Users can change their web search tactics: Design guidelines for categorized overviews

Bill Kules*
Department of Computer Science, Human-Computer Interaction Laboratory,
University of Maryland at College Park, College Park, MD 20742
kules@cua.edu

Ben Shneiderman
Department of Computer Science, Human-Computer Interaction Laboratory,
and Institute for Advanced Computer Studies
University of Maryland at College Park, College Park, MD 20742
ben@cs.umd.edu

Please send correspondence to Bill Kules (kules@cua.edu)
(202) 319-6278 (Kules)
(301) 405-2680 (Shneiderman)
(240) 599-7671 (fax)

Abstract

Categorized overviews of web search results are a promising way to support user exploration, understanding, and discovery. These search interfaces combine a metadata-based overview with the list of search results to enable a rich form of interaction. A study of 24 sophisticated users carrying out complex tasks suggests how searchers may adapt their search tactics when using categorized overviews. This mixed methods study evaluated categorized overviews of web search results organized into thematic, geographic, and government categories. Participants conducted four exploratory searches during a 2-hour session to generate ideas for newspaper articles about specified topics such as "human smuggling." Results showed that subjects explored deeper while feeling more organized, and that the categorized overview helped subjects better assess their results, although no significant differences were detected in the quality of the article ideas. A qualitative analysis of searcher comments identified seven tactics that participants reported adopting when using categorized overviews. This paper concludes by proposing a set of guidelines for the design of exploratory search interfaces. An understanding of the impact of categorized overviews on search tactics will be useful to web search researchers, search interface designers, information architects and web developers.

* Author's current address: The School of Library and Information Science, The Catholic University of America, Washington, DC 20064.

Prioritized Open Issues/Questions/To Dos

- Finish related work section
- Convert to IP&M submission format
- Consider moving tables 4-7 to appendix or deleting. If we delete, we should revise description of qualitative analysis to explain that codes are available in dissertation.
- Consider another guideline: provide clear feedback on selected/narrowed categories
- Should we include any appendices? E.g. questionnaires. I'm inclined not to.
- Verify consistent use of "categories" instead of "classifications", except where we are specifically discussing classifications
- Down to 12,600 words… still long, but probably okay for submission.

1. Introduction

Categorized overviews of web search results are a promising way to support user exploration, understanding, and discovery. These search interfaces, also referred to as faceted or guided search interfaces, combine a metadata-based overview with the list of search results to enable a rich form of interaction.

The strategies and tactics that searchers use are affected by the capabilities provided by the search interface (M. 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 (M. J. Bates, 1979; Gary Marchionini, 1995). Searchers can take numerous actions while examining search results (M. Bates, 1990; Fidel, 1985; Garcia & Sicilia, 2003; Gary Marchionini, 1995; Shneiderman & Plaisant, 2004; Wildemuth, 2004)

Designers build interfaces to support specific strategies, based on intuition or analysis. But the effect of new capabilities on search tactics may not be what designers anticipate. Unexpected problems may negate expected benefits. Serendipitous possibilities may present to searchers. In response, searchers may adapt their tactics and strategies as they become familiar with the capabilities. Our research seeks to understand how exploratory search systems with rich user interfaces change the way that searchers think about and pursue their searches. What strategies and tactics do exploratory search interfaces enable? And, ultimately, do they enable searchers to achieve their higher-level objectives?

1.1 Research questions

Three 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 behaviors do searchers exhibit when categorized overviews are available?
3. In what ways could the presence of categorized overviews affect the quality of the search outcome?

Evaluation of exploratory search systems is an exciting research challenge (R. White, Muresan, & Marchionini, ; R. W. White, Kules, Drucker, & schraefel, 2006). The situated nature of exploratory search tasks can lead to many different task outcomes for different searchers. This can make it difficult to specify quantitative 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, dead-ends, 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.

This study adopted a mixed methods approach. It shows how a combination of qualitative and quantitative methods can address research questions related to exploratory search. Following a brief description of related work, the experiment design is described Section 3. Section 4 presents the results. Section 5 discusses the results, identifying seven tactics that searchers adopted. Section 6 proposes eight design guidelines suggested or refined by the study. Section 7 concludes with a summary of the contributions and suggestions for future research.

2. Related Work

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.

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 (P. 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).

Task-based evaluation of exploratory search systems using controlled experiments has been effective for showing subjective satisfaction differences between systems, but less effective at showing objective differences in task performance, particularly in task outcomes. (Kabel et al., 2004; Yee et al., 2003). Evaluations have assessed and rated the quality of a task outcome to generate quantitative measures on a lesson plan creation task (Kabel et al.) or measured incidental learning that occurred during a search session (P. Pirolli et al., 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 (Toms, Freund, Kopak, & Bartlett, 2003; Yee et al., 2003). Controlled experiments and in-depth case studies are two approaches to evaluation of exploratory search systems.

Consider:
(G. Marchionini, Plaisant, & Komlodi, 1998)
(Spink, 2002)
(Vakkari, 1999)
(Jansen, Spink, & Saracevic, 2000)
(Wang, Hawk, & Tenopir, 2000)
(Käki, 2005)
(Kules & Shneiderman, 2004, submitted)
(Johnson, Griffiths, & Hartley, 2003)
(Garcia & Sicilia, 2003)

(Kules, 2006a) This position paper describes two approaches to evaluation of exploratory search systems. A mixed method approach was used to evaluate categorized overviews of search results. The longitudinal approach is proposed to extend the mixed methods.

At least one commercial search engine (Exalead.com) has implemented categorized overviews of web search results using an adaptation of the ODP. However we are not aware of any studies of this approach.

3. Experimental Design

Based on our previous research (Kules & Shneiderman, submitted), we expected to observe quantifiable and significant differences in specific behaviors and preferences. For example, we expected that searchers would explore deeper in their result lists using the categorized overview. These were formulated as hypotheses that were empirically tested. We also anticipated that the interface would prompt additional behavioral changes, but there was no *a priori* list. A qualitative approach used a combination of observation and semi-structured interview questions to identify phenomena not modeled by the research variables. Due to space limitations, this paper focuses on the qualitative results, highlighting selected quantitative findings. The study was initially designed to also investigate the effect of broad and narrow topics, but that aspect was problematic and is not discussed here. Details of the complete study can be found in Kules (2006b).

