factors. Except for the human smuggling topic, searchers viewed pages more deeply using the Categorized overview system. The largest change between systems was for the workplace allergies topic.

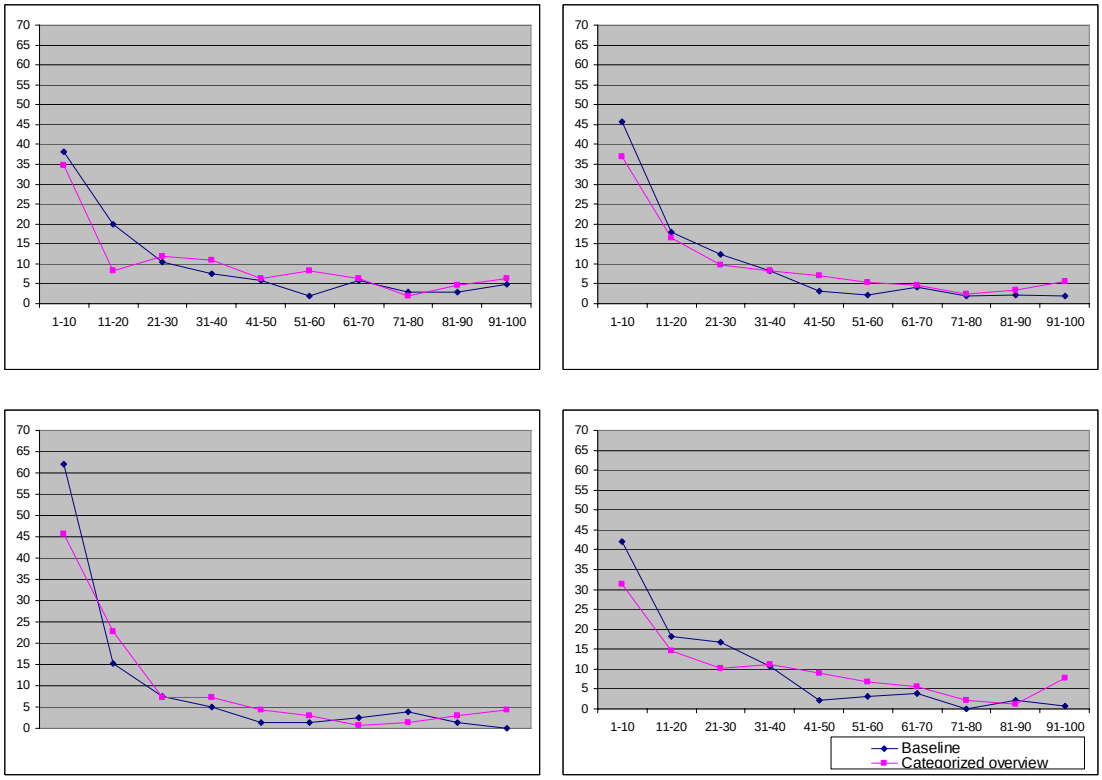

Figure 47. For each topic, percent of pages viewed by original location of page within search results.

5.10.1.3 Original location of collected pages in search result list

Searchers collected 611 pages from the search results during their searches. As with the viewed pages, the raw data are skewed, but the log-transformed values appear to have a normal distribution (Figure 48). The results of a 2 (system) by 4 (topic) factorial analysis were not significant, although there was a marginally significant effect for topic $F(3, 603) = 2.48$, $p<0.10$. The mean (median) depth of collected pages was 26.1 (16) for both the baseline and categorized overview. The histograms in Figure 50 and the plot in Figure 50 show that the distribution of original locations is similar between the two systems.

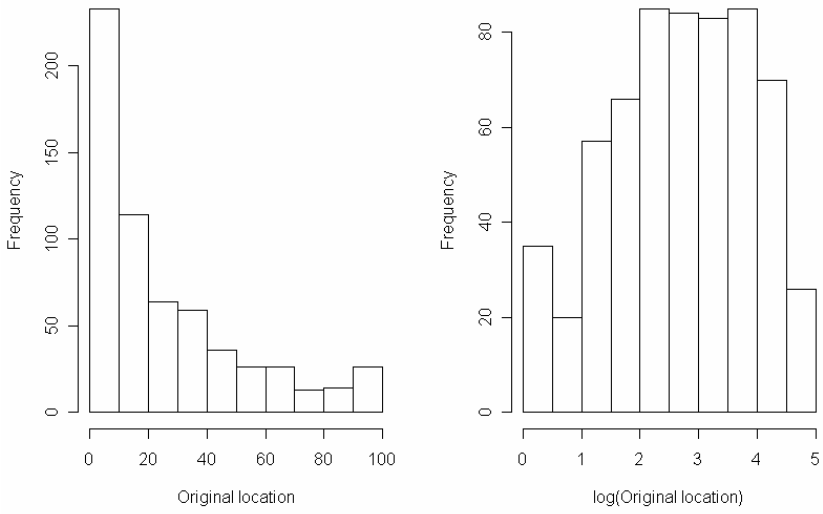

Figure 48. Histograms of a) original location of collected pages, and b) log(original location).

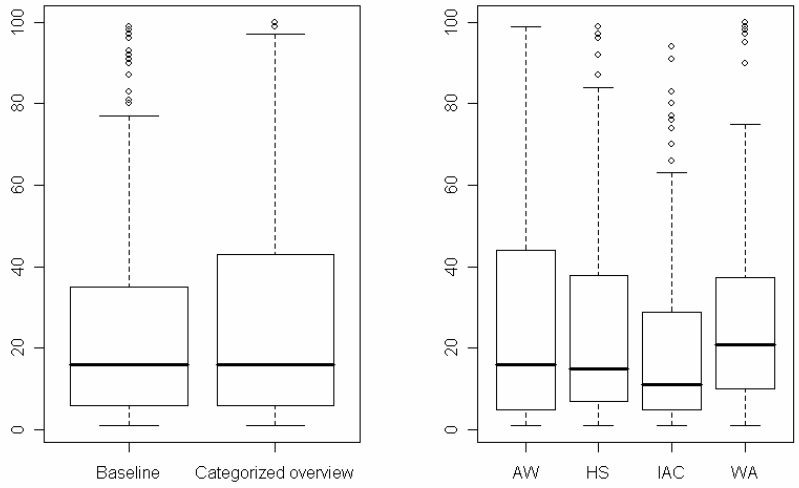

Figure 49. Original location of collected pages, a) by System, and b) by Topic^{+} (N=611).

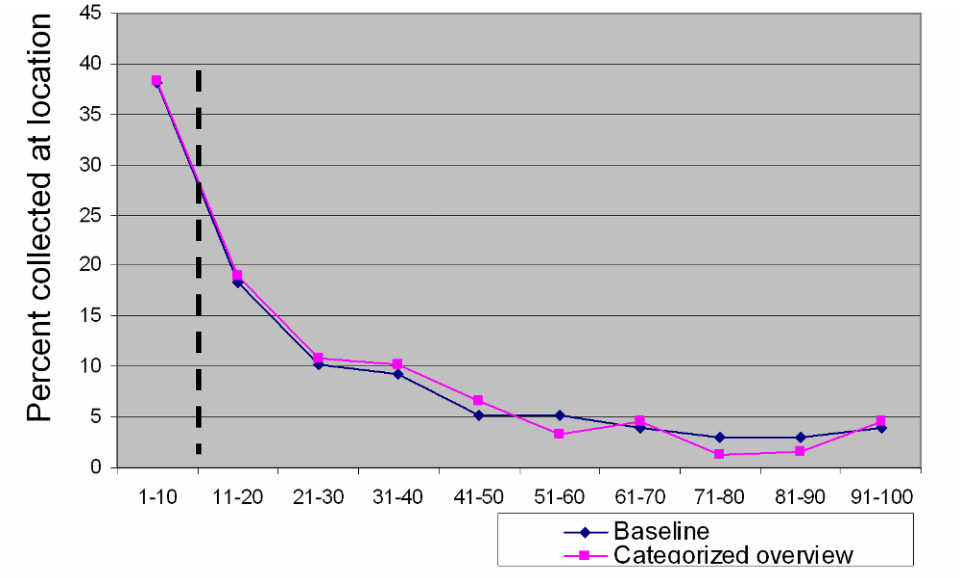

Figure 50. Percent of pages collected by original location of page within search results. The interface displayed approximately 10 results per screen. The dashed line shows the initial screen break.

The interaction diagram in Figure 46 shows that the largest change in mean depth of viewed pages between systems occurred with the workplace allergies topic. To see if this change was reflected in the pages they chose to collect, the original location of collected pages was computed for each topic. Figure 51 shows that for this topic participants collected more pages from the lower 40 locations in the result list and fewer from the top 30 locations.

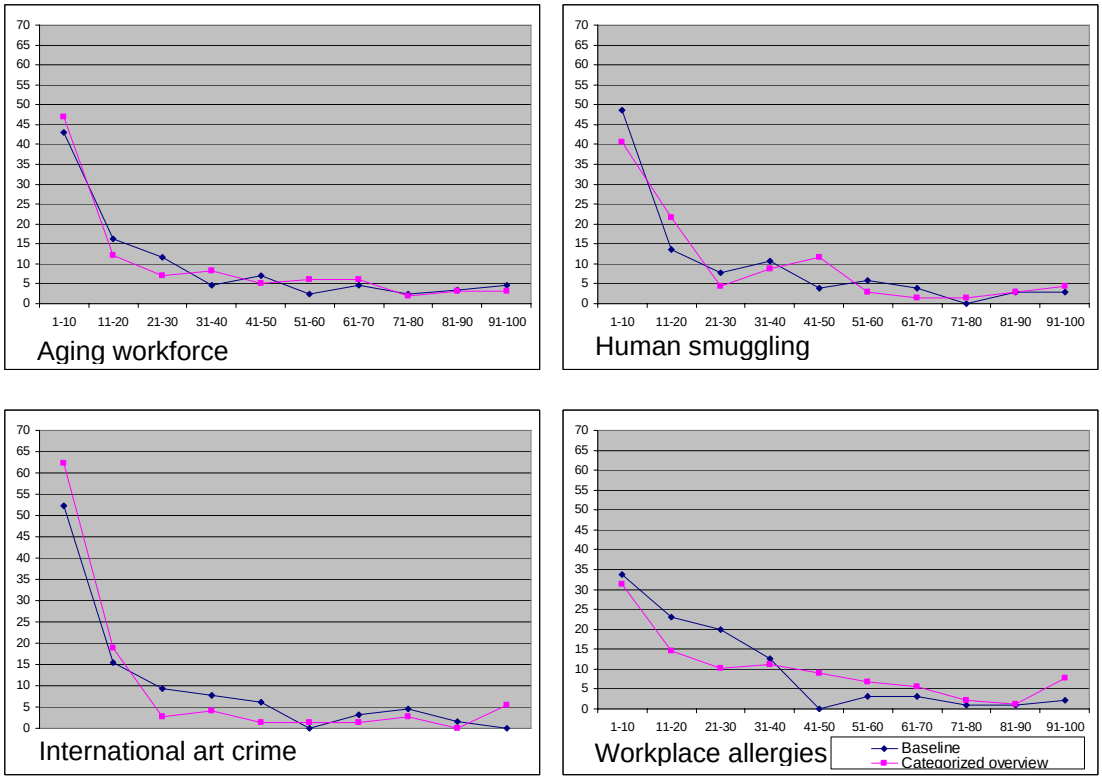

Figure 51. For each topic, percent of pages collected by original location of page within search results.

5.10.1.4 Proportion of pages collected from categorized facets

Searchers collected a total of 679 pages, including pages that were not in any search results (e.g. pages found by following links). For each collected page a boolean variable (InAnyFacet) was computed indicating whether the page was found in any of the facets (topic, geographic, or government). The proportion of categorized pages differed significantly by System, $\chi^2(1, N = 679) = 5.11$, $p < .05$, and Topic, $\chi^2(1, N = 679) = 18.00$, $p < .001$. The difference for the System factor is 7.5 percentage points, suggesting that the categorized overview biased searchers toward categorized pages. The percentage for the workplace allergies topic is substantially different from the

other topics (12-17 points). This may be related to the relative ease of finding information about the workplace allergies topic, which several participants commented on.

Table 23. Percent of collected pagesthat had been categorized, by System*.

System	Percent Categorized
Baseline	75.4 %
Categorized overview	82.7 %

Table 24. Percent of collected pages that were categorized, by Topic*.

Topic	Percent Categorized
Aging workforce	85.3 %
Human smuggling	80.2 %
International art crime	82.7 %
Workplace allergies	68.5 %

5.10.1.5 Number of queries issued during searches

Searchers conducted a total of 96 searches. All subjects except one issued at most 10 queries. One subject issued 15 queries during a search, and that outlier is removed from the following analysis. The raw data are skewed, but the log-transform values are somewhat more normally distributed (Figure 52). The results of a 2 (system) x 4 (topic) factorial analysis indicated a significant difference by system $F(1,87)=7.15$, $p<0.01$ and by topic $F(3,87)=3.63$, $p<0.05$. The Normal Q-Q plot shows that the residuals are reasonably distributed (Figure 53). The mean (median) number of

queries per search was 3.0 (2) for the categorized overview system and 3.5 (3) for the baseline. Tukey post-hoc tests on topic indicated there were significant differences between the WA and IAC topics. The mean (median) number of queries per search was 3.3 (3) for the WA topic and 3.1 (3) for the IAC topic. The magnitudes of differences, although statistically significant, are modest.

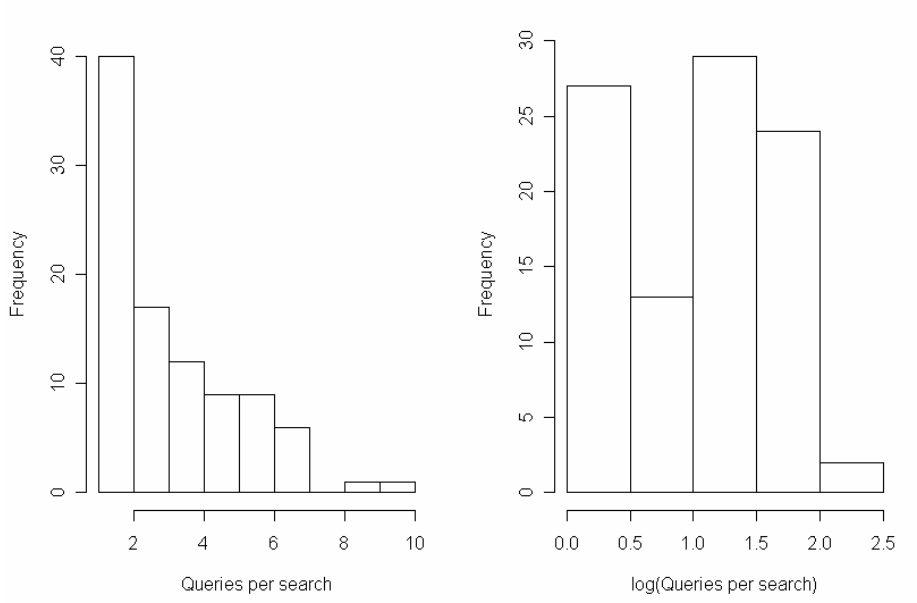

Figure 52. Histograms of a) queries per search and log(queries per search).

