Browsing and Searching in a Faceted Information Space: A Naturalistic Study of PubMed Users' Interaction With a Display Tool

Muh-Chyun Tang
National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 106, Taiwan R.O.C. E-mail: mctang@ntu.edu.tw

The study adopts a naturalistic approach to investigate users' interaction with a browsable MeSH (medical subject headings) display designed to facilitate query construction for the PubMed bibliographic database. The purpose of the study is twofold: first, to test the usefulness of a browsable interface utilizing the principle of faceted classification; and second, to investigate users' preferred query submission methods in different problematic situations.

An interface that incorporated multiple query submission methods—the conventional single-line query box as well as methods associated the faceted classification display was constructed. Participants' interactions with the interface were monitored remotely over a period of 10 weeks; information about their problematic situations and information retrieval behaviors were also collected during this time. The traditional controlled experiment was not adequate in answering the author's research questions; hence, the author provides his rationale for a naturalistic approach.

The study's findings show that there is indeed a selective compatibility between query submission methods provided by the MeSH display and users' problematic situations. The query submission methods associated with the display were found to be the preferred search tools when users' information needs were vague and the search topics unfamiliar.

The findings support the theoretical proposition that users engaging in an information retrieval process with a variety of problematic situations need different approaches. The author argues that rather than treat the information retrieval system as a general purpose tool, more attention should be given to the interaction between the functionality of the tool and the characteristics of users' problematic situations.

Introduction

The potential of using a faceted classification in organizing information in a networked environment has long been recognized (Anderson, 1990, 2002; Bates, 2002; Broughton, 2002; Ellis & Vasconcelos, 2000). It also has been suggested that a thorough faceted analysis applied in query formulation is conducive to favorable results (Drabenstott, 2001; Soergel, 1985). There has been empirical evidence suggesting better performance by faceted queries. For example, Kekäläinen and Järvelin (1998) showed that structured queries that represented multiple facets performed better than unstructured ones in a best match setting. Vakkari, Jones, MacFarlane, and Sormunen (2004) also found that the more the users' queries covered the facets of the search topic, the better the search results. Just as faceted analysis has been used to remind the indexer of the different aspects by which a document can be represented, a faceted display of the classification might also encourage users to articulate different aspects of their information needs. Yet it remains an empirical question how the users might interact with a faceted classification. Here this question is approached from the theoretical standpoint that different problematic situations demand different types of need representation methods (Belkin, Oddy, & Brooks, 1982). Framed this way, the question becomes "What types of problematic situations can the faceted classification best support?"

Theoretical Standpoint

The act of representation, that is, utilizing a representational device to stand for the represented objects in some aspects or capacity, inevitably entails the loss of information. Each representation device is enabling in some aspects and constraining in others. As Kwasnik (1999) demonstrated, the approach to information representation has significant consequences on knowledge access and discovery. Yet as Soergel (1994) pointed out, inferring retrieval performance solely from indexing characteristics can be misleading as it fails to take into consideration the other end of the search process, which is the query formulation. He further argued that one of the important factors often overlooked in an artificial test situation is "the adaptation of the query formulation to the characteristics of the retrieval system."

Received January 28, 2006; revised January 5, 2007, February 13, 2007; accepted February 14, 2007

 ● Published Online 12 September 2007 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/asi.20689

On the other hand, previous studies indicate that users often engage in different search strategies when interacting with an information retrieval (IR) system, depending on various contextual factors such as their problematic situations and work constraints. It follows that there is no single mode of representation that can optimally meet all users' purposes. The ideal representation method of users' needs should be a function of their problematic situations and task constraints at a certain juncture in the process of carrying out his or her work task. The question then arises whether we can identify the compatibility between different representation modes and user's search situations. What is the ideal representation mode for a certain search situation?