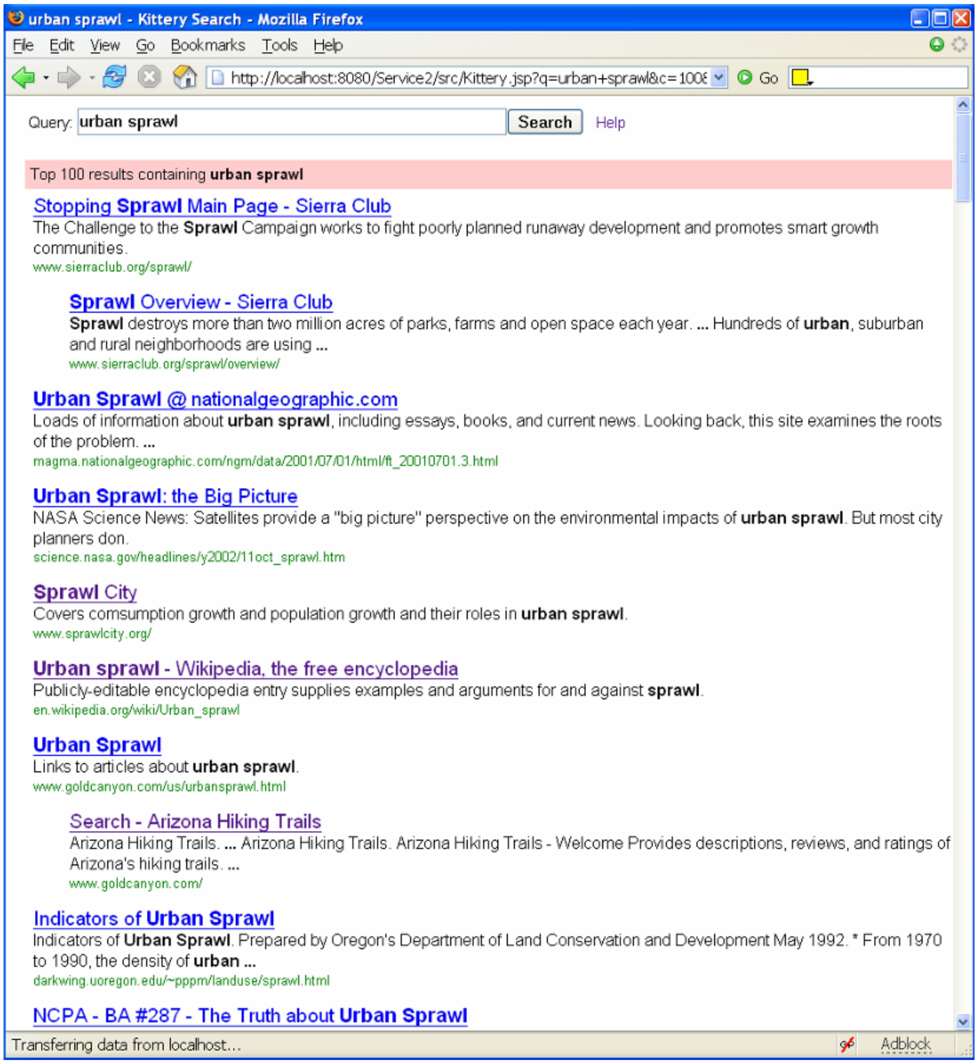
3.1 Experimental conditions

This study compared web search results with and without categorized overviews. The categorized overviews were based on three facets: Topic, Geography and US Government. The topical facet, extracted from the Open Directory Project (ODP) web directory (www.dmoz.org), classified web sites according to the 14 top-level categories shown in Table 1. The geographic facet was extracted from the ODP top-level category, Region. A database of federal government web sites was used to create the Government Agency facet. Web sites were categorized into the top 2 levels of each hierarchy. The categorized overviews were thus comprised of three 2-level facets.

Table 1. 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	

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. This study used the SERVICE (SEarch Result Visualization and Interactive Categorized Exploration) search interface (Kules, 2006b). The SERVICE system was designed to be a flexible, extensible architecture and framework for research in categorizing search interfaces. For this study, it was configured to present two interfaces: a baseline interface that displays a typical list of search results, similar to Google (Figure 1), and a categorized overview interface that displays the overview to the left of the result list (Figure 2).


urban sprawl - Kittery Search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Search
Help
Adblock

http://localhost:8080/Service2/src/Kittery.jsp?q=urban+sprawl&c=100&

Query: urban sprawl

Top 100 results containing **urban sprawl**

Sprawl Stopping **Sprawl** Main Page - Sierra Club
The Challenge to the **Sprawl** Campaign works to fight poorly planned runaway development and promotes smart growth communities.
www.sierraclub.org/sprawl/

Sprawl Overview - Sierra Club
Sprawl destroys more than two million acres of parks, farms and open space each year. **...** Hundreds of **urban**, suburban and rural neighborhoods are using **...**
www.sierraclub.org/sprawl/overview/

Urban Sprawl @ nationalgeographic.com
Loads of information about **urban sprawl**, including essays, books, and current news. Looking back, this site examines the roots of the problem. **...**
magma.nationalgeographic.com/ngm/data/2001/07/01/html/ft_20010701.3.html

Urban Sprawl: the Big Picture
NASA Science News: Satellites provide a "big picture" perspective on the environmental impacts of **urban sprawl**. But most city planners don.
science.nasa.gov/headlines/y2002/11oct_sprawl.htm

Sprawl City
Covers comsumption growth and population growth and their roles in **urban sprawl**.
www.sprawlcity.org/