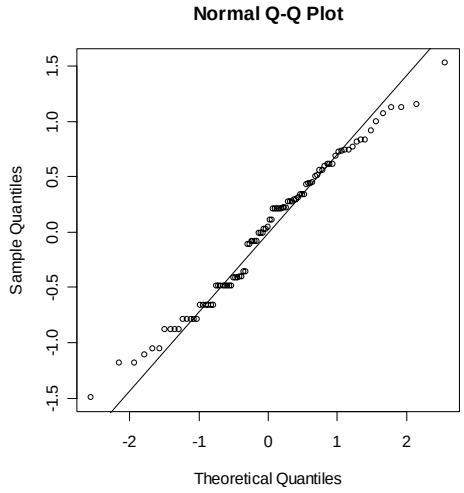

Figure 53. Normal Q-Q plot of residuals for the number of queries per search.

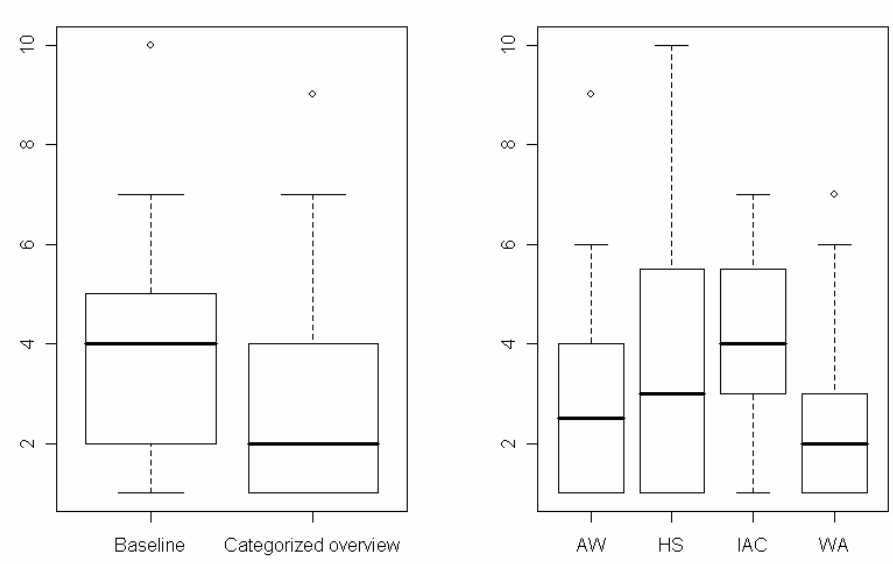

Figure 54. The number of queries per search, a) by System*, and b) by Topic* (N=95).

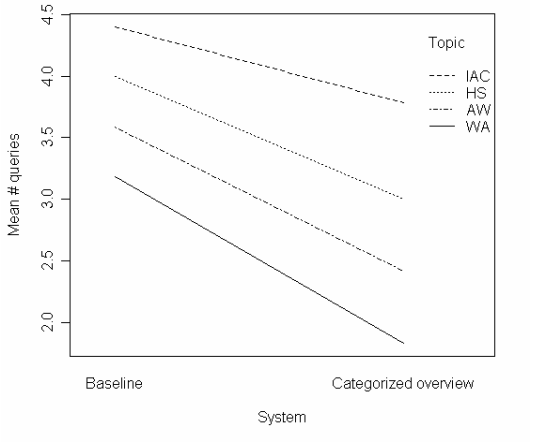

Figure 55. Interaction plot of mean number of queries per search for System and Topic factors.

5.10.1.6 Ease of exploration of search results

The results of a 2 (system) x 4 (topic) factorial analysis showed a marginally significant difference by system $F(1,88)=2.99$, $p<0.10$ and no significant difference by topic.

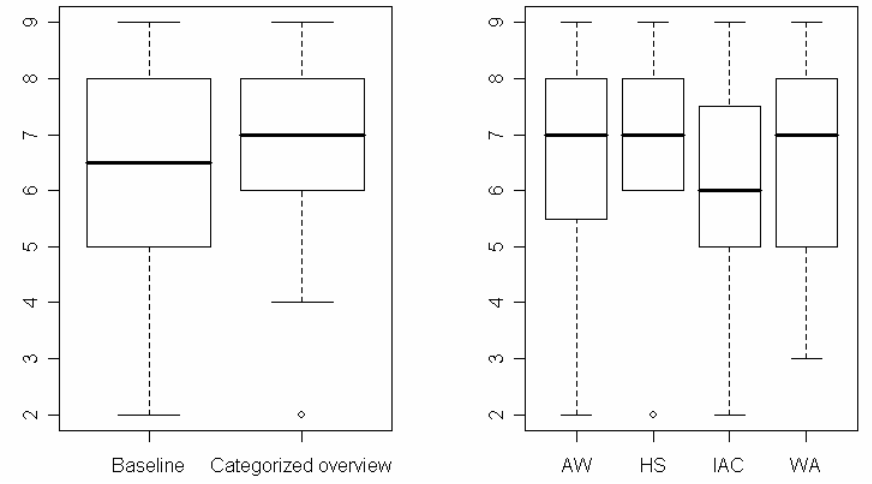

Figure 56. Ease/difficulty (1=difficult, 9=easy) of exploring search results, a) by System^{+}, and b) by Topic (N=96).

5.10.1.7 Got a good overview of the information available on the Web for topic

The results of a 2 (system) x 4 (topic) factorial analysis showed no significant difference by system and a marginally significant difference by topic $F(3,88)=2.57$, $p<0.10$.

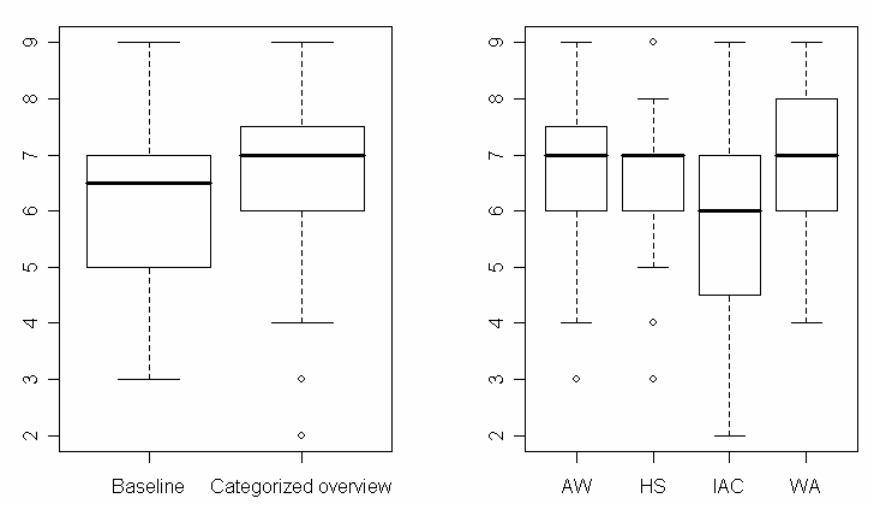

Figure 57. Agreement that they got a good overview of the topic, a) by System, and b) by Topic[+] (N=96).

5.10.1.8 Organization of search results

The results of a 2 (system) x 4 (topic) factorial analysis indicated a significant difference by system $F(1,88)=42.11$, $p< 0.001$ and no significant difference by topic. The Normal Q-Q plot shows that the residuals are normally distributed. The mean agreement for the categorized overview system was 7.4, and the mean agreement for the baseline system was 4.9. The corresponding medians were 7 and 5.

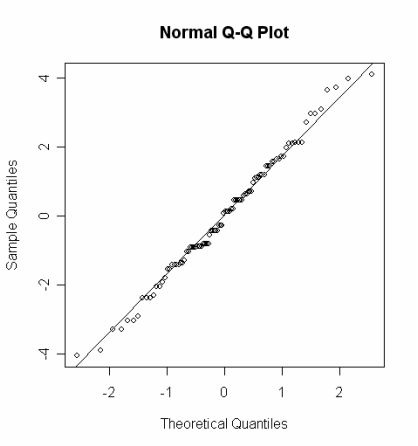

Figure 58. Normal Q-Q plot of residuals for the organization of search results measure.

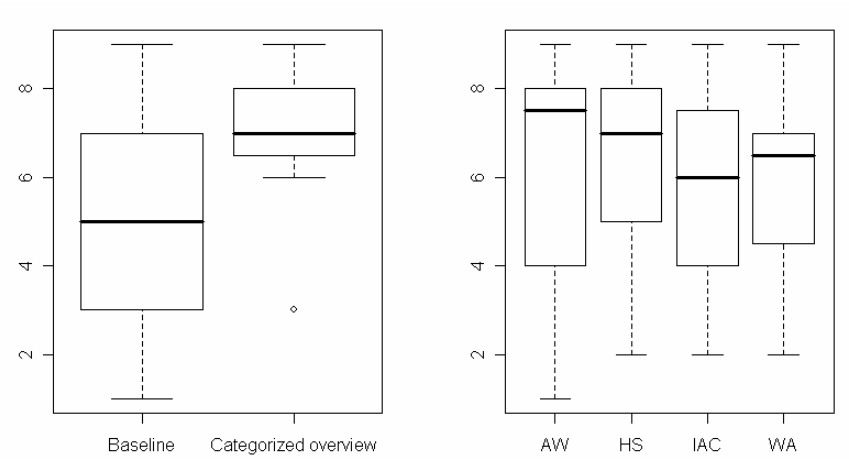

Figure 59. Agreement that system organized results well, a) by System*, and b) by Topic (N=96).

5.10.1.9 Agreement that system helped assess results and decide what to do next

The results of a 2 (system) x 4 (topic) factorial analysis indicated a significant difference by system $F(1,88)=13.63$, $p<0.001$ and no significant difference by topic. The Normal Q-Q plot shows that the residuals are slightly skewed, but this is unlikely to affect the overall significance of the system difference. The mean agreement for the categorized overview system was 6.5, and the mean agreement for the baseline

system was 5.3. The corresponding medians were 7 and 5.

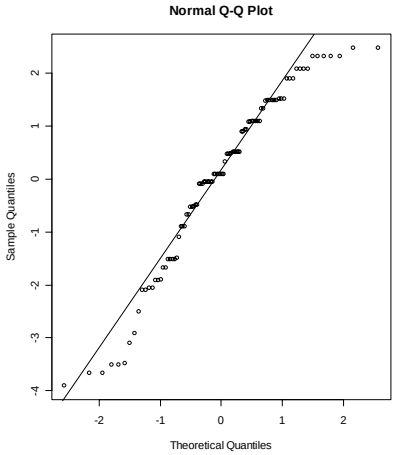

Figure 60. The normal Q-Q plot shows a slightly skewed distribution of residuals.

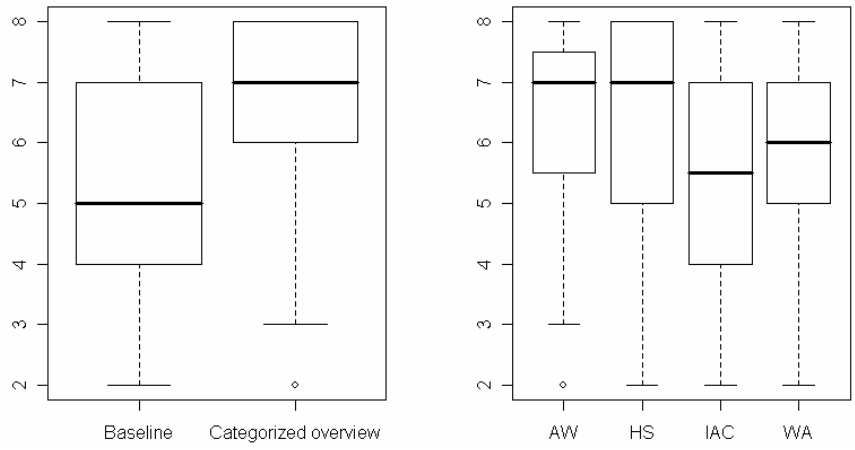

Figure 61. Agreement that interface helped assess results, a) by System*, and b) by Topic (N=96).

5.10.1.10 Adjectives to describe system

The 2 (system) x 4 (topic) factorial analysis for each of the eight system adjectives (semantic differentials) indicated no significant differences by topic for any measure,

but three measures showed significant differences by system: Terrible/wonderful $F(1,88)=7.05$, $p<0.01$; dull/stimulating $F(1,88)=13.73$, $p< 0.001$; and disorganized/organized $F(1,88)=45.7$, $p<0.001$. The analysis indicated marginally significant differences by system for the frustrating/satisfying measure $F(1,88)=3.03$, $p<0.10$.

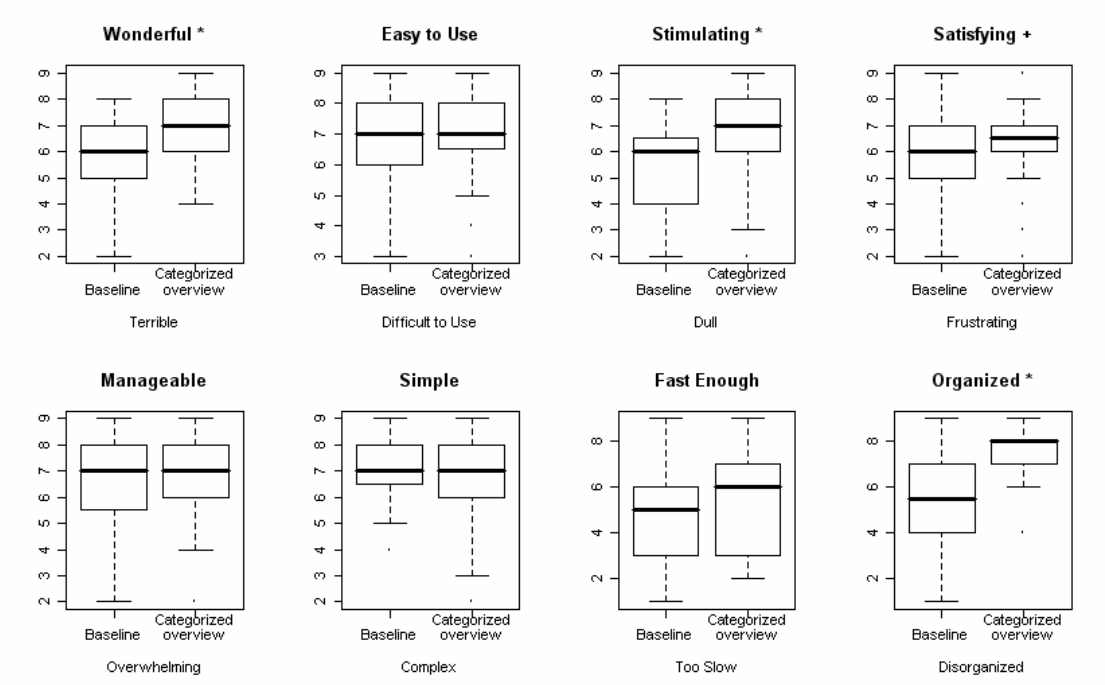

Figure 62. Adjectives by System.

5.10.1.11 Familiarity with topic

The results of a 2 (system) x 4 (topic) factorial analysis indicated no significant difference by system and a significant difference by topic $F(3,88)=5.71$, $p<0.01$. The normal Q-Q plot shows that the residuals are reasonably distributed. The mean change in familiarity was 3, that is, searchers rated their familiarity 3 points higher on a 9 point scale after the search. Tukey post-hoc tests on topic indicated significant

differences between the workplace allergies topic and the other three topics. The mean changes for AW, HS, IAC, and AW were 2.5, 2.1, 2.3, and 4.2, respectively. This is consistent with comments by several participants, who commented on the relative ease of finding information about the workplace allergies topic.