Categorization of Users' Problematic Situations

Following Belkin et al. (1982), problematic situations are defined as a cognitive deficiency experienced by a person when attempting to carry out a work task. Kelly and Belkin (2002) adopted four aspects of users' problematic situations in their study of implicit modeling of users' information behaviors: topic familiarity, topic persistence, task endurance, and problem solving stages. There also has been rich evidence that individuals engage in different information seeking activities in different search or research stages (Kuhlthau, 1991; Vakkari, 2001; Vakkari & Hakala, 2000; Vakkari, Pennanen, & Serola, 2003; White, 1975). Marchionini (1995) listed specificity, quantity, and timeliness as three basic dimensions that an information seeking task can be characterized. Within traditional OPAC (online public access catalog) literature, user goals have been roughly categorized as known-item and subject searches. Using grounded theory methodology, Carol Hert (1996) identified four different user goals when accessing OPAC: looking for known entities, looking for unknown entities, looking for information about entities, and unspecified information. In Saracevic, Kantor, Chamis, and Trivison (1988) search questions were classified by the attributes such as domain, clarity, specificity, complexity, and presupposition.

In this study, the participants were asked to characterize their problematic situations before they conducted their searches, resulting in two nominal variables: search goals (i.e., known item search, search for a specific question, and browse without a specific question) and types of searches by topic familiarity and comprehensiveness needed; and four numerical variables: familiarity with the topic, specificity of search, search stage, and thoroughness for the results desired. The characterization here of searchers' problematic situations is not intended to be thorough and definitive, but to highlight the aspects that might interact with the characteristic of the faceted display. We are interested in seeing how these variables might influence users' querying behaviors (i.e., their use of query submission methods).

The Interface

An interface was built for the display of MeSH (Medical Subject Headings) tree on PubMed. PubMed is chosen because it is one of few databases where faceted analysis for indexing is performed. It should be noted that MeSH's top categories are not purely facets as they are not entirely mutually exclusive to one another. It does, however, present a multidimensional structure that serves as a "scaffolding" device for information need representation (Jacob, 2001).

One of the distinct features of the proposed interface is its attempt to incorporate both browsing and searching modes of access. The other prototypical systems that have also attempted to utilize the faceted approach to classification display all rely heavily on browsing and direct manipulation of the classification structure (Allen, 1995; Hearst, 2000; Pollitt, 1998). It is arguable that keyword searching has been the access mode to which most users have become accustomed. Instead of relying solely on browsing, the proposed interface preserves the search mode of access while providing the browsable thesaurus in support of searching.[1]

Methodology

In light of the research questions discussed above, a naturalistic, longitudinal research design is better suited for this study. Yuan (1997) reported a one-year longitudinal study of searching behaviors of a group of law students who began as novices of the IR system. She found that search experience affected several aspects of user behaviors including, most notably, the increase in types of commands and strategies used and the increase in search speeds. The same might be expected in searchers' adoption of a faceted display. One of the major methodological issues to be tackled is the familiarity effect when comparing customary features with a relatively novel one. The user might prefer or perform better with the tool she or he is most familiar with simply out of habit. In their study of the usability of three visualization tools, Heo and Hirtle (2001) concluded that a usability test needs to take into consideration the mental developmental aspect of tool use. They suggest that "further work on usability needs to extend the practice with the tool far beyond a single session, so that the tool and its benefits can be fully developed by the user" (Heo & Hirtle, 2001, p. 674). Such sensitivity to the developmental aspect of tool use is in line with the basic assumption of activity theory (see, e.g., Kaptelinin, 1996) that there is a dialectic relationship between individuals' mental activities and their cultural and technical environments. It is likely that, when facing the faceted display in a real-life situation, the user might not be able to perceive such an advantage without initial exploratory interactions with the tool. Richard Cordes (2001) expressed a similar view when he commented that ". . . this 'learning the capabilities' of a product and how they match user needs is an important component of usability that rarely receives evaluation in laboratory-based usability studies" (p. 416). He called for employing user-defined tasks in place of product-supported tasks to avoid task-selection bias.

[1] A demo of the interface can be accessed at http://morris.lis.ntu.edu.tw/~muhchyun/mapdemo.avi

DOI: 10.1002/asi

We hoped that a longitudinal study would be a sufficient solution to the familiarity effect because it enables us to observe the learning and assimilation of interface processes by the users, which is often unavailable in a strictly controlled environment. An extended period gave us an opportunity of sampling different problematical situations with relatively fewer participants so the relationships between and problematic situations and search strategies could be observed.