Urban sprawl - Wikipedia, the free encyclopedia
Publicly-editable encyclopedia entry supplies examples and arguments for and against **sprawl**.
en.wikipedia.org/wiki/Urban_sprawl

Urban Sprawl
Links to articles about **urban sprawl**.
www.goldcanyon.com/us/urbansprawl.html

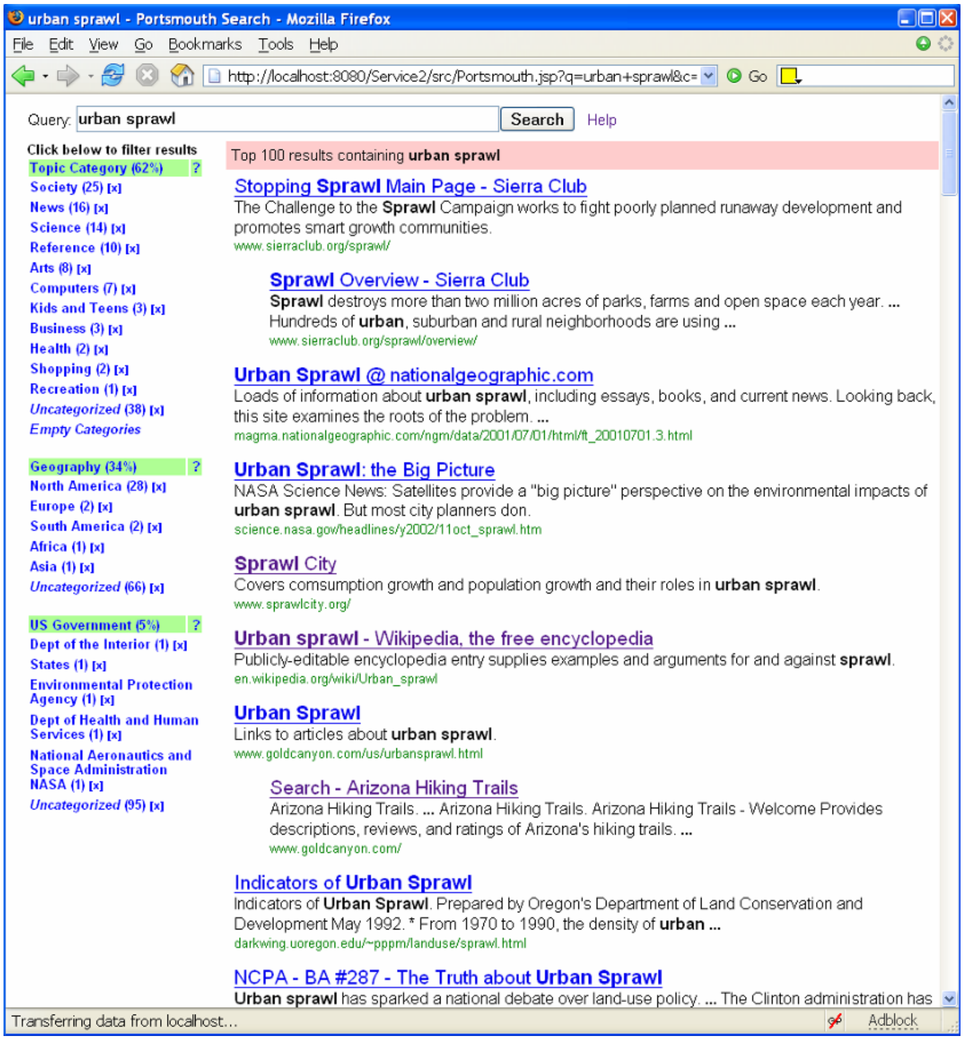
Search - Arizona Hiking Trails
Arizona Hiking Trails. **...** Arizona Hiking Trails. Arizona Hiking Trails - Welcome Provides descriptions, reviews, and ratings of Arizona's hiking trails. **...**
www.goldcanyon.com/

Indicators of **Urban Sprawl**
Indicators of **Urban Sprawl**. Prepared by Oregon's Department of Land Conservation and Development May 1992. * From 1970 to 1990, the density of **urban ...**
darkwing.uoregon.edu/~pppm/landuse/sprawl.html

NCPA - BA #287 - The Truth about **Urban Sprawl**

Transferring data from localhost...

Figure 1. 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.



Top 100 results containing **urban sprawl**

Stopping **Sprawl** Main Page - Sierra Club
The Challenge to the **Sprawl** Campaign works to fight poorly planned runaway development and promotes smart growth communities.
www.sierraclub.org/sprawl/

Sprawl Overview - Sierra Club
Sprawl destroys more than two million acres of parks, farms and open space each year. **...** Hundreds of **urban**, suburban and rural neighborhoods are using **...**
www.sierraclub.org/sprawl/overview/

Urban Sprawl @ nationalgeographic.com
Loads of information about **urban sprawl**, including essays, books, and current news. Looking back, this site examines the roots of the problem. **...**
magma.nationalgeographic.com/ngm/data/2001/07/01/html/ft_20010701.3.html

Urban Sprawl: the Big Picture
NASA Science News: Satellites provide a "big picture" perspective on the environmental impacts of **urban sprawl**. But most city planners don.
science.nasa.gov/headlines/y2002/11oct_sprawl.htm

Sprawl City
Covers comsumption growth and population growth and their roles in **urban sprawl**.
www.sprawlcity.org/

Urban sprawl - Wikipedia, the free encyclopedia
Publicly-editable encyclopedia entry supplies examples and arguments for and against **sprawl**.
en.wikipedia.org/wiki/Urban_sprawl

Urban Sprawl
Links to articles about **urban sprawl**.
www.goldcanyon.com/us/urbansprawl.html

Search - Arizona Hiking Trails
Arizona Hiking Trails. **...** Arizona Hiking Trails. Arizona Hiking Trails - Welcome Provides descriptions, reviews, and ratings of Arizona's hiking trails. **...**
www.goldcanyon.com/

Indicators of **Urban Sprawl**
Indicators of **Urban Sprawl**. Prepared by Oregon's Department of Land Conservation and Development May 1992. * From 1970 to 1990, the density of **urban ...**
darkwing.uoregon.edu/~pppm/landuse/sprawl.html

NCPA - BA #287 - The Truth about **Urban Sprawl**
Urban sprawl has sparked a national debate over land-use policy. **...** The Clinton administration has

Figure 2. 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.