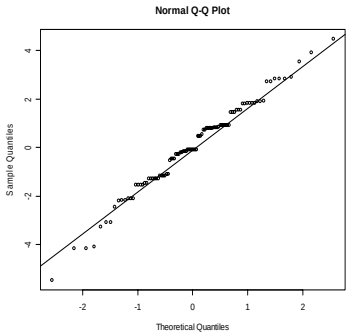

Figure 63. The normal Q-Q plot shows a normal distribution of residuals, indicating a good fit for the model.

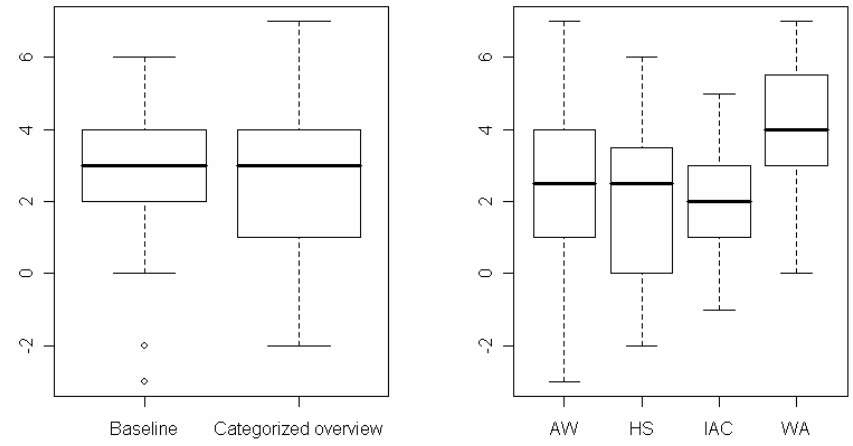

Figure 64. Change in familiarity after search, a) by System, and b) by Topic* (N=96).

5.10.1.12 Finding useful information

The results of a 2 (system) x 4 (topic) factorial analysis showed no significant difference due to either factor.

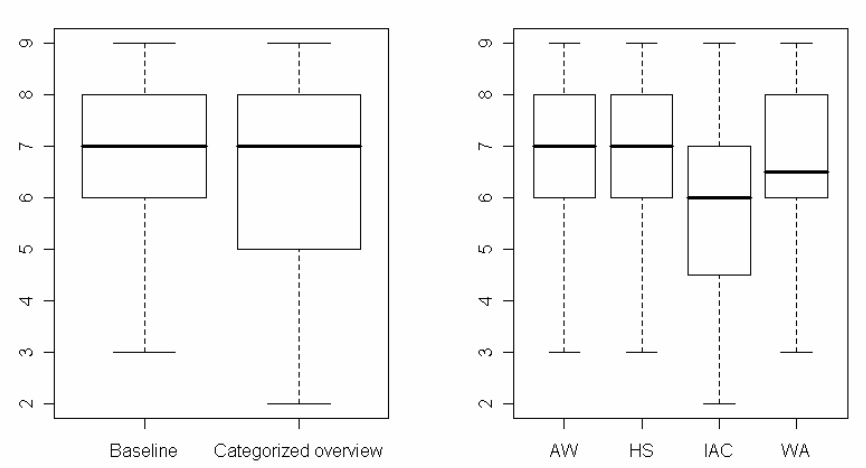

Figure 65. Useful information responses, a) by System, and b) by Topic (N=96).

5.10.1.13 Progress toward scenario goal

The results of a 2 (system) x 4 (topic) factorial analysis showed no significant difference due to either factor.

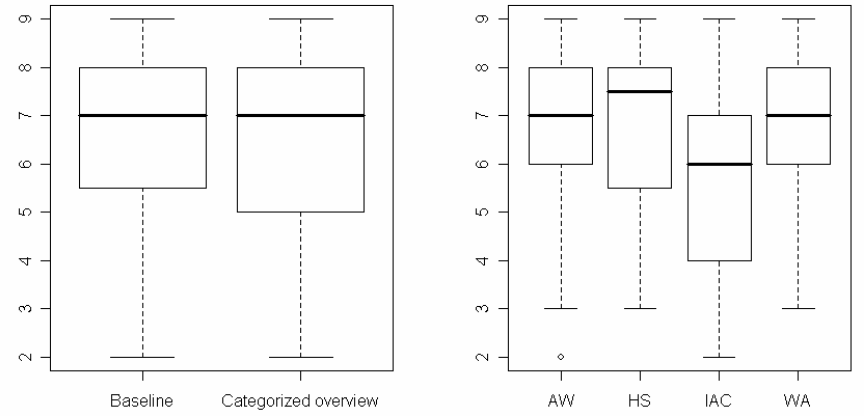

Figure 66. Progress toward scenario goal, a) by System, and b) by Topic (N=96).

5.10.1.14 Idea quality

Searchers generated a total of 679 ideas. Idea quality was generally low, at least in part because of the time limit, which several participants commented on. Although a nine-point scale was used (1 = poor, 9 = excellent), the highest rating assigned was 5. A Wilcoxon rank sum test did not detect a significant difference in Idea Quality by System. A Kruskal-Wallis test detected a marginally significant difference by Topic, $p<0.10$.

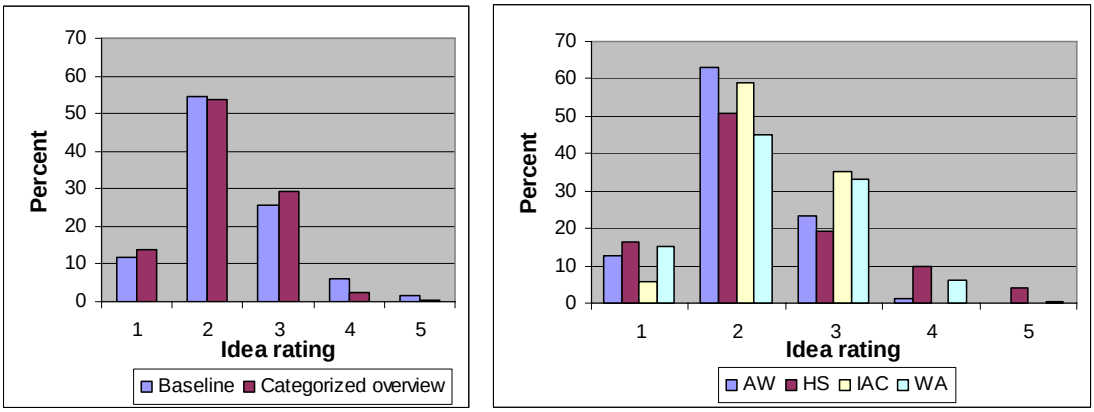

Figure 67. Distribution of idea quality ratings, a) by System, and b) by Topic[+] (N=679; idea rating 1 = poor, 9 = excellent).

The mean idea ratings were computed for the four non-journalism students and compared to the journalism students. They were almost equivalent. The mean idea rating was 2.27 for journalism students and 2.29 for non-journalism students.

5.10.1.15 Category use

Searchers used an average of 5.4 categories per search with the categorized overview. Most of the selected categories were used only once or twice, which suggests strong variation in individual assessments of the utility of each category, except for a few

highly used categories. Of the 259 instances of category use, 68 were for categories only selected once, and 44 were for categories selected twice. The most popular categories are shown in Table 25.

Table 25. Top 3 categories used for each topic.

Topic	Category	Distinct users	Absolute use
AW	/Health	7	7
	/Home	5	5
	/Business	4 (tie)	5
	/Society	4 (tie)	
HS	/Society	6	7
	/North America	5	6
	/News	3	5
IAC	/Arts	9	10
	/News	6	10
	/Reference	6	8
WA	/Health	9	12
	/Business	6	7
	/Society	6	6

5.10.1.16 Preferred system for task types

During the exit interview for the last 12 sessions, participants were asked which system they would rather use for four new tasks. They could respond with a system or say no preference. The sequence of the four tasks was randomized. The responses suggest that searchers would prefer a categorized overview for the comparison and

exploratory task. They would prefer the baseline system for the known item task, and were evenly divided for the simple informational task (Table 26).

Table 26. System preferences for known item, simple informational, comparison, and exploratory tasks.

	Preferred system		
Task (type)	**Baseline**	**No preference**	**Categorized overview**
a) Find the home page for the daily newspaper in Concord, NH, The Concord Monitor. (known item)	7	3	2
b) Find information on caring for a pet gerbil. (simple informational)	4	4	4
c) Start looking for information to help you select and buy a new digital camera. (comparison)	3	2	7
d) Learn about U.S. business investment in Africa. (exploratory)	3	0	9

5.10.1.17 Understanding of selected categories

Early in the sessions, I observed that participants had particular difficulty understanding the Open Directory's top-level Computers category during the training task. There seemed to be a discrepancy between what they expected to see under that category and the actual results. This prompted the addition of a question to the exit interview, starting with twelfth session, asking participants to describe what they would expect to find in three selected categories in the context of a search for "leonardo da vinci." They were shown the search results and asked to answer without using the mouse. After they answered, the pointer was placed over the Computers

category, which showed the subcategories in a small pop-up (Figure 68). They were then asked to answer the question again for the Computer category. Answers were informally evaluated to determine whether it was correct, partially correct, or incorrect/did not know and tabulated (Table 27). The Computers category was clearly problematic. A review of the pages in this category suggests that many pages are placed here because they are in computer-related web-sites. For example, Wikipedia pages are categorized in the /Computers/Open Source/Open Content/Encyclopedias/Wikipedia category (as well as others). This is an example of the ODP combining two kinds of relationships (*is-a* and *about*) and of minor problems that can be cause by limiting the depth of the hierarchy. These issues are discussed in section 0.

Figure 68. For the query "leonardo da vinci", placing the pointer over the top-level category Computer opened a small pop-up window with the five populated subcategories.

Table 27. Accuracy of participant understanding for selected categories (Kids and Teens, Reference, and Computers).

Category	Correct	Partial	Incorrect/ Did not know
Kids and Teens	13	5	0
Reference	11	5	2
Computers – without pop-ups	3	2	12
Computers – with pop-ups	9	6	3

5.10.2 Qualitative results

The relatively long (2 hours) study time enabled participants to consider their tactics and produced novel insights into the search processes of sophisticated searchers coping with challenging tasks. The qualitative results are organized into five sub-sections: Behavioral differences, cognitive and affective impacts, judgments of outcome, facet usage, and miscellany. These sub-sections include observations, quotes and comparisons between participants to highlight differences that the quantitative results do not capture.

5.10.2.1 Behavioral impacts

In confirmation of expectations, participants indicated that they used the overviews to filter, narrow, refine and explore their results. One participant was particularly effusive about the ease of narrowing her results, appreciating the immediacy of the interaction and commenting on two aspects of relevance that the overview enhanced for her.

> *I loved it. I was in love with that. I wish Google had that…That really helps if you can narrow it down by geography or if you're really looking for a credible source and you wish to go for government. The government sources are right there. It's just one click of the button and you have your government source…With 3 clicks you have 5 pieces of information. That's all you need to look through. (Participant 220)*

Participants were observed reading contents of the subcategory pop-up windows, which provided a form of query preview (Tanin, Plaisant, & Shneiderman, 2000), before clicking on that category or moving the pointer to a different category. Some commented during their searches and in a separate exit interview question on how they used the list of subcategories in the pop-up window to help decide whether to explore specific categories.

Two participants felt that they used fewer queries and five felt that their queries were more general when they used the categorized overview, but they had varying reactions to these changes. The average query length did differ slightly between systems for the aging workforce and workplace allergies topics, although not for the other two topics (Table 28). Most participants apparently considered this reduction in work a positive effect.

I knew that if I did a broader word it could be divided by the categories; I didn't necessarily have to be so specific. (222)

Rather than narrow down my search by adding additional search words I found myself narrowing my search by exploring categories and subcategories. (211)

Table 28. Mean (SD) query length by topic and system.

Topic	System	Mean (SD) Word Count
Aging workforce	Baseline	3.35 (1.15)
	Categorized overview	2.96 (1.18)
Human smuggling	Baseline	2.65 (0.78)
	Categorized overview	2.60 (0.77)
International art crime	Baseline	3.57 (1.04)
	Categorized overview	3.60 (1.05)
Workplace allergies	Baseline	2.96 (1.40)
	Categorized overview	2.77 (1.57)

Two participants expressed reservations about the change in their tactics:

I didn't use as many queries, which is part of the reason why I didn't get as good information. (216)

Maybe it made me a little bit lazy. But I felt like I had to do less because it would do more….it didn't take as much from me because they were gonna sort

through them and organize them for me.. I guess I changed by doing less. (213)

Two participants indicated that they adopted a tactic wherein they looked at the top of the search results first, then looked at the overview. One participant commented that it provided, "sort of a search within a search. That was very cool" (203). Six participants indicated that they used the overviews more on their second search and two felt they used it less. Of the two who used it less, one attributed this to encountering a very useful hub page (a page with many links). The other participant felt that the overview did not help and opted to use more queries instead.

Table 29. The 6 behavioral codes. Plus signs indicate that participants considered this a positive aspect. Negative signs indicate they considered it a negative aspect of their interaction. Neutral or mixed opinions are indicated by a 0. The count is the number of participants who made this type of comment.

Description	**+/-/0**	**Count**
Overview helped to filter or narrow list	+	7
Issued more general queries	0	5
Issued fewer queries	0	2
Ping-ponged – alternated between using the overview and the list	0	2
Explore – used the overview to explore the results	+	1
Used the overview to refine search	+	1

5.10.2.2 Cognitive and affective impacts

Thirty-four comments related to cognitive or affective impacts were gathered. The placement of pages within categories generated numerous comments. Eight

participants commented on pages that did not belong within a category at all, judging them as incorrectly categorized, whereas eight people indicated that they found unexpected pages in a category. This persisted even though the instructions emphasized that it was typically the web sites that were categorized, not the specific web pages. The prevalence of these concerns suggests that searchers may not remember the nature of the relationship entailed by category membership.

> *I wasn't exactly sure what I thought Shopping would be but I didn't think it was going to be here is where you can buy things like mold remover...whatever I thought it wasn't a web site where you can go shopping. (202)*

One participant particularly noted this problem in the geographic facet.

> *In the human smuggling one, because that one has a lot to do with geography but I noticed that in the geography sections you'd click on Europe but it wouldn't be about Europe, it'd be like.. like I said, companies based in Europe talking about human smuggling anywhere, you know? It wasn't always exactly what you'd think it would be... yeah, it could be a BBC story talking about something in Asia but it still categorized as Europe… It would be hard to fix that… I don't think it was a big problem, you just have to know that something could sort of have a double meaning like a geographic location. (208)*

Seven participants commented on the structure or organization of a facet as being confusing or non-intuitive.

> *Personal Finance under home I guess that makes sense but it's not something I would go to intuitively. I might have gone to...Business if I was looking at finance, but business is more like the corporate world and home would be your personal world, so after viewing it I can see the logic but it wouldn't have been there for me initially. (203)*

> *Why did they put News and Media under Computers? Publications under Shopping? (215)*

Five participants commented on confusing categories; two people felt that the topical categories were too general and one person felt that they were ambiguous. Interestingly, for all of the above codes, about half of the respondents indicated that the problems were minor or not a hindrance; perhaps they were able to quickly compensate for this variability in a manner similar to which searchers compensate for other breakdowns on the Web. A review of two of the session videos seems to indicate that those who reported experiencing such problems would quickly continue their search in the face of typical Web errors. One person indicated that he specifically did not go to one page because it did not fall in the category he expected.

> *I was shocked at the category that it was under, and I didn't pursue it but, and I can't remember the specific... seemed like it was very strange that it would be under that category… I'm not going to that site. [laughs] I just kept moving, which is probably not the best thing to do because it might be worth investigating but that's what I did. (203)*

One person was concerned that she might have missed useful pages in categories she did not explore.