We also decided to let the participants search their own questions instead of using assigned tasks, for a couple of reasons. Even though it has been demonstrated that well-crafted, semantically open task descriptions were able to simulate enough participant involvements and yielded results comparable with those of the real tasks (Borlund, 2000; Borlund & Ingwersen, 1997), the user of textual representation is problematic for the purpose of evaluating an interface designed to facilitate query construction. The textual representation of the task would likely interfere, if not offset, the query construction functionality the interface designed to provide. It was feared that the participants would rely heavily on terms in the task narrative, which not only makes the browsable interface unneeded, but also puts the validity of the task categorization in jeopardy. This is especially so for search tasks in the highly specialized domains of biomedicine, for a couple of reasons: First, it would be difficult to create semantically open task descriptions because these fields are highly laden with technical terminology; and second, due to the specialized nature of the domains, it would be difficult to make the tasks engaging enough for the participants who are specialized in the topic. It is likely that the participants might use the superficial tokens in the task narrative for query terms, and take the exact same tokens as the evidence for document relevance.

Therefore, instead of taking a snapshot of a user's interaction with the interface as in most experimental settings, the participants were asked to perform search sessions over an extended period of time in a natural setting.

Research Procedures

To recruit the participants, recruit notes were circulated through institutional mailing lists and bulletin boards in health sciences departments at a large research university. Each participant was asked to conduct at least nine search sessions at times of their own choosing in the time period of 2½ months. After signing up for the study, the participants were directed to an online tutorial Web site where a demonstration of the functions of the faceted display and other search options was given. They were asked to create an account so they could access the interface remotely at their workplaces. Twenty-four PubMed users initially signed up for the study, however, only 19 participants managed to successfully complete the minimum of nine search sessions.

As the experiment began, the participants would log in and conduct the searches through a proxy server at their workplaces instead of coming to the laboratory. Each search session started with filling out a presearch questionnaire, followed by performing searches on the information problems of their own, and concluded with answering a postsearch questionnaire. The participants were given the option of either using the faceted display or using the traditional query box provided by PubMed. All the questionnaires were administered online to make remote monitoring possible.

Results

Participant Characteristics

Among the 19 participants, six participants listed their primary position as students, nine were researchers, one was a faculty member, and three were health care professionals. The faculty and the health care professional groups are likely to be underrepresented because the practical difficulty of recruiting participants from these groups. The fact that the majority (15 out of 19) of the participants are either students or researchers has a significant impact on the types of work tasks undertaken by the searches. The results reported here are most generalizable to the searches motivated by research projects, which account for 86% of the total work tasks surveyed in the presearch questionnaire (the other types of work tasks are teaching and primary care).

The participants were asked to indicate their familiarity with PubMed on a 7-point scale. In general, they were fairly familiar with PubMed, with 10 (over 50%) of them giving a score over 6.

However, the participants' familiarity with MeSH before the study began had been quite low, nor had they used MeSH to assist their searches on a regular basis. Nine of them answered that they had never used MeSH to assist their searches, and only one participant had used MeSH consistently. This seems to confirm the trend that keyword search has become the most dominant query submission method in the era of end-user search, as a previous study also showed that only 10% of MEDLINE users in primary care employed MeSH to refine their search (Cullen, 2002).

Query Submission Methods: Overall and Final

Two hundred one search sessions were completed by the 19 participants, with a total of 633 query submissions, on average, a little more than 3 query submissions occurred per session.

We initially identified five query submission methods based on the combination of search features used and the sources of the terms submitted (Table 1).

As each search session often comes with multiple query submissions, we faced the difficulty of incompatible units of analysis between query submission methods, which occurred multiple times in a search session, and other search session-based variables such as users' "problematic situations" and "satisfaction with the results." This makes it difficult to investigate the relationships between query submission methods and users' "problematic situations." The final submission method was used to create a search session

DOI: 10.1002/asi

TABLE 1. Categorization of query submission methods.

Query submission method		Definition
Not use the classification display	Single line	Subject terms submitted using the conventional query box
	Author/author subject	Terms includes author names and other bibliographic identifiers
Use the classification display	MeSH term select	Terms selected and submitted from MeSH
	Multi-facet input	Terms inputted by users into the multi-facet query boxes
	Combine select & input	Terms come from both MeSH and user inputs

Note. MeSH = Medical subject headings.

level variable of the query submission method. This solution was valid for several reasons. First, in the majority (78%) of the search sessions, the query submission methods were never changed. For those search sessions where query submission did change, it happened mostly when the multifacet interface was used, between the multifacet box input and the combination of selection and input. Second, being at the end of search session, the final query submission was presumably the ideal representation of users' information need.