3.2 Scenario and task design

A high-level scenario was constructed around an exploratory search task for journalists. The task involved generating ideas for newspaper articles. Journalists' information needs are often uncertain, and can change 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, where were reviewed by a journalism professor to ensure that they 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.*

The four topics used for the study were:

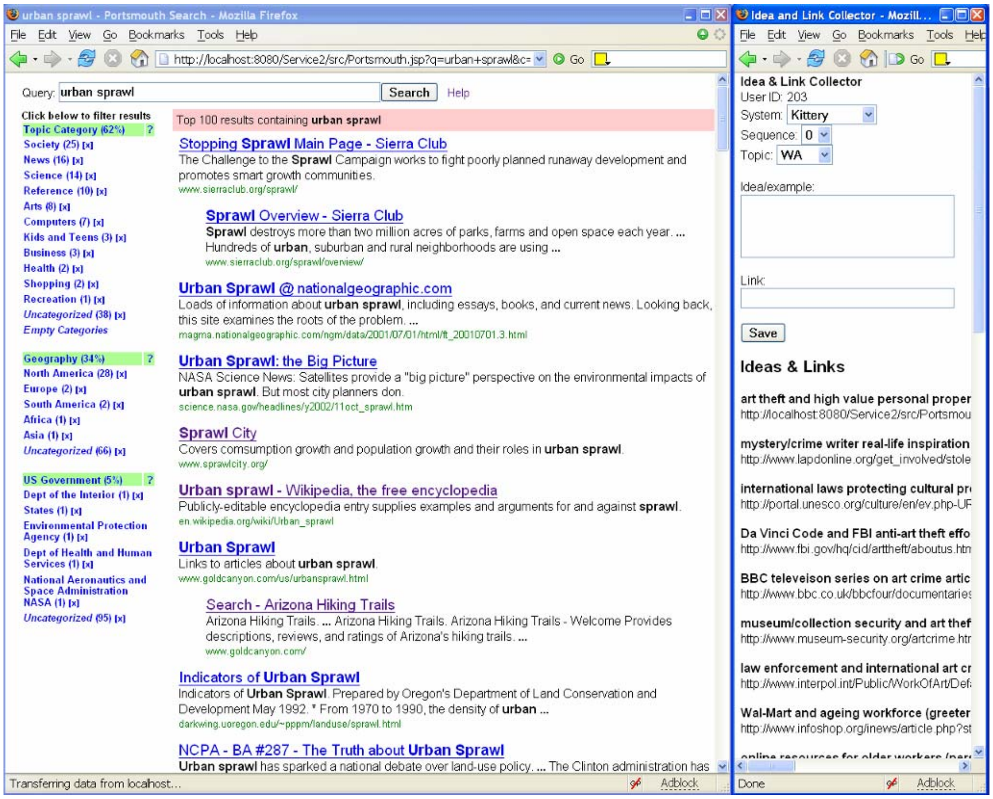
- Workplace allergies (WA)
- The aging workforce (AW)
- Human smuggling (HS)
- International art crime (IAC)

3.3 Participants

Twenty-four experienced web searchers (5 male, 19 female) were recruited primarily from the University of Maryland's College of Journalism and paid $30 for their participation. 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. All reported at least three years of search experience, and all but two reported searching at least once per day. All used the Google search engine.

3.4 Materials

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 3). 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.


urban sprawl - Portsmouth Search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Search
Help
Idea and Link Collector - Mozill...
File Edit View Go Bookmarks Tools Help
Go
Save
Adblock

http://localhost:8080/Service2/src/Portsmouth.jsp?q=urban+sprawl&c=

Query: urban sprawl

Click below to filter results

Topic Category (62%) ?
Society (25) [x]
News (16) [x]
Science (14) [x]
Reference (10) [x]
Arts (8) [x]
Computers (7) [x]
Kids and Teens (3) [x]
Business (3) [x]
Health (2) [x]
Shopping (2) [x]
Recreation (1) [x]
Uncategorized (38) [x]
Empty Categories

Geography (34%) ?
North America (28) [x]
Europe (2) [x]
South America (2) [x]
Africa (1) [x]
Asia (1) [x]
Uncategorized (66) [x]

US Government (5%) ?
Dept of the Interior (1) [x]
States (1) [x]
Environmental Protection Agency (1) [x]
Dept of Health and Human Services (1) [x]
National Aeronautics and Space Administration NASA (1) [x]
Uncategorized (95) [x]

Top 100 results containing **urban sprawl**

Stopping **Sprawl** Main Page - Sierra Club
The Challenge to the **Sprawl** Campaign works to fight poorly planned runaway development and promotes smart growth communities.
www.sierraclub.org/sprawl/

Sprawl Overview - Sierra Club
Sprawl destroys more than two million acres of parks, farms and open space each year. ... Hundreds of **urban**, suburban and rural neighborhoods are using ...
www.sierraclub.org/sprawl/overview/

Urban Sprawl @ nationalgeographic.com
Loads of information about **urban sprawl**, including essays, books, and current news. Looking back, this site examines the roots of the problem. ...
magma.nationalgeographic.com/ngm/data/2001/07/01/html/ft_20010701.3.html

Urban Sprawl: the Big Picture
NASA Science News: Satellites provide a "big picture" perspective on the environmental impacts of **urban sprawl**. But most city planners don.
science.nasa.gov/headlines/y2002/11oct_sprawl.htm

Sprawl City
Covers comsumption growth and population growth and their roles in **urban sprawl**.
www.sprawlcity.org/

Urban sprawl - Wikipedia, the free encyclopedia
Publicly-editable encyclopedia entry supplies examples and arguments for and against **sprawl**.
en.wikipedia.org/wiki/Urban_sprawl

Urban Sprawl
Links to articles about **urban sprawl**.
www.goldcanyon.com/us/urbansprawl.html

Search - Arizona Hiking Trails
Arizona Hiking Trails. ... Arizona Hiking Trails. Arizona Hiking Trails - Welcome Provides descriptions, reviews, and ratings of Arizona's hiking trails. ...
www.goldcanyon.com/

Indicators of **Urban Sprawl**
Indicators of **Urban Sprawl**. Prepared by Oregon's Department of Land Conservation and Development May 1992. * From 1970 to 1990, the density of **urban** ...
darkwing.uoregon.edu/~pppm/landuse/sprawl.html

NCPA - BA #287 - The Truth about **Urban Sprawl**
Urban sprawl has sparked a national debate over land-use policy. ... The Clinton administration has

Transferring data from localhost...