> *Some categories I didn't even look at, and there might have been something useful there. 'Cause… I mean…I guess… I really don't know what the person was thinking when they categorized it. So… I mean… They might have been thinking about something that, like, never occurred to me but that is perfectly relevant so I feel like that might have, uh, hidden some information from me. (223)*

Four people commented that the categories helped generate ideas.

> *I was just looking for general information about the aging workforce but on the side it gave like me what the government was doing about it and I was like "oh, that's a good idea to look for," and like social issues about it…*

Two people commented that they used the overview when they were stuck.

When I was stuck on something I could start a new search pretty much because I could go in there and click on a new topic and then go see everything that they listed. (214)

… like allergies in the workplace. It was tougher to find varying things so I used the categories more when I was kind of stuck. (201)

Three people felt that the categories exposed them to different aspects of the topic.

I think it kind of opened up my mind a little bit to investigate a little bit deeper. Without the categories I just saw a list and I just had this mentality that I didn't want to go ahead and search through all of them but the categories made me think of different possibilities so I was more opted [sic] to search through a variety of different pages versus just looking for specific factors. (204)

It definitely changed the way I searched, probably for the better for something like this because it made me look at a wide range of categories. (210)

Another participant said that the overview provoked an illuminating question.

> *For the art crimes one, when I clicked on, I saw science and it was just, "What does that have to do with art crimes?" So that made me click on it and I found out that science can help solve art crimes. So that was something that I probably wouldn't have picked up on if that subcategory hadn't been there. (225)*

One person indicated that she used the overview to get an overall sense of how results were distributed within or across top-level categories.

> *It also changed how I originally took in the results rather than reading the titles and descriptions. I looked to see how they were divided up, what main categories there where, because I thought it would be faster way to see what I had in front of me especially for this particular task where I'm looking for different angles within a larger topic I wanted to see," well, there's a social issue and a health issue and a business issue," so that lends itself very well to that. (211)*

Table 30. The 34 cognitive and affective codes.

Description	+/-/0	Count
Incorrectly categorized – Subject considered the page to be in the wrong category	-	8
Unexpected pages in category – Subjects did not initially expect the pages they found within that category, although they did not consider it incorrect	-	8
Classification structure undesirable or confusing	-	7
Confusing categories	-	5
Generated ideas	+	4
Takes experience	-	3
Overwhelming	-	3
More complex	-	2
Indicated frustration	-	2
Pages appeared In multiple categories	-	2
Subject had topic in mind	0	2
Overview helped organize results better	+	2
Categories suggested idea	+	2
Used overview when stuck	+	2
Experience was less overwhelming	+	2
Felt more comfortable 2nd time	0	2
Categories too general	-	2
Exposed searcher to different aspects of topic	+	2
Concern that they might miss something	-	2
Ambiguous categories	-	1
Misleading	-	1
Provoked a question	+	1
Distraction	-	1
Many uncategorized results	-	1
Difficult to change search style	0	1

Confusing interface	-	1
Less confusing	+	1
Was more cautious using overview	-	1
Was more careful using overview	-	1
Human editors cataloged pages	-	1
Idea of where pages fit in categories	+	1
Overview made subject look at wide range of categories	+	1
Showed how pages were distributed across categories	+	1
Did less work	0	1

5.10.2.3 Judgments of outcomes

Participant comments included judgments on the outcomes of their searches when using the categorized overview. During their responses to the questions, ten participants indicated that the categorized overview was helpful. Three felt it was unhelpful and one commented that it was mixed overall. Eight participants commented that the problems they encountered were minor or did not hinder their search. They typically also described their rationale for this assessment. The first comment here illustrates one line of reasoning.

> *[It wasn't helpful for] Amazon.com. But, like you said, it didn't really frustrate me, it just, I just had to keep in my mind that it's human-generated. So it's not the web site's fault that it's there, its just somebody categorized Amazon.com as shopping, or say they considered it computers cause its an internet web site. It's not their fault that I clicked on it when that web site is just categorized as that, so it's okay. (220)*

> *Did it hinder searching at all? I would say generally no because I would go to the results here [indicates the list] first and then use this [indicates overview] as sort of a backup to reorder or filter again sort of thing. So it's a helpful tool. (203)*

One participant observed that a new query would generate more results.

> *With that whole legislation thing, I looked under US Government and I didn't find anything so I realized that, "Oh, maybe it is a little bit more specific," so then I just did a whole new search for it…. I got a lot more when I actually did a separate search than when I just clicked on US Government and expected more stuff to be there… (206)*

One participant attributed his assessment of poorer results to the fact that he issued fewer queries with the categorized overview and followed unhelpful links. Another participant felt she got sidetracked because of the overview.

> *I didn't use as many queries which is part of the reason why I didn't get as good information.. It led me down paths I didn't need to go down, because of the links on the side. (216)*

Table 31. The 9 judgment codes.

Description	+/-/0	Count
Problems were not a hindrance	+	4
Problems were a minor hindrance	+	4
Saw something that wouldn't have been seen otherwise	+	4
Search went faster	+	3
Search went slower	-	1
Got more results from a new query	-	1
Search was more efficient	+	1
Found poorer quality information	-	1
Got side-tracked	-	1

5.10.2.4 Facet and category usage

All participants commented on aspects of their use of the topic facet. Several commented on use of government and geographic facet use. Participants found that these facets helped narrow results and focus their search in ways that the topic facet did not.

> *That really helps if you can narrow it down by geography, or if you're really looking for a credible source and you wish to go for government. The government sources are right there. Its just one click of the button and you have your government source. It's easier to cite it. You don't go looking for – like with Google – you'd go through what the US government has to say about*

workplace allergies. Here, it's in front of you, you know, Dept of Health and Labor. (220)

I like the government sites at the bottom, because I tended to look at government sites first. (224)

When I was doing the smuggling thing I focused more on the geography, because – human smuggling clearly is a social issue, an economic issue, well that's obvious, but then it's like, where is it in the world, so I looked under geography (206)

I was getting a lot of stuff about the US, so I clicked on Europe and it gave me stuff about the UK. (207)

As with the topic categories, participant comments indicated minor problems with the categorization, or their interpretation of the categorization rules. In this quote, the participant was evidently confused about what pages would be placed in the US government categories.

I think even though certain things are categorized under certain topics...things under US government might just mention US government. It might not be an actual government page. (207)

Table 32. Mentions of geographic or government category use.

Description	+/-/0	Count
Used geographic facet	0	7
Used government facet	0	4

5.10.2.5 Miscellaneous

Three participants commented that the topic had an effect on how much they used the overview.

> *I definitely used it more the second one because... It was also a tougher, tougher thing to find, like allergies in the workplace. It was tougher to find varying things so I used the categories more when I was kind of stuck. (201)*

> *The second topic was more conducive to that kind of thing because the workplace allergies sorted so well.. .there's health issues, there's business issues there's government issues. It worked really well with those categories, a little better than the human smuggling one because that [topic] doesn't fit well into like health or computers as the other one. It's a little bit more narrow probably that's why… [Workplace allergies] is more broad so it fits into the categories a little bit more, except that it doesn't fit into all of them.*

5.11 Discussion

5.11.1 Topic and task efficacy

The four topics used for the searches were intended to be matched pairs (two broad and two narrow) for the Topic Type factor. It became clear during the study that they were not well matched. In hindsight, the procedure used to select the topics was not sufficient to ensure a match. The evaluation of the candidate topics during the pilot test was not rigorous enough, in part because the broad/narrow concept was not operationally defined in a way that permitted an objective assessment of topic breadth. The lack of a clear definition also hindered the construction of the topics because there were no guiding criteria, and the resulting pairs of tasks were differentiated more in terms of difficulty than breadth. The two topics drawn from the TREC Robust track (international art crime and human smuggling) were generally perceived as more difficult than the other two (aging workforce and workplace allergies), although not universally. Participants varied in how they interpreted the topics, and some participants had knowledge that caused them to consider a topic easy. The exploratory nature of the task encouraged participants to apply their own experience and knowledge, and this amplified the variations. These factors contributed to the unmatched nature of the topics, and necessitated changes in the quantitative analysis. What was to have been analyzed as a 2 level, within-subjects Topic Type factor had to be analyzed as a 4-level, between groups Topic factor instead, and the Topic Type hypotheses (all the “B” hypotheses) could not be tested. The variability may also have reduced measured differences by System in dependant variables.

The combination of positive participant response to the interface with no differences in outcomes could indicate that the specific task was less dependent on gaining an overview than originally anticipated. Most participants appreciated and used the overview, but when that wasn't available, scanning the result lists and reformulating queries were reasonably effective tactics for generating article ideas to satisfy the assigned task. The quality ratings for all ideas were generally low and there were not noticeable differences between systems.

5.11.2 Differences in search behavior

The quantitative and qualitative data indicate that the overviews did change searcher behavior in several ways. The log data showed that participants explored significantly more deeply within the result list. This supports hypothesis H1a and is consistent with, but more modest than, previous studies (Käki, 2005). Overall, they did not collect pages significantly more deeply with the categorized overview. The mean depths of collected pages were the same for both conditions (as were the medians). Thus hypothesis H2a was not supported. Overall, for the given task and topics, the categorized overview did not have a significant effect. For the aging workforce topic, however, which showed the largest difference in depth of viewed links between systems, participants using the categorized overview did collect links from deeper in the result list. This could suggest that when they did explore deeper in the results, they did find useful pages more deeply.

With the categorized overview, participants did collect slightly more pages that were categorized (i.e., they collected fewer uncategorized pages), supporting hypothesis H3a. Thus the categorized overview biased participants toward pages that were found in at least one category. Whether this bias is positive or negative depends on the context of search, the number of uncategorized pages, the value of the uncategorized pages, and the negative impact of not viewing the uncategorized pages. A few participants were concerned that they might overlook something by using the categories. This implies that to minimize undesirable impacts, searchers should understand when they are limiting their search to categorized results, whether it is important for them to view uncategorized results, and how to do so. This suggests a need for better training and/or clearer indications to searchers that their results are being filtered. Participants did not always appear to comprehend this distinction. This finding has been incorporated into the design principles.

The participants issued fewer queries with the categorized overview, supporting hypothesis H4a. The categorized overview appeared to provide cues, similar to the notion of "information scent" (Pirolli & Card, 1995; Pirolli & Card, 1999), that induced participants to click on categories instead of refining their query. This is supported by the participant comments. Participants commented on submitting more general queries and then using the categories to explore or narrow their results. They did issue somewhat shorter queries for two topics (aging workforce and workplace allergies). It is possible that there was a confounding factor: Issuing a new query took substantially longer than simply exploring a category, which could have induced

participants to avoid query refinement. Participants were alerted in advance that there could be delays, and I asked them to "be patient and search as you normally would." None of the participants appeared to indicate (verbally or non-verbally) impatience at the delays incurred by the search or a reluctance to refine their query due to the time. Thus, the query time is unlikely to have been a confounding factor.

Participants clearly appreciated the categorized overview. Search engine operators might also benefit if they can serve the same number of searchers with fewer queries. Processing fewer transactions with larger result sets could be a desirable engineering trade-off. This would require that the client be able to receive the entire result set at once and then allow the searcher to interact with it. The SERVICE prototype currently implements the filtering logic on the web server, but with the use of client-side technology currently available (JavaScript, Ajax, etc.) it is feasible to implement the entire UI on the browser. Because the amount of data being transmitted in a set of search results is modest (100-150k for 100 results from Google), this could be accomplished in a single HTTP request per query without substantial delay. Even a small reduction in queries per search could be beneficial for high-volume search services. This could also improve interactive performance from the searcher perspective.

During the interview, participants commented that they changed their tactics to utilize the overview. One participant commented that he skipped a page because it was in an

unexpected category. Section 5.11.5, Differences in searcher thinking, discusses these changes.

5.11.3 Cognitive impact of categorized overviews

The overviews provided an alternative perspective on the search results that participants found helpful. In some cases the benefit derived from a reduction in work, for example by replacing a query refinement step with a single click. The query log data corroborate participant assessments of this effect. The subcategory pop-up windows provided contextual information and formed a query preview that helped searchers decide whether to explore a category. In other cases the participants concluded that the overviews suggested an idea or question or exposed them to a concept they would not have otherwise seen. They were speculating, of course, but they considered it a positive contribution to their search experience.

One of the most common complaints concerned the assignment of pages to categories. When page categorizations did not match participant expectations, they experienced frustration, confusion and doubt. The ODP classification generally captures what web sites are *about,* i.e. the topic. Fourteen of its 16 top-level categories are primarily topical, which was the rationale for using them to construct the Topic facet. However, three factors clearly reduced the categorization accuracy from the participants' perspective: encoding different relationships within the same facet, ambiguous categories, and the hierarchical structure of categories.

Pages from the British Broadcasting Corporation (BBC), for example, were categorized under /Category/Arts/Television, which is closer to encoding an *is-a* relationship than an *about* relationship. (It could more accurately be construed as the hosting organization, producer, host or even author relationship.) Thus, when a BBC web page about a human smuggling story was found under Television, it was puzzling to many participants. It did not match their expectations.

Participants also commented on the generality or ambiguity of the categories, particularly the topical categories. This could be attributed, at least in part, to the limited depth of the hierarchy that was used in the categorized overview. The depth of each facet was limited to two levels due to performance issues with the specific implementation. The ODP-assigned categories were frequently four or more levels deep. For example, the BBC web page mentioned earlier was assigned to categories in the topical and regional facets (Table 33). Truncating the categories to two levels removed useful contextual information. The end result was a more general category. This could also have contributed to the perception that pages were incorrectly categorized**.**

An alternative approach could have preserved the contextual information in the overview by promoting the lower level categories, thereby flattening the hierarchy. This approach could work well with larger displays, but is problematic in the limited space available for the overview. In early tests, this multiplied the number of second

level categories unacceptably; there were too many to fit on one screen in the overview.

Table 33. A BBC web page on human smuggling was categorized into eight categories in two facets, most of which were at least four levels deep. Truncating the categories to two levels removed useful contextual information.

	Original categories	**Two-level category**
Topical Facet	• /Arts/Television/Networks/Cable/BBC • /Arts/Television/Networks/Europe • /Arts/Art History/Artists/D/Da Vinci, Leonardo • /Science/Educational Resources	• /Arts/Television • /Arts/Art History • /Science/Educational Resources
Regional Facet	• /Europe/United Kingdom/Government/Culture, Media and Sport/Broadcasting • /Europe/United Kingdom/News and Media • /Europe/United Kingdom/Science and Environment/News and Media • /Europe/United Kingdom/Guides and Directories/Search Engines	• /Europe/United Kingdom

The category structure was sometimes problematic. Some participants did not initially expect to find the Television category under Arts, for example, and found this troubling. The example in Table 33 illustrates how the ODP uses the descriptor "News and Media" in two separate entries. A more rigorous approach to facet analysis (Soergel, 1974) could yield more nearly orthogonal facets and identify additional facets. This might yield substantial improvements in the perceived accuracy of the category assignments. A lightweight tool could allow experienced

indexers or “power searchers” with expertise in specific domains to customize the category structure, quickly edit hierarchies, splitting, merging, promoting, or hiding categories.

The occasional problems with category structure partly reflect the tension between ideals and practice in a classification that is used as a boundary object between communities of practice (Bowker & Starr, 1999). The ODP categories are used by 50,000 editors to catalog web sites. These editors have varying backgrounds, motivations and interpretations of category meaning. Moreover, the application of the ODP to organizing search results changes the context within which the categories are interpreted. The original context, within which the editors operated when cataloging, was a browseable directory of web sites. The primary concept was the classification structure. When used to organize search results, however, searchers have a different conceptualization of the context, in which the search goal and immediate task are primary, and the classification structure is secondary.