Therefore in the following analysis, the querying method used in the last submission in a search session will be analyzed along with other search session characteristics (See Table 2 for the distribution of the final query submission methods).

Types of Problematic Situations

By familiarity/comprehensiveness. The participants were asked to characterize their searches in one of the five categories made up along two dimensions: topic familiarity and comprehensiveness needed for the search. As shown in Table 3, over half of the search sessions (67%) were characterized by the participants as searching for background information, with "Searching for background information in a previously unfamiliar area" standing as the single most frequent (39%) type of search problem. "Comprehensive exploration of a previously unfamiliar area" happened most rarely, accounting for only 4% of all search sessions.

TABLE 2. Distribution of final query submission methods.

Query submission method		Frequency	%
Not use the classification display	Single line	67	33
	Author/author subject	27	13
Use the classification display	MeSH term select	27	13
	Multifacet input	49	24
	Combine select & input	31	15

Note. MeSH = Medical subject headings.

TABLE 3. Types of search by comprehensiveness and familiarity.

	Frequency	%
Background information in a familiar area	57	28
Background information in a unfamiliar area	78	39
Comprehensive exposition of a familiar area	37	18
Comprehensive exposition of an unfamiliar area	8	4
One single fact	19	10
System missing	2	1
Total	201	100

By search goals. Again, the participants were asked to characterize their search goals in one of three categories: known item search, search for previous unknown articles for a specific search, and browsing without a specific question in mind. As shown in Table 4, "Look for unknown articles for a specific question" was the most frequently indicated search goal (68%), whereas "Known article search" and "Browsing without a specific question" occurred at about the same frequency, accounting for 16% and 14% , respectively, of all searches.

Query Submission Methods and Types of Search

A two-way contingency table analysis was conducted to evaluate whether there was a relationship between participants' final query submission methods and types of problematic situations. Users' problematic situations are categorized by how familiar they are with the search topic and how comprehensive they need the search to be. They were found to be significantly related, Pearson χ^2 (16, N = 198) = 30.75, $p < .01$, Cramer's $V = .197$, indicating a small to medium effect size.

From Table 5 we can see a rather uneven distribution of query submission methods across different types of search. The users opted for the single-line query box predominantly (42%) when searching for background information in familiar areas. Yet the single-line query box was used rarely when searching for comprehensive exploration. It was used only 17% of the time when a comprehensive exploration for a familiar area was sought and was never used when a comprehensive exploration of a previously unfamiliar area was needed, whereas the three submission methods involving the

TABLE 4. Search goals

	Frequency	%
Known articles	32	16
Unknown articles for a specific question	136	68
Browse without a specific question	29	14
System missing	4	2
Total	201	100

DOI: 10.1002/asi

TABLE 5. Cross-tab between query submission methods and search situations. (The two most frequently used methods for each type of search problems are singled out.)

	Submission				
	Not use the display		Use the classification display		
Search types	Single line	Author/author subject	MeSH select	Multifacet input	Select & input
Background familiar	1st (42%)			2nd (28%)	
Background unfamiliar	1st (39%)		2nd (21%)		
Comprehensive familiar	(17%)			1st (31%)	2nd (28%)
Comprehensive unfamiliar	0%		2nd (25%)	1st (50%)	(13%)
Single fact	1st (32%)	1st (32%)	0%		

Note. MeSH = Medical subject headings.

use of the classification display (MeSH selection, multifacet query box input, and the combination of selection and input) account for 88% of all submissions in this category.

It seems that comprehensiveness needed for the search influenced whether the participants adopted the classification display. As shown in Table 6, the participants were more likely (70% of the time) to use the three submission methods associated with the display when conducting a comprehensive search.

On the other hand, in search situations where a quick answer to a question was sought, the participants avoided the querying methods that involve thesaurus browsing, which demands more time and effort (See Table 5). The MeSH selection was never used when a single fact was sought, whereas a single-line query box was used most often (32% using subject search, another 32% using queries with at least some bibliographic elements), followed by multifaceted keyword input (26%).