Idea & Link Collector
User ID: 203
System: Kittery
Sequence: 0
Topic: WA

Idea/example:

Link:

Ideas & Links

art theft and high value personal proper
http://localhost:8080/Service2/src/Portsmou

mystery/crime writer real-life inspiration
http://www.lapdonline.org/get_involved/stole

international laws protecting cultural pr
http://portal.unesco.org/culture/en/ev.php-UF

Da Vinci Code and FBI anti-art theft effo
http://www.fbi.gov/hq/cid/arttheft/aboutus.htm

BBC televeison series on art crime artic
http://www.bbc.co.uk/bbcfour/documentaries

museum/collection security and art thef
http://www.museum-security.org/artcrime.htr

law enforcement and international art cr
http://www.interpol.int/Public/WorkOfArt/Def

Wal-Mart and ageing workforce (greeter
http://www.infoshop.org/inews/article.php?st

Done

Figure 3. The interface used by participants was comprised of the system under test (left) and the Collector form (right).

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. Three online questionnaires were used during the experimental sessions (see Appendix). 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. The exit interview questions were read to the participants from a paper form.

Participants used an IBM T42p laptop with a 15 inch display, 1 GB of RAM, a 1.8 GHz Intel Pentium M processor running Windows XP Professional, an external keyboard and mouse. Camtasia Studio 3 was used to capture screen video and audio, with a desktop microphone. The SERVICE web search system was configured to log all pages visited, plus detailed data on category and result list clicks, mouse movements, and scrolling. A Tomcat 5 server running on the laptop hosted the search application and the Collector application. Participants used the Mozilla Firefox browser (v. 1.0.7). The laptop was connected to the Internet via the campus T3 connection.

3.5 Procedure

The experiment sessions were individually conducted in an office on the University of Maryland campus. As participants arrived, they were welcomed and provided a short introduction to the study. 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.” A training checklist ensured that they used the basic system features on their own or with prompting. During both the training and measured tasks, they were encouraged to think out loud, using a think-aloud protocol (Ericsson & Simon, 1984).

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 video for 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.

Materials and procedures were pilot tested with 12 participants. Based on the pilot test, 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.

3.6 Analysis methodology

The quantitative data was analyzed using 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. Where the raw data did not follow a normal distribution, it was transformed using a logarithmic transform, an accepted technique for handling non-normal distributions (Jaccard, 1983). For all significant ANOVA results, the normal Quantile-Quantile (Q-Q) Plots were examined to confirm that the residuals were distributed normally.

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 us 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.

Three open-ended questions from the exit interview were chosen for analysis. These questions related directly to the research questions, and we 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:

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

In question two, 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. Each response was reviewed by one researcher, the lead author, 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. 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.

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. The use of a single researcher is an acknowledged limitation of this study; however, the analysis and results was peer-reviewed.

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.

4. 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 strengths and weaknesses of the categorized overviews.

4.1 Summary of quantitative results

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

Searchers viewed (clicked on) a total of 924 pages from the search results. 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. 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 4 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.

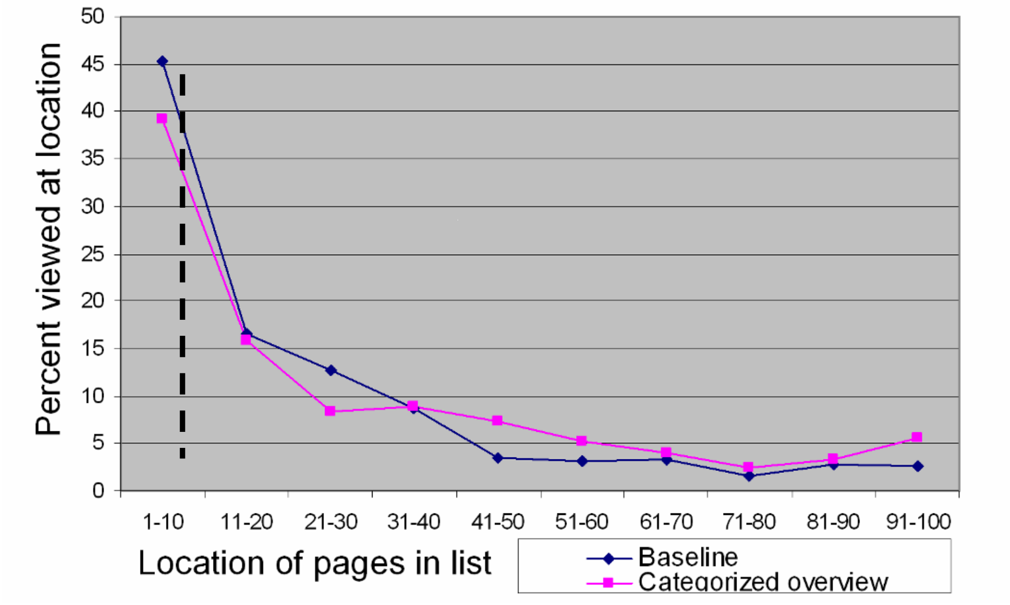


Figure 4. Percent of pages viewed by original location of page within search results, for each system. The interface displayed approximately 10 results per screen. The dashed line shows the initial screen break.