Fortunately, participants indicated that these problems were minor. Their comments on difficulties indicated that their problems were with details of the categorization and that they managed these by relying on the stability of the overall categorization scheme. They commented on being more familiar and comfortable with the categories and having a more accurate understanding of the categorization scheme by the second categorized overview task. This could be a benefit when compared to automatically

clustered or dynamically generated categories, which will differ for each set of search results.

The satisfaction data support this interpretation of the experimental results. Although the positive results for hypothesis H5a (searchers will find it easier to explore search results) were only marginally significant, participants agreed that the categorized overview organized the results well and helped them assess their results and decide what to do next, partially supporting hypothesis H8a. Hypothesis H6a (searchers will agree more strongly that the system provided a good overview of the information available on the Web) was not supported. It is likely that the topic breadth and difficulty contributed to the variability of this measure. Participants also found the categorized overview more generally appealing ("wonderful") and stimulating, supporting hypotheses H10a and H11a. There was no significant difference in ease of use ratings, so hypothesis H9a was not supported. The satisfaction ratings, which favored the categorized overview, were marginally significant. This suggests some support for hypothesis H12a. Participant satisfaction depended on many factors, including the information available as well as the search interface, so it is possible that participants' assessment of their modest progress and generally poor quality results (which many commented on) reduced their overall satisfaction.

There was no significant difference in the overwhelming/manageable or complex/simple measures, although some participants commented on the overview being more complex or overwhelming. This lack of support for hypothesis H13a

(searchers will rate the categorized overview more complex) is good news, because it suggests that the categorized overviews were not, in general, perceived as substantially more complex. This does not mean that complexity effects should be ignored. Indeed, one participant specifically asked if he could hide the overview because it was distracting. But for this task there were clear benefits to most participants. During the exit interview, the 12 participants who were asked which interface they would rather use for a range of fact finding to exploratory tasks, they indicated a preference for the baseline interface for fact finding and the categorized overview for the exploratory tasks. These results reinforce the value of providing searchers additional control over their search (Greene, Marchionini, Plaisant, & Shneiderman, 2000; Koenemann & Belkin, 1996; Shneiderman, Byrd, & Croft, 1998), including whether to include display or hide the categorized overview.

5.11.4 Differences by breadth of topic

Although the pilot testing suggested the topics were matched in terms of breadth, it became apparent during the experimental sessions that participants had highly varied interpretations of the breadth of the topic. Their knowledge of the topic, attitudes toward it, their response to specific web pages, all contributed to their perception of the breadth of the topic. More importantly, their perception of the topic difficulty varied widely, for similar reasons. These two factors clearly affected their assessment of their progress toward the scenario goal, along with the limited search time. Participants commented on these issues during the sessions and afterwards. In hindsight, the study would have benefited from a more rigorous and defined development of topic breadth and difficulty (Bell & Ruthven, 2004). This could have

contributed to more reliably measurable effects. Nevertheless, the varied individual interpretations of the topic provided useful data for the qualitative analysis contributing to the other research questions.

An additional post hoc analysis could be performed to stratify the cases by the perceived breadth or the topic difficulty. A two-factor analysis with system and perceived topic breadth as factors might show differences by topic breadth that the above analysis did not. Because of the small size of this data set, the likelihood of needing to use a less powerful non-parametric analysis technique, and the inherent variability of the data, it is unlikely that any differences would be significant. However, future studies could take this into consideration.

5.11.5 Differences in searcher thinking about search tactics

Participants commented on many interesting effects that the categorized overviews had on their thinking during searches. They confirmed expectations that they would change their tactics to utilize the overview. Some used it before looking at the result list, whereas others used it in an ancillary or backup role, for example, when they felt "stuck." Participants used the categorized overview to understand the distribution of the pages across categories. They also used the categories to confirm interest in a particular page seen in the result list. They used the query preview capability provided by the subcategory pop-up window to predict what would be in the category and help decide whether to view the results within that category. In these cases, they appreciated the categorized overviews, and several commented on feeling more efficient.

Several participants spoke of the difficulty of changing established search tactics. In the time allotted, some searchers changed their tactics rapidly, whereas others only started to change. Two participants did not appear to change their tactics at all during the session. Rather than use the overview to help guide their idea generation, they thought of specific ideas, and then searched for them. Sometimes they would simply issue queries specific to that idea, ignoring the overview. At other times, they would use the overview to filter the results to pages that were related to the desired topic.

Participants often thought that they used the overview more during the second categorized overview search. They actually clicked on categories slightly more during their first categorized overview search (132 vs. 127), but they appeared to be exploring the interface and probing categories. Some participants specifically said that's what they were doing, and comments like "let's see what this is" were frequent. By the second categorized overview search, it appeared that most participants were taking advantage of the overview, although many were still exploring the categories and revising their search tactics. The robustness with which participants responded to the problems discussed in previous sections also suggests that they quickly began to adapt their search tactics to take advantage of the categorized overview while compensating for its weaknesses.

They also commented on feeling cautious in using the categories, or of being more careful than usual, particularly after seeing a web page categorized in an unexpected

category. It is possible that these feelings would subside with greater use of the categorized overviews and increased familiarity with the categories.

5.11.6 Effect on quality of search outcome

None of the outcome-oriented hypotheses (H13a-H15a) were supported by the quantitative results. As noted earlier, many individual factors can affect search outcomes, particularly in exploratory searches. Participants perceived the breadth and difficulty of topics very differently. Their comments suggest that the challenging nature of the experimental task, the tight time limit and the topic difficulty all contributed to the difficulty in making progress toward their goal and the generally low quality of ideas.

The qualitative data suggest that ideas were provoked by the categorized overviews, and some participants felt that they would not have generated specific ideas without the overviews. The data also suggest one possible negative outcome on the quality of ideas. One participant indicated concern that idea quality was negatively affected, indirectly, by changes in his search tactics due to the overview. He felt that he was not getting as many good results because he relied on the categories instead of analyzing the results to identify new concepts and terms to refine his query. Although other participants did not directly comment on this, observations of their actions and comments while searching lends credence to this concern. When presented with a feature that reduced cognitive effort, some participants used it even if it produced non-optimal results. They found beneficial trade-offs in this satisficing behavior (Marchionini, 1995; Simon, 1979) due to the context of the search. In this case, the

low negative impact of poorer results, the non-trivial effort needed to generate high quality story ideas, and the limited search time, probably induced participants to accept the poorer outcome. In a bona fide context, they would probably be more motivated and have more time, which might produce better results.

5.12 Limitations

5.12.1 Subject population

This study was limited by the fact that the participants (N=24) were all students at the University of Maryland. Twenty were journalism students, so the scenario and task was appropriate for them (as most of them confirmed), but they might not be representative of the needs of other exploratory searchers. The journalism scenario was not relevant to the four non-journalism students, although the specific task appeared to be similar to tasks they had performed. The participants were all experienced searchers, and many appeared to have established sophisticated search tactics. They are unlikely to be representative of searchers with less experience.

5.12.2 Category structure and membership

The study was limited by several factors related to the categories: the specific facets used, the proportion of uncategorized results, and the structure of the categories within facets. Only three facets were used: topic, geography, and US government. They were selected because they could be practically extracted from existing, available data. The first two were selected because they categorized a broad set of web sites, providing a wide, but shallow, set of categories, and the third was selected because it provided a comprehensive categorization for a narrow domain that was

conceivably useful for the scenario. Other facets could have been chosen, representing different types of relationships. For example, the Last Time Visited classifier might be useful for searchers attempting to re-find a page if they had knowledge of when they had viewed the desired page. This would have yielded different quantitative and qualitative results.

The modest proportion of pages that were categorized was a limitation of the study. Typically 40-80% of the search results for a query were categorized, which left many uncategorized pages. This had negative cognitive and affective impacts, discussed in section 5.11.3, like complicating the search process by the need to consider uncategorized pages in decision making. For domains in which search results can be more comprehensively categorized, these negative effects might not be observed. The modest changes in behavior (e.g. depth of viewed pages) might be more pronounced. Overall, the categories used in the study were intended to provide a pragmatic assessment based on the amount and kind of information currently available for categorizing search results from general web search engines. They did not utilize traditional text classification techniques. Incorporating these techniques might improve categorization rates.

The structure of the topic facet was a limitation of the study. The ODP is not a well-structured, formal classification. It represents different types of relationships within the hierarchy, the relationships can be ambiguous or loosely defined, and their interpretation can differ between the ODP editors and the searcher. This had minor

negative cognitive and affective impacts, puzzling and frustrating searchers. In particular, searchers sometimes perceived pages as being incorrectly categorized or were surprised by their placement within categories. This was exacerbated by the limited depth of the hierarchy used (three levels). For domains in which the classifications are formally defined, these impacts might be less prevalent.

5.12.3 Scenario and task

This study is limited because only one scenario and task type was evaluated. Other exploratory search tasks may benefit more or less from the categorized overview. In fact, the task was an important limitation on the quantitative results of the study. The overview may not have been as important to task performance as originally expected. The task could be successfully completed with tactics that did not utilize the overview. Individual differences also affected the quality of the generated ideas. These factors probably reduced the quantitative impact of the two different interfaces. The task had the desired effect of encouraging participants to re-evaluate and revise their existing search tactics, and it encouraged participants to analyze and integrate search results with their own knowledge, which is an important component of exploratory search tasks.

This study was limited because the characteristics of the desired outcome were not fully described to participants. Although they were asked to generate a diverse set of article ideas, they were not told the specific newsworthiness criteria by which the ideas would be assessed. Providing this information might have helped participants

generate higher quality ideas overall, which might have in turn enabled a distinction to be seen between the two interface conditions.

5.12.4 Time constraints

The study was also limited by several time constraints. Although the training time was sufficient for subjects to learn the mechanics of using the categorized overview and the practice the task, it took time for subjects to reflect upon and revise their search tactics. They were often still in the process of refining their tactics at the end of the second categorized overview task. The time allocated to each task (12 minutes) was also short, which limited their ability to conduct more thorough searches and generate high quality ideas. A longitudinal or multi-day study could overcome this shortcoming by giving searchers time to adapt before conducting the assessed tasks.

5.12.5 Interface design

The study was limited because the experimental design implemented only one design idea for presenting the overview, a textual list supported sequential selection of categories within a facet and simultaneous selections between facets. Table 20 lists nine dimensions of the design space which could be explored. Alternate approaches would have different trade-offs, possibly leading to different results. In particular, graphical elements could have been incorporated into the overview, a possibility that the Future Work chapter addresses.

5.12.6 Topic breadth

The topics were not matched, as discussed in section 5.11.1. This prevented quantitative investigation of the effect of topic breadth, and complicated the

qualitative analysis. It also limited the statistical power of the quantitative analysis of the Topic variable, which had to be analyzed as a 4-level between-groups factor instead of a 2-level within-subject factor.

5.12.7 Quantitative analysis

The statistical power of the quantitative analysis was limited by several factors. The non-matched nature of the broad and narrow topics has been noted. Individual differences appeared to be a factor in the variability of observed behaviors and the quality of the generated ideas. The overall statistical power was limited by the modest number of subjects, and may have been affected by the inclusion of the four non-journalism students.

5.12.8 Qualitative analysis

The qualitative analysis was limited in several important ways. The research was conducted in a laboratory setting rather than the participants' own workplaces, and the task was not of their own choosing. They were removed from their typical environment and asked to perform a task with artificial constraints. The detailed scenario was designed to provide a rich context for the task and to encourage them to draw on their own experience, but the experience was certainly only a facsimile of what it would be in practice. Participants did, however, show an awareness of these differences. They acknowledged the differences in during the exit interview, and they commented on the essential elements of the task that were common between the research setting and their workplace.

The qualitative analysis is limited because of the primary reliance on peer review of the interpretations and conclusions. A single researcher analyzed and interpreted the raw data. Conducting member checks was not considered feasible because of the cost and time required to recall participants after the intervening Christmas and New Year holidays. Using a second researcher to code the exit interview questions might have identified additional behaviors, tactics, and thoughts, or provided alternative interpretations. The study does make modest use of triangulation with the quantitative data, and it provides direct quotes to support interpretations. The phenomena being examined in this study was constrained by the laboratory environment and the task, which removed many external factors that could lead to variations in interpretations. The interpretations were closely tied to the raw data, often using the same language that participants used.

5.13 Summary

As a whole, this study and the two early studies provide qualitative support for the use of categorized overviews of search results based on meaningful and stable categories, and identify some possible limitations. Across two different domains, the three studies showed that searchers explored more deeply in their results, and were more satisfied with the experience, although they do not show objective differences in search outcomes. Searchers agreed that the categorized overviews helped them organize, explore and assess their results, and were not appreciably more complex than typical Google-like interfaces.

The early studies refined the design principles for exploratory search, and this study corroborated the principles by evaluating the SERVICE prototype, which was designed according to many of the principles described in section 4.2:

- Provide overviews of large sets of results
- Organize overviews around meaningful categories
- Tightly couple category labels to result list
- Arrange text for scanning/skimming
- Support multiple kinds of categories
- Make category structure visible
- Use separate facets for each type of category

One important implication of this study for search interface designers is that the hierarchy used in a categorized overview should be carefully analyzed and may need to be modified in two ways. First, different relationships encoded in the hierarchy (e.g. *is-a* vs. *part-of*) should be separated into separate top-level facets. Second, and more generally, parent-child (or broader-narrower) relationships that are clear when encountered while browsing a thesaurus or directory of web pages, will not always be clear when used in the context of a categorized overview of search results. The structure of the hierarchy will need to be changed in these cases. This suggested a new principle ("Use separate facets for each type of category") and refinement to the initial principle, "Visualize and clarify category structure." Practitioners should analyze at least the top two levels of a hierarchy, considering whether they need to be adjusted to provide the clearest overview.

The study suggested an additional design principle: “Ensure that full category labels are available.” It also suggested a refinement to the principle, “Tightly couple category labels to result list”: Provide clear indications to searchers when and how their results are being filtered.

The study suggested how categorized overviews affect cognitive processes, and illustrated ways that participants began to adapt their exploratory tactics to use the categorized overviews. The categorized overviews encouraged searchers to create fewer, and possibly broader, queries for the search tasks, which changed the tactics searchers used to (re-)formulate queries. Several different tactics for using the categorized overview emerged, including using it to organize the exploration of the results, alternating between the overview and the list, and using the overview simply as a backup or secondary tool. The study highlighted the difficulty that some participants had in adapting their existing search tactics to take advantage of the new capabilities. The study provided several examples of searchers apparently satisficing by using the categorized overview. These results helped to refine the analysis.

Evaluating exploratory search task outcomes is challenging, and these studies did not detect quantitative differences in search outcomes. The results do provide qualitative indications that categorized overviews suggest ideas and questions to searchers that would not surface with the baseline system. They also raise cautionary questions about possible negative impacts on the quality of search results.

One important economic implication of the study for search engine developers is that they might serve more searchers with fewer transactions by providing larger result sets with categorized overviews. This assumes that the category information is available at query time.

Chapter 6: Contributions

6.1 Benefits of categorized overviews

The qualitative analysis of study 3 identified changes in how searchers think about and interact with search results when a categorized overview is available. It identified seven tactics that searchers adopted in response to the categorized overviews. Study participants agreed significantly more that the search system helped them assess their search results and determine the next steps in their search process with the categorized overviews than without. Study participants found the categorized overview interface significantly more organized than the baseline system.