Query Submission Methods and Search Goals

Similarly, a two-way contingency table analysis was conducted to evaluate whether there is a relationship between types of query submission methods and types of search goals. Again, they were found to be significantly related, Pearson $\chi^2\,(8, N = 196) = 43.63\; p < .01$, Cramer's $V = .334$, indicating a medium to large effect size.

From Table 7, we can see that the faceted display was seldom used in known items search; author search and author combined with subject search account for 69% of all submissions in this category. MeSH term selection was used most frequently when browsing without a specific question in mind (35%). It is also interesting to see that single-line query box and multifacet query box input were used rather evenly across three types of search goals, both of which do not involve browsing the thesaurus, which suggest a wide applicability of these two submission methods.

TABLE 6. Comprehensiveness and the use of the classification display.

	Submission	
Search types	Not use the display	Use the display
Background information 135 (100%)	68 (50%)	67 (50%)
Comprehensive exposition 44 (100%)	13 (30%)	31 (70%)

The data collected from the exit interview also corroborates the findings cited above; most of the participants considered the browsable thesaurus most useful when they were unfamiliar with the topics and had only a very general idea about what to look for. When they had a specific topic to search for they did not find the MeSH selection method particularly useful. The most frequently cited reason for not using the faceted display was that the users already came equipped with their own specific query terms.

In summary, the results indicate that users' problematic situations, conceptualized either by search goals and types of search, did influence the sources of query terms and submission methods.

Other Problematic Situation Variables

Other than the aforementioned two nominal variables, four numeric variables were also created to characterize the participants' problematic situations, along four dimensions: topic familiarity, self-assessed search stage, thoroughness of the results desired, and the specificity of the question on 7-point scale (Table 8 presents the mean and standard deviation for these variables).

Table 9 presents the zero-order correlation coefficients for all pairs of the four variables.

There is a high correlation (Spearman's rho = .68) between participants' self-assessed search stage and familiarity with the search topic, which should not come as a surprise. Users grow more familiar with the search topic as the search stage evolves. A high correlation (Spearman's rho = .49) is also present between familiarity with the search topic and the specifiedness of the search question, which seems to suggest that users are able to pose more specified questions when they become more familiar with the topic. It is interesting to see that there is also a correlation (Spearman's rho = .39) between self-assessed stage and specifiedness of the question,

DOI: 10.1002/asi

TABLE 7. Cross-tab between query submission methods and search goal. (The two most frequently used methods for each type of search goals are singled out.)

	Submission				
	Not use the display		Use the classification display		
Search goal	Single line	Author/author subject	MeSH select	Multifacet input	Select & input
Known-item	2nd (28%)	1st (41%)			
Search for specific question	1st (37%)			2nd (26%)	
Browsing without specific question	2nd (28%)		1st (35%)		

Note. MeSH = Medical subject headings.

TABLE 8. Numerical variables for users' problematic situations ($N = 199$).

	1	~	7	*M*	*SD*
Self-assessed search stage	Initial		Finalized	3.55	1.92
Familiarity with the search topic	Not at all		Extreme	3.90	1.60
Thoroughness of the results	One good doc.		Thorough review	4.09	2.08
"Specificity" of the search	General		focused	4.92	1.91

TABLE 9. Correlations coefficients for numerical variables for problematic situations.

		Self-assessed search stage	Familiarity with the search topic	Thoroughness of results desired
Spearman's rho	Self-assessed search stage			
	Familiarity with the search topic	68(*)		
	Thoroughness of results desired	.10	.08	
	"Specificity" of the search	.39(**)	.49(*)	−.13

*Correlation is significant at the 0.01 level (2-tailed), $N = 199$.

which seems to confirm previous studies (e.g., Vakkari, 2001) that as users' information seeking process evolves, the search questions become more specific.

Discussion

The Relationships Between Problematic Situations and Query Submissions

The interaction between problematic situations and query submission methods was tested on two fronts, the participants' actual usage patterns, and their satisfaction with the results. Analyzing the usage patterns, patterns were observed between query submission methods preferred by the participants and types of problematic situations they had, whereas the findings on user satisfaction were more ambiguous.