4.1.2 Proportion of pages collected from categorized facets

Not all pages were categorized in the available facets, and we were interested in whether searchers were more likely to collect categorized pages when using the categorized overview. Searchers collected a total of 679 pages. 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.

Table 2. Percent of collected pagesthat had been categorized, by System*.

System	**Percent Categorized**
Baseline	75.4 %
Categorized overview	82.7 %

4.1.3 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 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 mean (median) number of queries per search was 3.0 (2) for the categorized overview system and 3.5 (3) for the baseline.

4.1.4 Perceived organization of search results

Subjects were asked to rate agreement with the statement, “The system helped me organize my search results,” (1 = strongly disagree, 9 = strongly agree). 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 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.

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

Subjects were asked to rate agreement with the statement, “The system helped me assess the results of my queries to decide what to do next,” (1 = strongly disagree, 9 = strongly agree). 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 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.

4.1.6 Adjectives to describe system

Subjects were asked to rate eight aspects of the systems. The 2 (system) x 4 (topic) factorial analysis for each of the eight system adjectives (semantic differentials) identified three measures that 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. No significant differences by topic were identified.

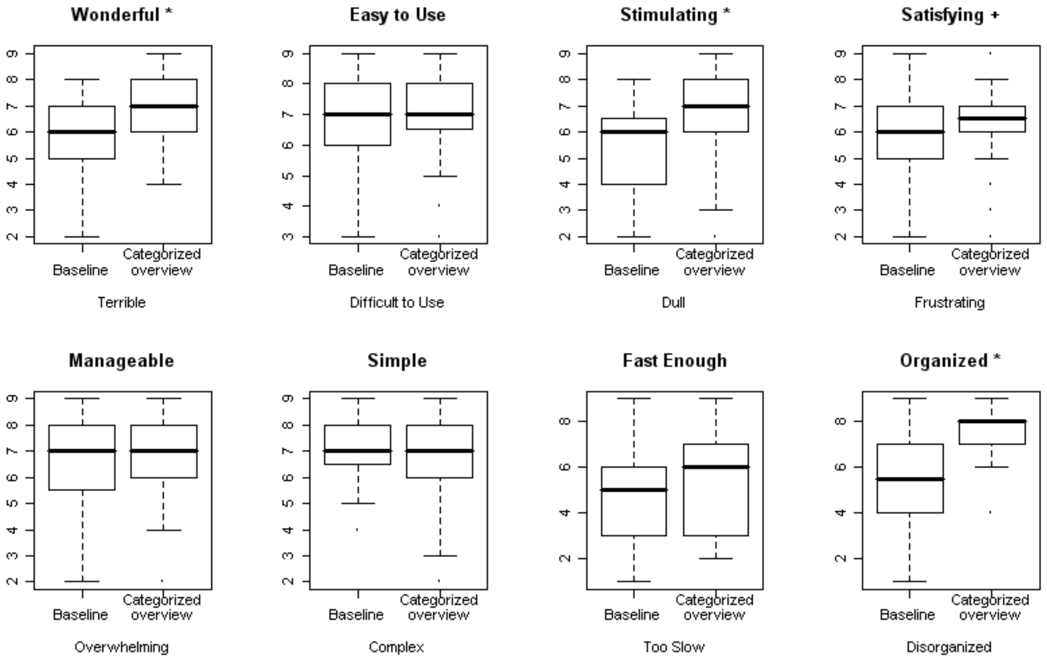


Figure 5. Adjectives by System.

4.1.7 Idea quality

The quality of the generated ideas was assessed blind by a single researcher. 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 and potential impact. Two passes were made through the ideas for each topic to gain familiarity the ideas before assigning a final quality rating.

Searchers generated a total of 679 ideas. Idea quality was generally low, perhaps 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, $p<0.10$.

4.2 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 sub-sections based three of the derived code groups: Behavioral differences, cognitive and affective impacts, and judgments of outcome. These sub-sections include observations, quotes and comparisons between participants to highlight differences that the quantitative results do not capture.

4.2.1 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.

> *I loved it. I was in love with that. I wish Google had that…With 3 clicks you have 5 pieces of information. (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. 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)*

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.

Table 3. 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

4.2.2 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 occurred 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, however, about half of the respondents indicated that the problems were minor or not a hindrance. 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)*

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)*

One person was concerned that she might have missed useful pages in categories she did not explore. Four people commented that the categories helped generate ideas. Two people commented that they used the overview when they were stuck.

Table 4. 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

4.2.3 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.

> *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 5. 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

4.2.4 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 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 6. Mentions of geographic or government category use.

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

5. Discussion

5.1 Differences in search behavior

The quantitative and qualitative data indicate that the overviews did change searcher behavior in several ways. The quantitative data showed that participants explored significantly more deeply within the result list. This is consistent with previous studies (Käki, 2005). With the categorized overview, participants did collect slightly more pages that were categorized (i.e., they collected fewer uncategorized pages). 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 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.

The participants issued fewer queries with the categorized overview and commented on this during the interview. The categorized overview appeared to provide cues, similar to the notion of "information scent" (P. Pirolli & Card, 1995; P. L. Pirolli & Card, 1999), that induced participants to click on categories instead of refining their query.

5.2 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 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.

When page categories did not match searchers' expectations, they experienced mild frustration, confusion or doubt. Three factors reduced the categorization accuracy from the participants' perspective: categories that represent different kinds of relationships within the same facet, ambiguous categories, problems with the hierarchical structure of categories, and.

Pages from the British Broadcasting Corporation (BBC) were categorized under /Category/Arts/Television, which is closer to encoding an *is-a* relationship than an *about* 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 (two levels) in the categorized overview. The ODP-assigned categories were frequently four or more levels deep. For example, the BBC web pages were assigned to categories in the topical and geographic facets. Truncating the categories to two levels, as shown in Table 7, removed useful contextual information.

Table 7. 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**
Topic 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
Geographic 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 7 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.

