Dissertation Proposal
"Browsing in a faceted information space: a longitudinal study of PubMed users' assimilation of a novel display tool"

Doctoral Candidate
Muh-Chyun Tang
School of Communication, Information and Library Studies
Rutgers, the State University of New Jersey,
4 Huntington Street,
New Brunswick, NJ 08904
muhchyun@scils.rutgers.edu

Advised by Dr. Nicholas J. Belkin
nick@belkin.rutgers.edu

Motivation of the problem

One of the key components in interactive IR (Information Retrieval) theory is the recognition that the users' query is only an imperfect representation of their actual information needs. Based on the rather short and superficial description of users' information needs, the matching function performed by current IR systems is essentially a matter of approximation. Belkin's (1982) concept of the Anomalous State of Knowledge (ASK) succinctly captures the dilemma faced by users who are often ill-equipped in articulating their information needs.

The issue of bridging users' ASK and the information source only becomes more crucial in the era of end-user searching. End users today enjoy direct access to a wide spectrum of electronic information sources, yet most of them have difficulty utilizing the full capacity of what modern IR systems can offer. Without effective external aids, users are left to their own devices when it comes to representing their information needs for retrieving and managing the search results.

One of the approaches to alleviate these difficulties in query formulation is through the display of indexing vocabularies generated by either humans (Hearst et al., 2002) or algorithms (Chen et al., 1995). Classification display has long been used to facilitate users in clarifying their information needs and supporting browsing. In classification literature, the advantages of the faceted analysis over the traditional enumerative scheme have long been recognized (Ingwersen and Wormell 1992, Svenonius 1992, 2000; Star, 1998; Vickery, 1960). It has also been suggested that a thorough faceted analysis applied in query formulation is conducive to favorable results (Drabenstott 2001). Yet despite the growing interest in using faceted classification to

support structured browsing on the Web (Allen, 1995; Anderson, 2003; Anderson, 1990; Bates, 2002; Broughton, 2001; Ellis & Vasconcelos, 2000; Hearst, 2000; Pollitt, 1998), little has been empirically investigated so far on how the faceted approach might be applied to the representation of users' information needs.

Research Questions

The overarching motivation of the study is to investigate how useful a faceted classification display might be in facilitating access to a relatively complex knowledge domain such as health science. Two groups of research questions can be drawn from the general purpose of the study.

The first group of research questions involves users' perception and acceptance of a faceted classification, display as compared with the search system they have been accustomed to. This will entail the investigation of users' assimilation of the display tool over time. This part of the inquiry will be directed by activity theory where the internalization of the external tool is highlighted (Vygotsky, 1978; Nardi, 1996). The research questions raised specifically by the assimilation of the novel tool are:

Q1: Can users learn to use the faceted classification display effectively in a short period of time?

Q2: How does the user's perception of the faceted display evolve over time? Will users, who have become accustomed to the traditional subject search strategy, grow to be able to construct their quests in terms of faceted categories?

Q3. How effective is the classification display in eliciting more queries and queries from different semantic categories (facets)?

Q4: What impact, if any, might users' familiarity with the faceted display have on their search query forming process?

The other aspect of this study concerns the compatibility between the user's problematic situations and the perceived strength of the tools. This part of the investigation will be informed by studies and theories in interactive IR and information seeking behaviors. Two aspects of information seeking contexts will be addressed here: firstly, at the external or institutional level, the relationships between searchers' work tasks and the patterns of problematic situations (Byström & Järvelin, 1995). And secondly, at the immediate cognitive level, the relationships between the problematic situations and the search tool and strategies used will be explored (Belkin et al. 1993; Kelly & Belkin 2002; Saracevic 1988; Vakkari 2001a, 2001b, 1999). The research questions regarding to the relationships between information seeking contexts and IR interactions are:

Q4. What kinds of problematic situations are best supported by the faceted display?

Q5. Can relationships be drawn between users' work tasks and their problematic situations during interaction with the IR system?