Our findings that different aspects of users' problematic situations interacted with the interface features support the theoretical proposition that multiple representations methods are needed in anticipation of diverse information needs (Belkin et al., 1982; Ingwersen, 1994). Ingwersen expressed this view when he commented on the necessity of multiple cognitive structures for the system to infer users' needs:

> In other words, from a cognitive perspective as many and as different cognitive structures as possible should be made available and applied during IR interaction, however, in accordance with an estimation which allows for a controlled or calculated selection of exactly such structures that are regarded most appropriate to the current retrieval situation. The issue of estimation is not necessarily seen as a mathematical one but rather a behavioural and psychological issue (p. 105).

Ingwersen' argument was made mainly in the context of the dynamic nature of information needs during the information retrieval process; however, this study looked more specifically into users' initial needs and how they could be better supported by different query representation methods. Nardi and O'Day (1999) expressed a similar view when they commented that "information needs come in many flavors and categories, and there must be a corresponding diversity in tools and services to meet different needs" (p. 96).

The Variety of User Expressions

Most users found the interface easy to use; nevertheless, several of the participants mentioned that the terms in the thesaurus were not specific enough for their purpose.

DOI: 10.1002/asi

The two most frequent complaints about the browsable thesaurus were the lack of specificity of the thesaurus descriptors and the difficulty in finding a term in the hierarchies. Hersh, Haynes, and McKibbon (1994) showed that up to one fourth of physician-generated MEDLINE queries were not represented in the UMLS (unified medical language system) metathesaurus. Even with the vast number of terms in MeSH, the database still barely covers the range of query expressions the users might come up with. Bates (1989) has proposed using an end-user thesaurus to allow a greater variety of search queries. In this regard, the MeSH translation table feature of PubMed that automatically translate users' query terms into MeSH terms or text words can be seen as an more effective approach to accommodate users' expressions.

Conclusion

The Case for Interactive Information Retrieval Evaluation

Since the early days of the Cranfield experiment (Cleverdon & Keen, 1966), much effort has been devoted to contriving better algorithms to improve information retrieval performance, as measured by system-oriented metrics such as precision and recall. With recent advances in interactive technology and the prevalence of end-user search, it has become clear that the Cranfield model of evaluation is ill-equipped to address all the evaluation needs. It has been a great concern in the evaluation of interactive IR systems to balance the objective measurability and realism (Robertson & Hancock-Beaulieu, 1992). Inspired by the user-centered, cognitive view in IR, researchers have started looking into the context of using an IR system and how IR impacts real-life tasks of the users. Recent calls for tighter integration between the two research areas, information seeking behaviors and IR, highlight the felt need to make information seeking research relevant to the improvement of system performance.

Arguably the most significant contribution so far made by the user-centered view of IR is the incorporation of real users into the system evaluation process. Breaking away from the batch-mode style of system evaluation, the TREC (Text REtrieval Conference) interactive track adopted an evaluation paradigm where not only the performance, but also the interactive process between the user and system could be observed and analyzed. The interactive paradigm has grown out of the need for evaluating retrieval techniques whose success hinges on users' perception of and active involvement with the tool (e.g., relevance feedback and browsing-based search aids). As end-user search becomes more prevalent, issues such as interface usability and users' mental model of the system have started receiving more attention.

The Significance of Task in Information Retrieval Evaluation

The other methodological complication entailed by the user-centered approach is the construction of search tasks. Among the major components of the system evaluation: users, tasks and system features, systems or system features are the ones that are most frequently subject to experimental manipulation as the comparison of system performance is the most common focus of evaluation. In traditional IR evaluation, every effort is made to make sure that everything is equal other than the system components being compared. User, tasks , and their interaction effects are regularly treated as random variants the experimenters strive to control and minimize.

Such a controlled-experiment approach is very effective if the systems or system features compared are conceptualized as tools of general purposes. For a general purpose search tool, little can be discerned from the differentiation and categorization of search tasks. The assigned tasks are created mostly in an ad hoc manner, without theorizing task characteristics and how these characteristics might interact with the system features concerned. Yet with more and more interactive search features (e.g., visualization and clustering, automatic thesaurus construction, relevance feedback, etc.) being introduced, it is easy to imagine search tools becoming specialized in certain aspects of search process or certain types of search tasks. Their strengths might not be easily translated into achieving a higher precision or recall, the traditional gold standards of system performance, or be readily observable across all types of tasks. On the other hand, as the search features/aids become more diverse, there is also a growing awareness of the diversity of search tasks users are engaged in when searching for online information sources. The awareness of the diversity is partly due to the permeation of online searching in our daily life, and partly due to the theoretical advances in human information seeking behaviors (See, for example, Bystrom & Hansen 2005; Vakkari, 2003).