Studies 1 and 3 confirmed previous findings that searchers view pages deeper in their search results when overviews are available (Käki, 2005). Study 1 extended these findings by providing quantitative (albeit not statistically significant) indications that the categorized overviews also helped searchers *find* relevant and useful pages deeper in the results for an exploratory search task ("Find 3 web pages providing different aspects of or perspectives on this topic".

Studies 1 and 3 confirmed previous findings that searchers were more satisfied with their experience when using the categorized overview than without it.

6.2 Limitations of categorized overviews

Study 3 found no differences in the outcomes of an exploratory search task (generate newspaper article ideas). Analysis of the results suggested that several factors

contributed to this. First, task performance for that task may not be dependent on an overview, even though searchers appreciated it. Second, a large number of uncategorized results may have limited the effectiveness of the overview. Third, flaws in the hierarchical structure of the categories may have limited the effectiveness of the overview. This suggested that when categories are incorporated from existing knowledge structures, such as the Open Directory, the hierarchical structure should be carefully analyzed and may need to be modified for use in the categorized overview. This yielded several design principles and suggested refinements for future studies.

Study 2 indicated that automated clustering techniques supported an exploratory search task that involved generating ideas for newspaper articles. Participant comments indicate that the words and phrases in the cluster labels suggested article ideas.

6.3 Analysis of search tactics with categorized overviews

This dissertation presents an analysis of search with categorized overviews. It proposes a model of the exploratory search process (Figure 25), identifies four lightweight actions available to searchers when evaluating search results with categorized overviews (Table 11), and describes six beneficial tactics that searchers can adopt when categorized overviews are available (Table 12). This provides theoretical support for a set of principles for the design of exploratory search interfaces. The analysis helped guide the design of the SERVICE categorizing search system.

The analysis should stimulate research into the delicate interplay between the presentation of categorized overviews and the search results, the forms of interaction available to the searcher, the learned tactics that searchers employ, and the fundamental human and machine constraints that affect search. This analysis, narrowly focused on one step (examining search results) and one form of interface (categorized overview) should be seen as one step in understanding how exploratory searchers search.

6.4 Design principles for categorized overviews of search results

This dissertation proposes a set of design principles for exploratory search interfaces, supported and refined by the empirical studies:

- Provide overviews of large sets of results
- Organize overviews around meaningful categories
- Clarify and visualize category structure
- Tightly couple category labels to result list
- Ensure that the full category information is available
- Support multiple types of categories and visual presentations
- Use separate facets for each type of category
- Arrange text for scanning/skimming
- Visually encode quantitative attributes on a stable visual structure

These principles will be useful for digital library and web search designers, information architects, and web developers because they provide guidance for the appropriate integration of visual overviews with search result lists, and particularly

for the textual surrogates embedded in result lists. These principles embed a strong call for the surfacing of structure – which is often used internally by search engines, but less often exposed at the user interface – without abandoning the tried and true value of text.

6.5 Fast feature classifiers

This research contributes a framework in three dimensions (fast-feature/full-feature, rich/lean, online/offline) to analyze techniques for categorizing web search results. It describes nine Fast-feature, online classifiers that integrate information available in web search results with external data sources to categorize search results into meaningful and stable categories. The implementation and analysis of the Fast-Feature classifiers shows their potential for use in categorized overviews for web search results. An analysis of search results from queries based on 250 TREC Robust topics showed that an average of 66% of the top 100 and 61.6% of the top 350 results for each query could be categorized in a rich thematic hierarchy based on the Open Directory.

6.6 Enriching search result interaction with brushing and linking

The general web search interface enabled novel, lightweight interactions with web search results by incorporating a brushing and linking technique. Specifically, brushing the pointer over a category label in the overview had the effect of highlighting any of the currently visible results in that category. Brushing the pointer over a result highlighted the categories that it was in. In study 3 participants did not find the system with the categorized overview significantly more complex than a ranked list of results. This demonstrated that searchers can use and appreciate

lightweight interactions that support, but do not get in the way of, their search tactics and actions.

6.7 Design space of categorized overviews

The description of the SERVICE design decisions and the summary of the design space for categorized overviews (Table 20) will help to guide designers as they develop categorized overview interfaces. The design space summary helps to identify decisions they will need to make during the design process. The design space can serve as a framework for additional research.

6.8 Working system for categorized overviews of web search results

The final contribution of this dissertation research is the SERVICE architecture and implementation technology, which supports two working categorizing search interfaces: AOL music search (Figure 30) and general web search (Figure 1). The SERVICE architecture defines a common Java interface to support easy plug-in of alternate category schemes. The SERVICE technology is comprised of approximately 40 Java class files, which implement nine classifiers plus the two search interfaces. The two search interfaces use JavaServer Pages (JSP), hosted by an Apache Tomcat servlet container. The system runs on Windows and Linux, and uses JDBC to integrate with MySQL and MS-Access databases. The system also implements a client-side logging facility that supports capture of any JavaScript events, including scrolling, mouse clicks and mouseovers, passing the timestamped events back to a Java-based logging tool. Four external data resources containing over 500 MB of data were processed to extract category information, using Java, Perl and PHP. The ideas embedded in the user interface will be useful to designers of other search interfaces,

and the SERVICE system is available to researchers at the categorized overview project page (http://www.cs.umd.edu/hcil/categorizedoverview/). This will provide a flexible, extensible platform for additional research in categorizing search interfaces.

Chapter 7: Future work

7.1 Evaluation of exploratory search interfaces

Evaluation of exploratory search interfaces is an exciting research challenge (White, Muresan, & Marchionini, 2006; White, Kules, Drucker, & schraefel, 2006). Task-based evaluation of exploratory search interfaces using controlled experiments has been effective for showing subjective satisfaction differences between interfaces, but less effective at showing objective differences in task performance, particularly in task outcomes. (Kabel, Hoog, Wielinga, & Anjewierden, 2004; Yee, Swearingen, Li, & Hearst, 2003). Controlled experiments and in-depth case studies are two approaches to evaluation of exploratory search interfaces.

Three factors may have contributed to the lack of objective differences in study 3: the proportion of uncategorized results, the structure of the hierarchies, and the degree to which the task depended on an overview. Controlled experiments may help quantify the effect of each factor in an exploratory search context. Future research in this area should carefully construct the topics to ensure that they are indeed distinguishable by breadth. The broad/narrow concept should be operationally defined in terms of specific criteria, such as searcher perception, or in relation to a specific set of categories (e.g. distribution of search results), and tested with pilot subjects. Studies of exploratory search should also account for individual differences. Differences in cognitive abilities, cognitive styles, and problem-solving styles have been shown to affect search behavior and outcome (Kim & Allen, 2002; Wang, Hawk, & Tenopir,

2000). This appeared to be particularly true for the exploratory search tasks used in these three studies.

The situated nature of exploratory search tasks can lead to many different, but successful, task outcomes for different searchers. In-depth, longitudinal case studies have been used to evaluate information visualization interfaces and creativity support tools (Shneiderman et al., 2006; Shneiderman & Plaisant, 2006). These techniques integrate ethnographic and quantitative methods, using participant observation, surveys, interviews, and usage logs to study users performing complex tasks with individually defined goals. These techniques may be beneficial for investigating how searchers adapt their tactics when rich web search interfaces like categorized overviews are available.

7.2 Structure of category hierarchies for search results

Research on web directories generally indicates that broad, shallow hierarchies are desirable. These studies have typically used known-item or other narrowly defined search tasks. Does the exploratory search task benefit from a different seat of breadth, depth and size trade-offs? Does the content domain affect these trade-offs?

Zaphiris, Shneiderman & Norman (2002) found that expandable menus outperformed sequential menus on hierarchies of depth 2 or 3, but performed poorer than sequential menus with hierarchies of depth 4. As with other studies, they used narrowly defined search tasks with a single correct answer. They speculate that fully expanded hierarchy (of depth 4) became unwieldy for users. Supporting hierarchy

customization operations “on-the-fly” as users explore search results may ameliorate that by allowing them to promote and move sub-trees of interest. However, that benefit could be offset by the additional training and possible cognitive effort. A comparison of sequential menus versus expandable outliners in this problem domain could yield different results than Zaphiris, et al. observed, and could deepen our understanding of the trade-offs inherent hierarchical displays.

7.3 Graphical overviews of search results

Graphical displays of web search results, inspired by the success of information visualization for abstract data, are a promising way to improve information retrieval. They have yielded mixed results to date, though. This dissertation has argued that designers of first generation tools (e.g., Grokker and Kartoo) overlooked the ongoing importance of text in their zeal to reap the perceptual benefits of graphical displays. The analysis and principles begin to address the graphical elements of categorized overviews, but have not yet been theoretically or empirically validated. Compact graphical overviews, paired with search result lists, are one promising research direction. This approach does impose moderate to severe size constraints on the graphical elements. Information visualization techniques like GRIDL (Shneiderman, Feldman, Rose, & Grau, 2000), SuperTable (Klein, Müller, Reiterer, & Eibl, 2002), and WebTOC (Nation, Plaisant, Marchionini, & Komlodi, 1997), and the treemaps used in study 1are starting points, and may provide additional opportunities for lightweight interaction with search results, in the spirit of dynamic queries. Additional research is needed to better understand the tasks as well as the fundamental perceptual and cognitive processes that will benefit.

7.4 Leveraging the Semantic Web

The Semantic Web community advocates the development of machine usable metadata to support automated resource discovery and reasoning, but there is growing recognition that both human and automated agents can benefit from interoperability between metadata standards. A plethora of proposals and standards purport to address the needs of classification users in multiple fields. Topic Maps, the Simple Knowledge Organization System (SKOS), and other proposals for the interchange of thesauri, classifications and ontologies promise a way to distribute classifications widely and maybe even to interconnect them at strategic points. Documenting and distributing the instantiated algorithms and rules for categorizing items into a classification has not yet been addressed. Additional research in this area could extend the fast-feature classifiers to take advantage of work done by projects such as the Dublin Core Metadata Initiative, the Open Archives Initiative, and CITIDEL to find, harvest, and integrate external metadata. Collaborative taxonomies, or folksonomies, like Flikr (flikr.com) and del.icio.us (del.icio.us) could be incorporated into categorized overviews. Collaborative taxonomies are not controlled vocabularies, but social forces encourage evolution toward a common set of tags. Both services provide application programmer interfaces to their tagging engines.

7.5 Lightweight customization of categories

The formative study results motivate a lightweight mechanism for customizing hierarchies. The need to restructure and reorganize hierarchies was highlighted by during the development of the SERVICE system and the final study. Existing taxonomic maintenance tools are designed to manage extensive metadata for

taxonomies and classifications. They are full-featured, complex and require a commitment of time to learn and use. There may be value to end-users and designers in a lightweight tool to customize rich category hierarchies. This could allow motivated end-users (perhaps “power users”) to customize hierarchies for niche uses.

Appendix A: Study 1 – Perspectives identified by subjects

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 34. 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-Michican	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 35. 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 36. 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**	

Appendix B: Study 1 – Unusual results identified by subjects

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 37. 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
incorrectly categorized 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 38. 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 39. 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
incorrectly categorized page		1		1
homeland security		1		1
computer aided manufacturing		1		1
why not more wacky sites			1	1
Total	**13**	**11**	**9**	

Appendix C: Study 3 – Paper materials

Study Introduction
IRB Project: User Interfaces for Public Access Information Systems

Thank you for agreeing to participate in this study. In a moment I'll ask you to fill out a few forms, but first I'll give you a little background.

Although people often use search engines to find a particular piece of information or a specific web page, they also search when they want to explore more generally, for example to find out what information is available about an issue of concern, or to start learning about an unfamiliar subject.

A journalist could be doing background research for an article or a person might want to learn more about a friend's health problem. We call this exploratory searching. Can you think of an instance when you did a search like that?

We are conducting this study to learn more about this kind of searching. To do that we have to observe how real people conduct these searches – how they explore, gather information, and make sense of what they find.

That's why your help is so important. In this study, you will perform several of these exploratory searches. I will provide a scenario, asking you to imagine that you have a particular need for information from the Web and then you will use a search engine to gather information to satisfy that need.

You will use two experimental systems that implement new ways of retrieving and presenting search results. You'll do 4 searches total, with a short break in the middle. Before and after each search, you will complete short questionnaires. At the end of the session, which will last about 2 hours, you will receive a $30 reward in recognition of your help.

Do you have any questions at this point? Okay, if you would please read and sign this informed consent form, we can get started. **[Informed Consent; offer water; turn off cellphones]**

Ok, could you fill out this short questionnaire? **[Entry questionnaire; 3-button mouse; tabbed browser]**

Now I'll show you a short video. **[training video (Portsmouth first => Portsmouth+Collector.avi, Kittery first => Collector.avi)]**

[Both note] The search engine returns 100 results, listed with the most relevant pages at the top, like a regular search engine.

[Portsmouth note] It's important to remember that the web sites for these pages **[point to results]** are cataloged, not necessarily the specific pages. And the cataloging is done by human editors, so there will always be some sites that haven't been categorized yet. Those pages might appear near the top of the results, and might be good pages, even though they in the Uncategorized group. The Empty item is a list of categories that don't have any results in them. Sometimes it's useful to see what categories your results are not coming from.

Do you have any questions?

[Training task]

Now we'll walk through the scenario that we're using for this session and give you a chance to use the system.

Imagine that you are a reporter for a national newspaper. Due to some recent events, your editor has just asked you to generate a list of ideas for a series of articles on [**urban sprawl]**. There's a meeting in an hour, so she doesn't need a lot of detail, but she wants a diverse list of 8-10 (or more) ideas for discussion.

They should cover many different aspects of the topic, to appeal to a broad range of readers. Unusual or provocative ideas are good. You have about 10 minutes to conduct a short web search to find out what information is available and generate the ideas.

Your results will be judged (by your imaginary editor) on the quality and diversity of ideas. For example, "public health impact" would be an okay idea. and "obesity as a public health impact of urban sprawl" would be even better, because it is a bit more specific.

As you use the search engine to explore and generate article ideas, enter them in the Collector form and include the web page that inspired your idea. It is important that you enter the ideas, not notes like "a good page". Think of this list **[point to the Collector]** as a bullet list for the discussion.

Please think out loud as you take each action, for example, when you enter a query, click on something, or scroll a page. Briefly say why you did it and then tell me your reaction to the system's response.

I'm also interested in what's good or bad, problems or insights, and anything confusing. You don't have to describe what you are doing, since we're recording it. We'll spend a few minutes on this now.

[Start Camtasia. Encourage them to explore the system and generate 2-3 ideas. If they haven't done all items on checklist, prompt them. After training task:] Do you have any questions? **[Stop Camtasia]**

Please remember that you are not being tested. Instead, you are helping us to evaluate these systems by using them to the best of your ability and producing the best results you can.

Participant ID: _____

Informed Consent Form

Title of the Project: User Interfaces for Public Access Information Systems

Purpose: This research and the experiment aim to improve user interfaces for browsing and searching data on computer.

Procedure: You will be asked to use different user interfaces to perform various tasks. This experiment is not for testing your ability to use the interfaces but to compare the interfaces themselves. For each interface you will perform a set of simple tasks. You will be asked to study computer displays or listen to sounds produced by a computer for each task. This process will last approximatively an hour.

Investigators: Ben Shneiderman, Catherine Plaisant, Haixia Zhao, Bill Kules

HCIL/UMIACS
University of Maryland, College Park, MD 20742
Tel: (301) 405 2768

Experimental Consent Agreement

1. I am older than 18 years.
2. I have freely volunteered to participate in this experiment.
3. I have been informed in advance as to what my task(s) would be and what procedures would be followed.
4. I have been given the opportunity to ask questions, and know when I will be able to ask more questions.
5. I am aware that I have the right to withdraw consent and discontinue participation at any time, without penalty.
6. I understand that there are no known risks associated with participating in the experiment.
7. I understand that the benefit of this research will the development of interfaces that can deal with large and complex data sets, and that I will not gain any personal benefits.
8. I will receive a copy of the signed agreement.
9. My signature below may be taken as affirmation of all of the above, prior to participation.