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 may 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.

Fortunately, participants indicated that these problems were minor. Their comments about 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 subjective measures lend support to this interpretation of the experimental results. Participants agreed that the categorized overview organized the results well and helped them assess their results and decide what to do next. Participants also found the categorized overview more generally appealing ("wonderful") and stimulating. The satisfaction ratings, which favored the categorized overview, were marginally significant.

These results also suggest that the categorized overviews were no more difficult to use than the baseline in general. There was no significant difference in the overwhelming/manageable or complex/simple measures. 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, and a few participants did comment on complexity. But for this task there were clear benefits to most participants. 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.

5.3 Differences in search tactics

Participants commented on many interesting effects that the categorized overviews had on their tactics. 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. During their first categorized overview search they often 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. 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.

The seven tactics evidenced in this study are shown in Table 8. Studies have found that most searchers do not examine more than the first page of search results (Jansen et al., 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. The actions enabled by categorized overviews can lead to altered search tactics because they change the information available and the range of possible interactions. 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.

Table 8. 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.

5.4 Effect on quality of search outcome

Most participants appreciated and used the overview, but there were no observed differences in the quality of the story ideas they generated with the overviews. This could indicate that the task was less dependent on gaining an overview than originally anticipated. When the overview wasn’t available, scanning the result lists and reformulating queries were reasonably effective tactics for generating article ideas. Participant 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. 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 exploring the categories instead of analyzing the results to identify new concepts and terms to refine his query. Training could help users decide when to use the categorized overview and when to submit a new query.

5.5 Limitations

This study was subject to several important limitations. The participants (N=24) were primarily journalism students, so the scenario and task was appropriate for them, but they might not be representative of the needs of other exploratory searchers. The participants were all experienced searchers, and their experience will certainly differ from novice searchers. Only one scenario and task type was evaluated. Other exploratory search tasks may benefit more or less from the categorized overview. The amount of time available was substantial for each session, but not long enough for searchers to completely adapt their tactics. 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.

The study was limited by several factors related to the categories. Only three facets were used: topic, geography, and US government. Other facets could have been chosen, and this could have yielded different quantitative and qualitative results. The modest proportion of pages that were categorized (40-80%) was a limitation of the study. 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 qualitative analysis was limited in several important ways. The research was conducted in a laboratory setting rather than the participants' own workplaces. Participants did, however, show an awareness of these differences, and they commented on the essential elements of the task that were common between the research setting and their workplace. More importantly, a single researcher analyzed and interpreted the raw data. The study does make modest use of triangulation with the quantitative data, and it provides direct quotes to support interpretations. The interpretations were closely tied to the raw data, often using the same language that participants used. Additional studies are needed to confirm these results.

6. Design guidelines for categorized overviews

The results of this study have been used to refine a set of design guidelines that we are developing for categorized overview interfaces. They are particularly intended to support exploratory search. User interface design guidelines 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, and provide practical advice to interface designers. The design guidelines suggested or refined by this study are:

1. Provide overviews of large sets of results
2. Organize overviews around meaningful categories
3. Clarify and visualize category structure
4. Tightly couple category labels to result list
5. Ensure that the full category information is available
6. Support multiple types of categories and visual presentations
7. Use separate facets for each type of category
8. Arrange text for scanning/skimming

6.1 Provide overviews of large sets of results

During exploratory search, hundreds or thousands of results are potentially relevant. Shneiderman's 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. The ideal number will certainly depend on many factors, including the task domain, topic, the quality and quantity of documents, and search engine capabilities. The fact that many of the pages viewed in this study were ranked in the range of 50th-100th suggests that at least 100 results can be useful. The results of this study show that this can be done without introducing overwhelming complexity.

6.2 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. Meaningful categories support learning, reflection, discovery, and information retrieval (Kwasnik, ; Soergel). The results of this study suggested that categorized overviews based on topic, geography, and the US government reduced cognitive workload and supported beneficial search tactics. Categories based on document format, language, or Domain Name Service (DNS) domain may be useful (Kules, Kustanowitz, & Shneiderman, 2006). Numeric attributes such as date or size can be grouped into meaningful categories. Even abstract or computed attributes such as a journal impact factor (Garfield, 2005) can form the basis of meaningful, albeit controversial or limited, categories.

This study also suggests that stable categories will allow searchers to reuse category knowledge on subsequent searches. Dynamic categories, such as those generated by automated clustering techniques, change with each query. Thus the learning benefits of stable categories may accrue less, but they may provide other benefits (Kules & Shneiderman, submitted).

6.3 Clarify and visualize category structure

If the categories are drawn from a classification, taxonomy, or ontology, the structure should be made visible. 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.

This study suggests that practitioners should review at least the top two levels of a hierarchy, considering whether they need to be adjusted to provide the clearest overview. The study results show that 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.

6.4 Tightly couple category labels to result list

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 to synchronize two views of the results, an overview and a detailed list (Klein, Reiterer, Müller, & Limbach, 2003). Judicious use of this technique can support richer interactions between category information and individual results. The SERVICE system provides two examples of tight coupling: The category labels can be clicked to narrow or broaden the result list; and, pausing the pointer over a result in the list can highlight all the categories (in the overview) that contain the result. 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.

When this capability is implemented, it is important to provide clear feedback indicating which categories are currently applied. Observations from this study suggested that participants occasionally forgot or overlooked the fact that they were viewing a subset of their original query.

6.5 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 this study, participants wondered aloud what specific category results were in. As discussed in section 5.2, they were occasionally frustrated because only the top two levels of the category were visible in the overview. Providing the full category label could alleviate this problem. 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. A small hyperlink in each result may be an appropriate design compromise, although this was not implemented or evaluated. These alternatives should be investigated in future studies.

6.6 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 the Clusty and Exalead search engines are examples of search result interfaces that permit users to reorganize results using alternate sets of categories. During previous 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 (Kules & Shneiderman, submitted). 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 enable an experienced searcher to customize the overview and share it with others. Supporting multiple classifications and multiple visual presentations may enable users to view and explore search results from the perspectives most appropriate to their needs.