With the growing diversity of search tools and search types or search contexts, more attention ought to be given to the interaction effects between the two. Indeed, there is no lack of examples of retrieval techniques tailored to address certain types of search tasks and evaluated as such. For example, in the question-answering track and the novelty track within TREC, the systems are evaluated with criteria specifically designed to their respective contexts.

The strengths of the approach outlined here lie in the authenticity of the search problems and the naturalistic setting where the searches are performed. Yet the longitudinal methodology also introduces further complications to the analysis. The individual participant becomes the secondary sampling unit, where their problematic situations were repeatedly drawn from. In both χ^2 tests of two-way contingency table analyses, the relationships between the problematic situations and the query submission methods turned out to be significant. One might challenge the test results in that they did not include the participants' preference into the picture. Yet because there is little reason to believe that participants' preference might influence their problematic situations in a systematic manner that favor our hypothesis, it does not present a threat to the validity of the tests.

There is also a lack of objective evidence about the effectiveness of the interface as measured by traditional evaluation metrics in this study. This seems to be an inevitable

DOI: 10.1002/asi

consequence of sampling real information problems instead of assigning uniform tasks, as it precludes the use of comparable measures. One possible way of maintaining the balance between realism and comparable measurement in the future is utilizing the real life situation where multiple participants are assigned a uniform set of information problems such as classroom assignments or assignments for professional information analysts. Such a "group" or "team" real-life task environment will allow the use of both subjective and objective measures of performance. It also allows the further investigation of the relationships between these types of performance measures.

A more objective assessment of the effectiveness of the interface can also be achieved in a controlled experiment setting using simulated search scenarios to represent the problematic situations where the display tool is actually adopted by the participants in this study. The operationalization of search tasks remains, however, a challenge to be tackled within the future.

References

Allen, R.B. (1995). Retrieval from facets spaces. Electronic Publishing, 8(2&3), 247–257.

Anderson, J.D. (1990). Ad hoc, user-determined classification displays based on faceted indexing. In S. Humphrey & B. Kwasnik (Eds.), Proceedings of the 1st ASIS SIG/CR Classification Research Workshop (pp. 95–100). Silver Spring, MD: Information Today.

Anderson, J.D. (2002). Effective display of browsable classification on the WWW and other hypertext media. In J.-E. Mai, C. Beghtol, J. Furner, & B. Kwasnik (Eds.), Proceedings of the 13th ASIS&T SIG/CR Classification Research Workshop (pp. 110–123). Silver Spring, MD: Information Today.

Bates, M.J. (1989). The design of browsing and berrypicking techniques for online search interface. Online Review, 13(5), 407–424.

Bates, M.J. (2002). After the dot-bomb: Getting web information retrieval right this time. First Monday, 7(7). Retrieved July 11, 2002, from http://www.firstmonday.org/issues/issue7_7/bates/index.html

Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., et al. (2003). Query length in interactive information retrieval. In J.P. Callan, F. Crestani, & M. Sanderson (Eds.), Proceedings of SIGIR 2003 (pp. 205–212). New York: ACM.

Belkin, N.J., Oddy, R., & Brooks, H. (1982). ASK for information retrieval: Part I. Background and theory. Journal of Documentation, 38(2), 61–71.

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

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, 53(3), 225–250.

Broughton, V. (2002). Facet analytical theory as a basis for a knowledge organization tool in a subject portal. Challenges in Knowledge Representation and Organization for the 21st century: Integration of Knowledge Across Boundaries: Proceedings of the Seventh International Conference (pp. 135–142) Granada, Spain/Wurzburg: Ergon Verlag. Retrieved on May 20, 2002, from http://www.ucl.ac.uk/fatks/paper2.htm

Bystrom, K., & Hansen, P. (2005). Conceptual framework for tasks in information studies. Journal of the American Society for Information Science 56(10), 1050–1061.

Cleverdon, C., & Keen, E. (1966). Factors determining the performance of indexing systems (vol. 1: Design, vol. 2: Results). Cranfield, England: Aslib Cranfield Research Project.