PRINTED NAME	DATE	SIGNATURE
_____	_____	_____

Renewal 2005--Plaisant/Shneiderman -- IRB#01120

Training Task Checklists
IRB Project: User Interfaces for Public Access Information Systems

Participant ID: _____ Sequence: ____/____/____/____ Date: _____

Check mark (✓) if they do it on their own, "P" if prompted.

Portsmouth

__ Enter query

__ Pointer over category; view subcats & highlighted results

__ Pointer over result; view highlighted cats

__ Filter on category

__ Exclude category

__ Pointer over Empty pseudo category; view list

__ Show all results

Collector

__ Collects idea & link (from a web page)

__ Collects idea & link (results list)

__ Collects idea & link (URL location bar)

__ Generates 2+ appropriate ideas

Study Procedural Checklist
IRB Project: User Interfaces for Public Access Information Systems

Participant ID: _____ Sequence: ____/____/____/____ Date: _____

Start Time	**Cap-ture**	**Step**
		Introduction
		Informed consent form
		Entry questionnaire
		Video: __ Collector.avi __ Portsmouth+Collector.avi
	✓	Training task – Time limit: 8
		Task 1: __Br1 __Br2 __N1 __N2
		Pre-search questionnaire
	✓	Search – Limit: 12 End: _____
	✓	Post-search questionnaire
		Task 2: __Br1 __Br2 __N1 __N2
		Pre-search questionnaire
	✓	Search – Limit: 12 End: _____
	✓	Post-search questionnaire
		Break **(restart Eclipse)**
		Video: __Portsmouth.avi __None
	✓	Training task: __Yes __No
		Task 3: __Br1 __Br2 __N1 __N2
		Pre-search questionnaire
	✓	Search – Limit: 12 End: _____
	✓	Post-search questionnaire
		Task 4: __Br1 __Br2 __N1 __N2
		Pre-search questionnaire
	✓	Search – Limit: 12 End: _____
	✓	Post-search questionnaire
	✓	Exit interview
		Reward; receipt; copy of form

Exit Interview Questions
IRB Project: User Interfaces for Public Access Information Systems

Participant ID: _____ Date: _____

1. **Which system would you rather use for these tasks (Kittery, Portsmouth or no preference) [order: ____/____/____/____]:**
   a. **[K, P, no-op] Find the home page for the daily newspaper in Concord, NH, The Concord Monitor.**
   b. **[K, P, no-op] Find information on caring for a pet gerbil.**
   c. **[K, P, no-op] Start looking for information to help you select and buy a new digital camera.**
   d. **[K, P, no-op] Learn about U.S. business investment in Africa.**
2. **How do you feel about the quality of ideas that you generated for each task? Rank (worst) ____ - ____ - ____ - ____ (best)**
3. **Did the categorized overview change the way you searched? Can you describe an example? Why?**
4. **Can you describe an example where the categorized overview [helped/hindered, frustrated or mislead – whichever not indicated in previous question]?**
5. **Did you notice any difference in how you used the categorized overview each time? Can you describe an example?**
6. **Did you ever read the category pop-ups' subcategories? Why/why not? How did that help you decide?**
7. **[Show Leonardo da Vinci] What kind of results do you expect to see if you did this search and clicked on the Kids and Teens link? Computers? Reference? K&T _____ Computer _____ / _____ Reference _____**
8. **Did you ever use the Uncategorized pseudo-category? Example? Empty?**
9. **These systems display 100 results at a time. Do you have any thoughts on this? Would you typically use all 100?**
10. **How similar or dissimilar is this scenario what a journalist would really need to do? How does it differ?**
11. **How similar or dissimilar are the searches that you did to a search that would do? How/how not?**
12. **Could you compare the difficulty finding information on the topics? Rank them Rank (easiest) ____ - ____ - ____ - ____ (hardest)**
13. **How much time would you have spent on the tasks if there were no specific time limit?**
14. **Did you notice any changes in your energy level or alertness over the course of the session?**
15. **Do you have any suggestions for additions or changes to the categories?**
16. **Do you have any suggestions for changes to the system… features? layout? interactions?**

Exit Interview Questions
IRB Project: User Interfaces for Public Access Information Systems

Participant ID: _____ Date: _____

On a scale from 1 to 9, please rate how narrow or broad each of the topics was:

	Narrow **Broad**
Aging workforce	**1 2 3 4 5 6 7 8 9**
Human smuggling	**1 2 3 4 5 6 7 8 9**
International art crime	**1 2 3 4 5 6 7 8 9**
Workplace allergies	**1 2 3 4 5 6 7 8 9**

Appendix D: Study 3 – Online questionnaires

Entry Questionnaire
User Interfaces for Public Access Information Systems
Web Search Study 3
Investigator: Bill Kules

This questionnaire provides us with background information that helps us analyse the answers you give in later stages of this experiment.

Questions marked with a * are required.

*1. Participant ID (provided by experiment monitor)

*2. Your age

*3. Your gender

Female
Male

*4. Your occupation (if student, department/major)

*5. Highest level of education achieved

High school
Part way through undergraduate program
Undergraduate degree
Part way through graduate program
Graduate degree (e.g. Masters, PhD)

*6. Web searching experience – How long have you used search engines to look for information on the Web?

Less than 6 months
6-12 months
1-3 years
More than 3 years

*7. Search frequency – How often do you use a search engine to search for information on the Web?

Less than once a week
1-2 times per week, but less than once a day
At least once a day

*8. How do you rate your searching skills?

1	2	3	4	5
Novice				Expert

*9. When you search on the Web, how often do you find the information you are looking for?

Never or almost never
Rarely
Some of the time
Most of the time
Always or almost always

*10. What web search engines do you use frequently (select all that apply)?

Google
Yahoo!
MSN
AOL
Other:

*11. What type of information do you normally search for on the web (select all that apply)?

Research for classes
Research for work
Job searching
Entertainment/recreation
Places or products
News or information on events
Locate people (email, addresses, phone numbers, etc.)
Commerce, travel, employment or economy

Society, culture, ethnicity or religion
Government information
Other:

Pre-Search Questionnaire

Questions marked with a * are required.

*1. Participant ID (provided by experiment monitor)

*2. Sequence number (provided by experiment monitor)

*3. System (provided by experiment monitor)

Kittery
Portsmouth

*4. Topic (provided by experiment monitor)

*5. How familiar are you with this topic now?

1	2	3	4	5	6	7	8	9
Not at all								Very

*6. How interested are you in this topic now?

1	2	3	4	5	6	7	8	9
Not at all								Very

*7. How confident are you that you can find useful information about your topic on the Web?

1	2	3	4	5	6	7	8	9
Not at all								Very

*8. How do you feel about your ability to complete the task at this point?

	1	2	3	4	5	6	7	8	9	
Very uncertain										Very certain
Pessimistic										Optimistic
Confused										Clear
Doubtful										Confident

Post-Search Questionnaire

Questions marked with a * are required.

*1. Participant ID (provided by experiment monitor)

*2. Sequence number (provided by experiment monitor)

*3. System (provided by experiment monitor)

Kittery
Portsmouth

*4. Topic (provided by experiment monitor)

5. What are your thoughts at this point (you may write or comment out loud)?

*6. How familiar are you with this topic now?

	1	2	3	4	5	6	7	8	9	
Not at all										Very

*7. How interested are you in this topic now?

	1	2	3	4	5	6	7	8	9	
Not at all										Very

*8. How confident are you that you can find useful information about your topic on the Web?

1 2 3 4 5 6 7 8 9
Not at all Very

*9. How do you feel about your ability to complete the task at this point?

1 2 3 4 5 6 7 8 9
Very uncertain Very certain
Pessimistic Optimistic
Confused Clear
Doubtful Confident

*10. The search I performed was:

1 2 3 4 5 6 7 8 9
Stressful Relaxing
Boring Interesting
Tiring Restful
Difficult Easy

*11. How much progress did you make on generating good ideas?

1 2 3 4 5 6 7 8 9
None I'm ready to give my editor the list

*12. How useful was the information you found?

1 2 3 4 5 6 7 8 9
Not at all useful Very useful

*13. How difficult was it to explore / navigate the results of your search?

1 2 3 4 5 6 7 8 9
Very hard Very easy

*14. I was able to get a good overview of the information available on the Web for this topic:

1 2 3 4 5 6 7 8 9

Strongly disagree Strongly agree

*15. The system helped me organize my search results:

1 2 3 4 5 6 7 8 9
Strongly disagree Strongly agree

*16. The system helped me find useful pages:

1 2 3 4 5 6 7 8 9
Strongly disagree Strongly agree

*17. The system helped me assess the results of my queries to decide what to do next:

1 2 3 4 5 6 7 8 9
Strongly disagree Strongly agree

*18. Please indicate how well these descriptions apply to this system:

1 2 3 4 5 6 7 8 9	
Terrible	Wonderful
Difficult to use	Easy to use
Dull	Stimulating
Frustrating	Satisfying
Complex	Simple
Too Slow	Fast Enough
Overwhelming	Manageable
Disorganized	Organized

Bibliography

1. Aguilar, F. J. (1988). *General Managers in Action*. New York, NY: Oxford University Press.

2. Ahlberg, C., Shneiderman, B. (1993). Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 313-317). New York: ACM Press.

3. Allen, R. (1995). Two digital library Interfaces that exploit hierarchical structure, *DAGS95: Electronic Publishing and the Information Superhighway*.

4. Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

5. Ask.com. (2005). *About Ask.com: IQ*. Retrieved April 18, 2006, from http://sp.ask.com/en/docs/iq/iq.shtml.

6. Aula, A. (2004). Enhancing the readability of search result summaries. In *Proceedings Volume 2 of the Conference HCI 2004: Design for Life, Leeds, UK*. Retrieved April 27, 2006, from http://www.cs.uta.fi/~aula/aula_summary.pdf.

7. Aula, A., Jhaveri, N., & Käki, M. (2005). Information search and re-access strategies of experienced web users. In *Proceedings of the 14th International Conference on the World Wide Web, Chiba, Japan* (pp. 583-592). New York: ACM Press.

8. Bates, M. (1990). Where should the person stop and the information search interface start. *Information Processing and Management, 26*(5), 575-591.

9. Bates, M. J. (1979). Information search tactics. *Journal of the American Society for Information Science, 30*, 205-214.

10. Bates, M. J. (1989). The design of browsing and berrypicking techniques for the on-line search interface. *Online Review, 13*(5), 407-431.

11. Becks, A., Seeling, C., & Minkenberg, R. (2002). Benefits of document maps for text access in knowledge management: A comparative study. In *Proceedings of the 2002 ACM Symposium on Applied Computing* (pp. 621-626). New York: ACM Press.

12. Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5*, 133-143.

13. Bell, D. J., & Ruthven, I. (2004). Searchers' assessments of task complexity for Web searching. In S. Macdonald & J. Tait (Eds.), *Proceedings of the 26th BCS-IRSG Euopean Conference on Information Retrieval* (pp. 57-71). Berlin: Springer-Verlag.

14. Bhavnani, S. K., & Bates, M. J. (2002). Separating the knowledge layers: Cognitive analysis of search knowledge through hierarchical goal decompositions. In *Proceedings of the American Society for Information Science and Technology Annual Meeting* (Vol. 39, pp. 204-213). Medford, NJ: Information Today.

15. Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation, 56*(1), 71-90.

16. Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research, 8*(3), paper no. 152. Retrieved April 17, 2006, from http://informationr.net/ir/8-3/paper152.html.

17. Bowker, G., & Starr, S. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge MA: MIT Press.

18. Broder, A. (2002). A taxonomy of web search. *SIGIR Forum, 36*(2), 3-10.

19. 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.

20. Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Fransisco: Morgan Kaufmann.

21. 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.

22. 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.

23. 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 2nd USENIX Symposium on Internet Technologies and Systems, Boulder, CO. Retrieved April 17, 2006, from http://www.sims.berkeley.edu/~hearst/papers/usits99/.

24. Chirita, P. A., Nejdl, W., Paiu, R., & Kohlschütter, C. (2005). Using ODP metadata to personalize search. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil* (pp. 178-185). New York: ACM Press.

25. Choo, C. W., Detlor, B., & Turnbull, D. (2000). *Web Work: Information Seeking and Knowledge Work on the World Wide Web*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

26. Cockburn, A., & Jones, S. (1996). Which way now? Analysing and easing inadequacies in WWW navigation. *International Journal of Human-Computer Studies, 45*(1), 105-129.

27. Cousins, S. B., Paepcke, A., Winograd, T., Bier, E. A., & Pier, K. (1997). The digital library integrated task environment (DLITE). In *Proceedings of the Second ACM International Conference on Digital Libraries, Philadelphia, Pennsylvania* (pp. 142-151). New York: ACM Press.

28. Cunha, C., Bestavros, A., & Crovella, M. (1995). *Characteristics of WWW client-based traces* (No. TR-95-010): Boston University. Retrieved January 24, 2005, from http://cs-www.bu.edu/faculty/crovella/paper-archive/TR-95-010/paper.html.

29. Dervin, B., & Nilan, M. (1986). Information needs and uses. In M. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 21, pp. 3-33). White Plains, New York: Knowledge Industries.

30. Drori, O. (2003). Display of search results in Google-based Yahoo! vs. LCC&K interfaces: A comparison study. *Proceedings of Informing Science 2003 Conference, Pori, Finland*. Retrieved April 27, 2006, from http://shum.huji.ac.il/~offerd/papers/drori062003-b.pdf.

31. Drori, O., & Alon, N. (2003). Using documents classification for displaying search results list. *Journal of Information Science, 29*(2), 97-106.

32. Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA* (pp. 277-284). New York: ACM Press.

33. 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.

34. Durand, D., & Kahn, P. (1998). MAPA: A system for inducing and visualizing hierarchy in websites. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia* (pp. 66-76). New York: ACM Press.

35. Egan, D. E., Remde, J. R., Gomez, L. M., Landauer, T. K., Eberhardt, J., & Lochbaum, C. C. (1989). Formative design evaluation of SuperBook. *ACM Transactions on Information Systems, 7*(1), 30-57.

36. Ellis, D. (1989). A behavioral model for information retrieval system design. *Journal of Information Science, 15*(4-5), 237-247.

37. Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

38. Fidel, R. (1985). Moves in online searching. *Online Review, 9*(1), 61-74.

39. Furnas, G. W. (1997). Effective view navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 367-374). New York: ACM Press.

40. Furnas, G. W., & Rauch, S. J. (1998). Considerations for information environments and the NaviQue workspace. In *Proceedings of the Third ACM Conference on Digital libraries, Pittsburgh, PA* (pp. 79-88). New York: ACM Press.

41. Garcia, E., & Sicilia, M.-Á. (2003). User interface tactics in ontology-based information seeking. *Psychology Journal, 1*(3), 242-255.

42. Garfield, E. (2005). *The agony and the ecstasy-The history and meaning of the journal impact factor.* Paper presented at the International Congress on Peer Review and Biomedical Publication, Chicago, IL. Retrieved April 16, 2006, from http://garfield.library.upenn.edu/papers/jifchicago2005.pdf.

43. Ginsburg, M. (2004). Visualizing digital libraries with open standards. *Communications of the Association for Information Systems, 13*, 336-356.

44. Golovchinsky, G. (1997). Queries? Links? Is there a difference? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA* (pp. 407-414). New York: ACM Press.