6.7 Use separate facets for each type of category

When a rich set of categories encodes multiple types of relationships, presenting them as separate visual 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 classification organizes a domain into orthogonal sets of categories, which are ideally homogeneous, mutually exclusive, and represent a single characteristic of division (Vickery, 1960). It has been used to organize catalogs, classifications, and thesauri (Soergel, 1974; Vickery, 1960), information spaces on the Web (Louie, Maddox, & Washington, 2003), and non-web search interfaces (Yee et al., 2003). The importance of this principle was clarified during the development of the SERVICE system. During informal user tests, searchers experienced confusion when topical and geographic categories were used in the same facet. Separating geographic categories from topical categories in the final interface helped reduce this problem in this study. 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.

6.8 Arrange text for scanning/skimming

At a perceptual level, users of search results attempt to rapidly ingest large amounts of text. We observed searchers scanning category labels, titles, URLs, 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. 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).

6.9 Summary

These eight design guidelines for categorized overviews have been suggested or refined 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. These guidelines are not exhaustive or comprehensive. Evaluations and analysis of other exploratory search interfaces will certainly suggest additional guidelines.

7. Conclusions

As discussed in Section 5, this study suggests answers to the three research questions. It suggests important cognitive and tactical benefits of categorized overviews. Searchers explored more deeply in their results. They agreed that the categorized overviews helped them organize, explore and assess their results without being appreciably more complex than typical Google-like interfaces. The study also identifies limitations of categorized overviews, and raises a cautionary question about possible negative impacts on the quality of search results. The study identifies seven tactics that searchers began to adopt when the categorized overview was available.

These findings must be explored and validated with additional research. The study highlighted an important consideration for future evaluation of exploratory search systems: It takes time for searchers to reflect on their searches and refine their tactics. Future research should consider using an in-depth, longitudinal case study approach to address this. Longitudinal studies have been used to examine changes in tactics and query terms in relation to changes in searchers' information problem stage while developing a research proposal (Vakkari, 2000). 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 interactive categorized overviews are available. They present the opportunity to observe changes as searchers become familiar with an exploratory search system and tactics mature. They also present challenges, because search is often a means to an end, and individual searches may be initiated to satisfy a higher level task. The search sessions may not be readily organized into blocks of time that can be scheduled with a researcher.

Finally, the study suggested or refined a set of eight design guidelines for categorized overview interfaces to support exploratory search. We believe these guidelines 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. The guidelines must be tested, refined, and extended by additional studies. They will provide a touchstone for future researchers who wish to investigate the challenges of exploratory search.

8. Acknowledgements

This research was partially supported by an AOL Fellowship in Human-Computer Interaction and National Science Foundation Digital Government Initiative grant (EIA 0129978) “Towards a Statistical Knowledge Network.”

9. References

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.

Aula, A. (2004). Enhancing the readability of search result summaries. In Proceedings Volume 2 of the Conference HCI 2004: Design for Life, Leeds, UK.

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

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

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

Dumais, S., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. Proceedings of the SIGCHI conference on Human factors in computing systems, 277 - 284.

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

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

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

Garfield, E. (2005). In 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.

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.

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.

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

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.

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

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.

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.

Johnson, F.C., Griffiths, J.R., & Hartley, R.J. (2003). Task dimensions of user evaluations of information retrieval systems. Information Research, 8(4), paper no. 157.

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.

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.

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.

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.

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.

Kules, B. (2006a). Methods for evaluating changes in search tactics induced by exploratory search systems, ACM SIGIR 2006 Workshop on Evaluating Exploratory Search Systems. Seattle, WA.

Kules, B. (2006b). Supporting Exploratory Web Search with Meaningful and Stable Categorized Overviews (Unpublished doctoral dissertation): University of Maryland, College Park.

Kules, B., Kustanowitz, J., & Shneiderman, B. (2006). Categorizing web search results into meaningful and stable categories using Fast-Feature techniques. In Proceedings of the Sixth ACM/IEEE-CS Joint Conference on Digital Libraries, Chapel Hill, NC. New York: ACM Press.

Kules, B., & Shneiderman, B. (2004). In Categorized graphical overviews for web search results: An exploratory study using U.S. government agencies as a meaningful and stable structure (pp. 20-24). Paper presented at the Third Annual Workshop on HCI Research in MIS, Washington, DC.

Kules, B., & Shneiderman, B. (submitted). Using meaningful and stable categories to support exploratory web search: Two formative studies.

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

Louie, A.J., Maddox, E.L., & Washington, W. (2003). In Using faceted classification to provide structure for information architecture (pp. 203-212). Paper presented at the The 62nd ASIS Annual Meeting, Washington, D.C. ASIS.

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

Marchionini, G., Plaisant, C., & Komlodi, A. (1998). Interfaces and tools for the Library of Congress National Digital Library Program. Information Processing & Management, 34(5), 535-555.

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.

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.

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.

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

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.

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.

Shneiderman, B., Byrd, D., & Croft, W.B. (1997). Clarifying search: A user-interface framework for text searches. D-Lib Magazine.

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.

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.

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

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.

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

Soergel, D. (1999). The rise of ontologies or the reinvention of classification. Journal of the American Society for Information Science and Technology, 50(12), 1119-1120.

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. Information Processing & Management, 38(3), 401-426.

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.

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.

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.

Toms, E.G., Freund, L., Kopak, R., & Bartlett, J.C. (2003). The effect of task domain on search. Proceedings of the 2003 conference of the Centre for Advanced Studies on Collaborative research, 303-312.

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

Vakkari, P. (1999). Task complexity, problem structure and information actions: Integrating studies on information seeking and retrieval. Information Processing & Management, 35(6), 819-837.

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.

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

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.

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

White, R.W., Kules, B., Drucker, S.M., & schraefel, m.c. (2006). Supporting exploratory search. Communications of the ACM, 49(4), 36-39.

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.

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.