Cordes, R.E. (2001). Task-selection bias: A case for user-defined tasks. International Journal of Human Computer Interaction, 13(4), 411–419.

Cullen, R.J. (2002). In search of evidence: Family practitioners' use of the internet for clinical information. Journal of Medical Library Association, 90(4), 370–379.

Drabenstott, K.M. (2001). Web search strategy development. Online, 25(4),18–27.

Ellis, D., & Vasconcelos, A. (2000). The relevance of facet analysis for world wide web subject organization and searching. Journal of Internet Cataloging 2(3/4), 97–114.

Heo, M., & Hirtle, S.C. (2001). An empirical comparison of visualization tools to assist information retrieval on the web. Journal of the American Society for Information Science and Technology 52(8), 666–675.

Hersh, W.R., Hickam, D.H., Haynes, R.B., & McKibbon, K.A. (1994). A performance and failure analysis of sapphire with a Medline test collection. Journal of American Medical Information Association, 1, 51–60.

Hearst, M. (2000). Next generation web search: Setting our sites [Special issue]. IEEE Data Engineering Bulletin, 23(3), 38–48.

Hert, C.A. (1996). User goals on an online public access catalog. Journal of the American Society for Information Science, 47(7), 504–518.

Ingwersen, P. (1994). Polyrepresentation for information needs and semantic entities: Elements of a cognitive theory for information retrieval interaction. In W.B. Croft & C.J. van Rijsbergen (Eds.), SIGIR '94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (pp. 101–110). New York: ACM.

Jacob, E.K. (2001). The everyday world of work: Two approaches to the investigation of classification in context. Journal of Documentation, 57(1), 76–99.

Kaptelinin, V. (1996). Computer-mediated activity: Functional organs in social and developmental contexts. In B. Nardi (Ed.), Context and consciousness: Activity theory and human–computer interaction (pp. 45–68). Cambridge, MA: The MIT Press.

Kekäläinen, J., & Järvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 130–137). New York: ACM.

Kelly, D., & Belkin, N.J. (2002). A user modeling system for personalized interaction and tailored retrieval in interactive IR. In Proceedings of Annual Conference of the American Society for Information Science and Technology (pp. 316–325). Colorado Spring, MD: American Society for Information Science and Technology.

Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. Journal of American Society for Information Science, 45(5), 361–371.

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

Marchionini, G. (1995). Information seeking in electronic environment. Cambridge, UK: The University of Cambridge Press.

Nardi, B., & O'Day, V. (1999). Information ecologies: Using technology with heart. Cambridge, MA: The MIT Press.

Pollitt, A.S. (1998). The key role of classification and indexing in view-based searching. International Cataloguing and Bibliographic Control, 27(2), 37–40.

Robertson, S.E., & Hancock-Beaulieu, M.M. (1992). On the evaluation of IR system. Information Processing and Management, 28(4), 457–466.

Saracevic, T., Kantor, P., Chamis, A.Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. Journal of the American Society for Information Science, 39(3), 161–176.

Soergel, D. (1985). Organizing information: Principles of data base and retrieval systems. San Diego, CA: Academic Press Professional.

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. Journal of the American Society for Information Science, 45(8), 589–599.

Vakkari, P. (2001). Changes in search tactics and relevance judgments where preparing a research proposal: A summary of the findings of a longitudinal study. Information Retrieval, 4, 295–310.

DOI: 10.1002/asi

Vakkari, P. (2003). Task-based information searching. In B. Cronin (Ed.), Annual Review of Information Science and Technology (Vol. 37, pp. 413–464). Medford, NJ: Information Today.

Vakkari, P., & Hakala, N. (2000). Change in relevance criteria and problem stages in task performance. Journal of Documentation 56(5), 540–562.

Vakkari, P., Jones, S., MacFarlane, A., & Sormunen, E. (2004). Query exhaustivity, relevance feedback and search success in automatic and interactive query expansion. Journal of Documentation, 60(2), 109–127.

Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. Information Processing and Management, 39(3), 445–463.

White, M.D. (1975). The communications behavior of academic economists in research phases. Library Quarterly, 45(4), 337–354.

Yuan, W. (1997). End-user searching behaviors in information retrieval: A longitudinal study. Journal of the American Society for Information Science, 48(3), 218–234.

DOI: 10.1002/asi