45. Greene, S., Marchionini, G., Plaisant, C., & Shneiderman, B. (2000). Previews and overviews in digital libraries: Designing surrogates to support visual information-seeking. *Journal of the American Society for Information Science, 51*(3), 380-393.

46. Guba, E. G., & Lincoln, Y. S. (1982). Epistemological and methodological bases of naturalistic inquiry. *Educational Communication and Technology, 30*(4), 233-252.

47. Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 59-66). New York: ACM Press.

48. 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.

49. Hearst, M. A. (1999). The use of categories and clusters for organizing retrieval results. In T. Strzalkowski (Ed.), *Natural Language Information Retrieval* (pp. 333-373). Boston: Kluwer Academic Publishers.

50. Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM, 49*(4), 59-61.

51. Hearst, M. A., & Karadi, C. (1997). Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 246-255). New York: ACM Press.

52. Hearst, M. A., & Pedersen, J. O. (1996). Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland* (pp. 76-84). New York: ACM Press.

53. Hendry, D. (to appear). Workspaces for Search. *Journal of the American Society for Information Science and Technology*. from http://faculty.washington.edu/dhendry/docs/jasis2004.pdf.

54. Hendry, D., & Harper, D. (1997). An informal information-seeking environment. *Journal of the American Society for Information Science, 48*(11), 1036-1048.

55. Hert, C. (2002). *Developing and evaluating scenarios for use in designing the National Statistical Knowledge Network.* Retrieved February 15, 2006, from http://ils.unc.edu/govstat/papers/scenario_paper_nov_14_2002.doc.

56. Hochheiser, H., & Shneiderman, B. (1999). Performance benefits of simultaneous over sequential menus as task complexity increases. *International Journal of Human Computer Interaction, 12*(2), 173-192.

57. Jaccard, J. (1983). *Statistics for the Behavioral Sciences*. Belmont, CA: Wadsworth Publishing Company.

58. Jacob, E. (2004). Classification and categorization: A difference that makes a difference. *Library Trends, 52*(3), 515-540.

59. Janecek, P., & Pu, P. (2005). An evaluation of semantic fisheye views for opportunistic search in an annotated image collection. *Journal of Digital Libraries, 5*(1), 42-56.

60. 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.

61. Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management, 36*, 207-227.

62. Järvelin, K., & Ingwersen, P. (2004). Information seeking research needs extension towards tasks and technology. *Information Research, 10*(1), paper 212. Retrieved April 27, 2006, from http://informationr.net/ir/10-1/paper212.html.

63. Kaasten, S., & Greenberg, S. (2001). Integrating Back, History and Bookmarks in Web Browsers. In *CHI '01 Extended Abstracts on Human Factors in Computer Cystems* (pp. 379-380). New York: ACM Press.

64. Kabel, S., Hoog, R. d., Wielinga, B. J., & Anjewierden, A. (2004). The added value of task and ontology-based markup for information retrieval. *Journal of the American Society for Information Science and Technology, 55*(4), 348-362.

65. 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.

66. Käki, M. (2005). Findex: search result categories help users when document ranking fails. In *Proceeding of the SIGCHI Conference on Human Factors in Computing Systems, Portland, OR* (pp. 131-140). New York: ACM Press.

67. Kim, K.-S., & Allen, B. (2002). Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology, 53*(2), 109-119.

68. Kleiboemer, A., Lazear, M., & Pedersen, J. (1996). Tailoring a retrieval system for naive users. In *Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval*.

69. Klein, P., Müller, F., Reiterer, H., & Eibl, M. (2002). Visual information retrieval with the SuperTable + Scatterplot. In *Proceedings of the Sixth International Conference on Information Visualisation (IV '02)* (pp. 70-75). New York: IEEE Computer Society.

70. Klein, P., Reiterer, H., Müller, F., & Limbach, T. (2003). Metadata visualisation with VisMeB. In *Proceedings of the Seventh International Conference on Information Visualization (IV'03)* (pp. 600-605). New York: IEEE Computer Society.

71. Koenemann, J., & Belkin, N. J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems: Common Ground, Vancouver, British Columbia, Canada (pp. 205-212). New York: ACM Press.

72. 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.

73. Kules, B., Kustanowitz, J., & Shneiderman, B. (to appear). Categorizing web search results into meaningful and stable categories using Fast-Feature techniques. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*.

74. Kules, B., & Shneiderman, B. (2003). Designing a metadata-driven visual information browser for federal statistics. In *Proceedings of the 2003 National Conference on Digital Government Research* (pp. 117-122). Retrieved April 27, 2006, from http://hcil.cs.umd.edu/trs/2003-08/2003-08.pdf.

75. 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.* Paper presented at the Third Annual Workshop on HCI Research in MIS, Washington, DC. Retrieved April 27, 2006, from http://hcil.cs.umd.edu/trs/2004-38/2004-38.html.

76. Kunz, C. (2003). SERGIO - An Interface for context driven knowledge retrieval. In *Proceedings of eChallenges, Bologna, Italy, 2003*. Retrieved April 27, 2006, from http://www.hci.iao.fraunhofer.de/uploads/tx_publications/Kunz2003_SERGIO_Proceedings_of_eChallenges.pdf.

77. Kunz, C., & Botsch, V. (2002). Visual representation and contextualization of search results – List and Matrix Browser. In *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities* (pp. 229-234): Firenze University Press. Retrieved April 27, 2006, from http://www.bncf.net/dc2002/program/ft/poster10.pdf.

78. Kwasnik, B. H. (1999). The role of classification in knowledge representation and discovery. *Library Trends, 48*(1), 22-47.

79. Lamping, J., & Rao, R. (1996). The Hyperbolic Browser: A focus + context technique for visualizing large hierarchies. *Journal of Visual Languages and Computing, 7*(1), 33-55.

80. Larson, K., & Czerwinski, M. (1998). Web page design: Implications of memory, structure and scent for information retrieval. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 25-32). New York: ACM Press.

81. Louie, A. J., Maddox, E. L., & Washington, W. (2003). *Using faceted classification to provide structure for information architecture.* Paper presented at the The 62nd ASIS Annual Meeting, Washington, D.C. Retrieved April 17, 2006, from http://depts.washington.edu/pettt/presentations/conf_2003/IASummit.pdf.

82. Marchionini, G. (1995). *Information Seeking in Electronic Environments*: Cambridge University Press.

83. 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.

84. Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 129*, 592-613. Retrieved April 19, 2006, from http://www.psy.utexas.edu/psy/faculty/Markman/PB03.pdf.

85. 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.

86. Matsuda, K., & Fukushima, T. (1999). Task-oriented World Wide Web retrieval by document type classification. In *Proceedings of the Eighth International Conference on Information and Knowledge Management, Kansas City, MO* (pp. 109-113). New York: ACM Press.

87. Milic-Frayling, N., Jones, R., Rodden, K., Smyth, G., Blackwell, A., & Sommerer, R. (2004). Smartback: supporting users in back navigation. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 63-71). New York: ACM Press.

88. Miller, D. (1981). The depth/breadth tradeoff in hierarchical computer menus. *Proceedings of the Human Factors Society*, 296-300.

89. Nation, D. A., Plaisant, C., Marchionini, G., & Komlodi, A. (1997). Visualizing websites using a hierarchical table of contents browser: WebTOC. *Proceedings of the Third Conference on Human Factors and the Web*. Retrieved April 27, 2006, from http://hcil.cs.umd.edu/trs/97-10/97-10.html.

90. Nielsen, J., Clemmensen, T., & Yssing, C. (2002). Getting access to what goes on in people's heads? - Reflections on the think-aloud technique. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction, Aarhus, Denmark* (pp. 101-110). New York: ACM Press.

91. Niemela, M., & Saariluoma, P. (2003). Layout attributes and recall. *Behaviour & Information Technology, 22*(5), 353-363.

92. Norman, K. (1991). *The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface*. Norwood, NJ: Ablex Publishing Corporation.

93. Nowell, L. T., France, R. K., Hix, D., Heath, L. S., & Fox, E. A. (1996). Visualizing search results: Some alternatives to query-document similarity. In *Proceedings of the Nineteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 67-75). New York: ACM Press.

94. Periakaruppan, R., & Nemeth, E. (1999). GTrace - A graphical traceroute tool. In *Proceedings of the 13th USENIX Conference on System Administration, Seattle, WA* (pp. 69-78): USENIX Association. Retrieved April 18, 2006, from http://www.caida.org/publications/papers/1999/GTrace/GTrace.pdf.

95. Perugini, S., McDevitt, K., Richardson, R., Manuel Perez-Quiones, Shen, R., Ramakrishnan, N., et al. (2004). Enhancing usability in CITIDEL: multimodal, multilingual, and interactive visualization interfaces, *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. Tuscon, AZ, USA: ACM Press.

96. Pirolli, P., & Card, S. (1995). Information foraging in information access environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 51-58). New York: ACM Press.

97. Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground, Vancouver, British Columbia, Canada* (pp. 213-220). New York: ACM Press.

98. Pirolli, P. L., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*(4), 643-675.

99. Pollitt, S. (1997). Interactive information retrieval based on faceted classification using views in knowledge organization for information retrieval, *Sixth International Study Conference on Classification Research*. University College London, 16-19 June 1997.

100. Pratt, W., Hearst, M. A., & Fagan, L. M. (1999). A knowledge-based approach to organizing retrieved documents. In *Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, FL* (pp. 80-85): American Association for Artificial Intelligence. Retrieved April 27, 2006, from http://www.sims.berkeley.edu/~hearst/papers/AAAI-99.pdf.

101. R Development Core Team. (2005). *R Language Definition*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved April 27, 2006, from http://cran.r-project.org/doc/manuals/R-lang.pdf.

102. Risden, K., Czerwinski, M., Munzner, T., & Cook, D. (2000). An initial examination of ease of use for 2D and 3D information visualizations of Web content. *International Journal of Human-Computer Studies,* 695 - 714.

103. 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. Retrieved April 27, 2006, from http://hcil.cs.umd.edu/trs/2003-36/2003-36.pdf.

104. Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web* (pp. 13-19). New York: ACM Press.

105. Sebrechts, M., Vasilakis, J., Miller, M., Cugini, J., & Laskowski, S. (1999). Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 3-10). New York: ACM Press.

106. Shiri, A. A., & Revie, C. (2000). Thesauri on the web: current developments and trends. *Online Information Review, 24*(4), 273-279.

107. Shneiderman, B., Byrd, D., & Croft, W. B. (1997). Clarifying search: A user-interface framework for text searches. *D-Lib Magazine*. Retrieved April 16, 2006, from http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html.

108. Shneiderman, B., Byrd, D., & Croft, W. B. (1998). Sorting out searching: A user-interface framework for text searches. *Communications of the ACM, 41*(4), 95-98.

109. Shneiderman, B., Feldman, D., Rose, A., & Grau, X. F. (2000). Visualizing digital library search results with categorical and hierarchial axes. In *Proceedings of the Fifth ACM International Conference on Digital Libraries (San Antonio, TX, June 2-7, 2000)* (pp. 57-66). New York: ACM Press.

110. Shneiderman, B., Fischer, G., Czerwinski, M., Resnick, M., Myers, B., Candy, L., et al. (2006). Creativity support tools: Report from a U.S. National Science Foundation sponsored workshop. *International Journal of Human-Computer Interaction, 20*(2), 61-77.

111. Shneiderman, B., & Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (4th ed.). Boston: Pearson/Addison-Wesley.

112. Shneiderman, B., & Plaisant, C. (2006). Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies, *Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*

(BELIV '06): A Workshop of the AVI 2006 International Working Conference. Venezia, Italy.

113. Simon, H. A. (1979). *Models of Thought*. New Haven, CT: Yale University Press.

114. Soergel, D. (1974). *Construction and Maintenance of Indexing Languages and Thesauri*. New York: Wiley.

115. 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.

116. Spink, A., Bateman, J., & Jansen, B. J. (1999). Searching the Web: A survey of EXCITE users. *Internet Research: Electronic Networking Applications and Policy*, *9*(2), 117-128.

117. Spink, A., & Jansen, B. J. (2004). *Web Search: Public Searching of the Web*. New York: Kluwer.

118. Spink, A., Wilson, T. D., Ford, N., Foster, A., & Ellis, D. (2002). Information seeking and mediated searching study. Part 3. Successive searching. *Journal of the American Society for Information Science and Technology*, *53*(9), 716-727.

119. 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.

120. Swan, R., & Allen, J. (1998). Aspect Windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 173-181). New York: ACM Press.

121. Tanin, E., Plaisant, C., & Shneiderman, B. (2000). Browsing large online data with Query Previews. In *Proceedings of the Symposium on New Paradigms in Information Visualization and Manipulation (NPIVM) 2000, Washington, DC*: ACM Press. Retrieved April 27, 2006, from http://citeseer.ist.psu.edu/tanin00browsing.html.

122. Taylor, A. (1999). *The Organization of Information*. Englewood, CO: Libraries Unlimited, Inc.

123. Teitelbaum, R. C., & Granda, R. E. (1983). The effects of positional constancy on searching menus for information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA* (pp. 150-153). New York: ACM Press.

124. Tullis, T. (1988). Screen design. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 377-411). Amsterdam, The Netherlands: Elsevier Science Publishers.

125. 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.

126. Vakkari, P. (2000). eCognition and changes of search terms and tactics during task performance: A longitudinal case study. In *Proceedings of the RIAO 2000 Conference*. Retrieved April 27, 2006, from http://www.info.uta.fi/vakkari/Vakkari_Tactics_RIAO2000.pdf.

127. 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.

128. Vickery, B. C. (1960). *Faceted Classification: A Guide to Construction and Use of Special Schemes*. London: Aslib.

129. Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. *Information Processing & Management, 36*(2), 229-251.

130. Watters, C., & Amoudi, G. (2003). GeoSearcher: Location-based ranking of search engine results. *Journal of the American Society for Information Science and Technology, 54*(2), 140-151.

131. Wen, J. (2003). Post-valued recall web pages: User disorientation hits the big time. *IT & Society, 1*(3), 184-194. Retrieved April 27, 2006, from http://www.stanford.edu/group/siqss/itandsociety/v01i03/v01i03a10.pdf.

132. White, R., Muresan, G., & Marchionini, G. (2006). *Evaluating Exploratory Search Systems - SIGIR 2006 Workshop Call for Papers*. Retrieved April 24, 2006, from http://www.umiacs.umd.edu/~ryen/eess.

133. White, R. W., Kules, B., Drucker, S. M., & schraefel, m. c. (2006). Supporting exploratory search. *Communications of the ACM, 49*(4), 36-39.

134. Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology, 55*(3), 246-258.

135. Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems, Ft. Lauderdale, FL* (pp. 401-408). New York: ACM Press.

136. Zamir, O., & Etzioni, O. (1998). Web document clustering: a feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia* (pp. 46-54). New York: ACM Press.

137. Zamir, O., & Etzioni, O. (1999). Grouper: a dynamic clustering interface to Web search results. *Computer Networks, 31*, 1361-1374.

138. Zaphiris, P., & Mtei, L. (1997). *Depth v. Breadth in the Arrangement of Web Links*. Retrieved April 27, 2006, from http://otal.umd.edu/SHORE/bs04.

139. Zaphiris, P., Shneiderman, B., & Norman, K. (2002). Expandable indexes versus sequential menus for searching hierarchies on the World Wide Web. *Behaviour & Information Technology, 21*(3), 201-207.

140. Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., & Ma, J. (2004). Learning to cluster web search results. In *Proceedings of the 27th Annual International Conference on Research and Dvelopment in Information Retrieval, Sheffield, United Kingdom* (pp. 210-217). New York: ACM Press.