Design Principle

To test the effectiveness of using a faceted classification in query formulation, a display of MeSH (Medical Subject Headings) tree is designed for the access of PubMed bibliographic database. PubMed has been chosen because it is one of few databases where faceted analysis for indexing is performed. 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 attempted to utilize the faceted

approach to classification display all rely heavily on browsing and direct manipulation of the classification structure (Hearst et al., 2002; Pollitt, 1998; Allen, 1995). Yet it is arguable that keyword searching has been the access mode to the electronic sources most users have come to be accustomed to. Instead of relying solely on browsing, the proposed interface preserves the search mode of access while providing a browsable thesaurus to facilitate query formulation.

In the implementation of the faceted display, all the 15 top-level categories of MeSH will be presented to the users throughout the search process. The users will be given the options of either browsing and selecting the terms within the MeSH tree, or conducting keyword search by submitting their own terms in any of the 15 query boxes corresponding to the top-level categories/facets.

The display has been designed so that when users click on a category, a separate window will open up to display the terms in the chosen category. Users can use the choosing and browsing iteratively to narrow down their exact search requests. At each stage, the terms chosen will be automatically fed into the corresponding query boxes (A tutorial for the interface can be found online at http://www.scils.rutgers.edu/~muhchyun/interfacedemo/).

Methodology

One of the major methodological issues that needs to be tackled is the familiarity effect when comparing relatively novel features with customary ones. The user might prefer or perform better with the tool s/he is most familiar with. In their study of the usability of three visualization tools, Heo and Hirtle (2001) concluded 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)." 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" (Cordes, 2001). He called for employing user-defined tasks in place of product-supported tasks to avoid task-selection bias.

Such sensitivity to the developmental aspect of tool use is in accordance with the basic assumption of activity theory that there is a dialectic relationship between individuals' mental activities and their cultural and technical environments. It is hoped that a longitudinal study will be a sufficient solution to the familiarity effect because it will enable us to observe the learning and assimilation of interface by the users, which is often unavailable in a strictly controlled environment.

Furthermore, allowing the participants to search for their own search problems in a naturalistic setting also helps us capture the "embeddedness" of tool use. A naturalistic design reflects better the fact that users' encounters with the information access system is hardly an end itself as it is always part of their effort to cope with a problematical situation at hand.

Research Procedures

A total of 20 participants from the fields covered by PubMed such as medicine, nursing, dentistry, veterinary medicine, the health care system, and the pre-clinical sciences will be recruited for the study. Each participant will be asked to conduct eight search sessions at times of their own choosing in the time period of two months. The participants will be asked to create an account and access the interface through a proxy

server at their workplace instead of coming to the laboratory. Before the experiment begins, the participant will be asked to complete an entry questionnaire, given detailed instructions of the research protocol and a demonstration of the functions of the faceted display and other search options. Each time the participants search on the interface they will be asked to fill out a pre-search questionnaire, perform searches on the information problems of their own, and answer a post-search questionnaire. They will be given the option of either using the faceted display or using the traditional search mode provided by PubMed. All the questionnaires will be administered online to make remote monitoring possible.

Data Collection and Analysis

The variables that we are interested in can be summarized into six groups: user characteristics, types of work tasks, characteristics of users' problematic situations, users' perceptions of the display, use of search tools and strategies, and user satisfaction (See Table 1 for a grouping of variables).

The comparison of the registered numbers of usage between two querying methods: the conventional query box and the faceted classification display, will give us the basic indication of users' preference. A categorical variable, "stages of exposure" will be created by dividing the search sessions into either two or three values (early/late, early/middle/late) so a Chi-Square Test can be performed to determine whether the distribution of the querying methods differ significantly in stages of exposure.

The querying methods can be further categorized into four basic types of "search modes," Chi-Square Tests will also be performed to test the relationship between "search

modes (See details in Table 1)" and "stages of exposure," and that between the subject's "search modes" and search goals.

Table 1

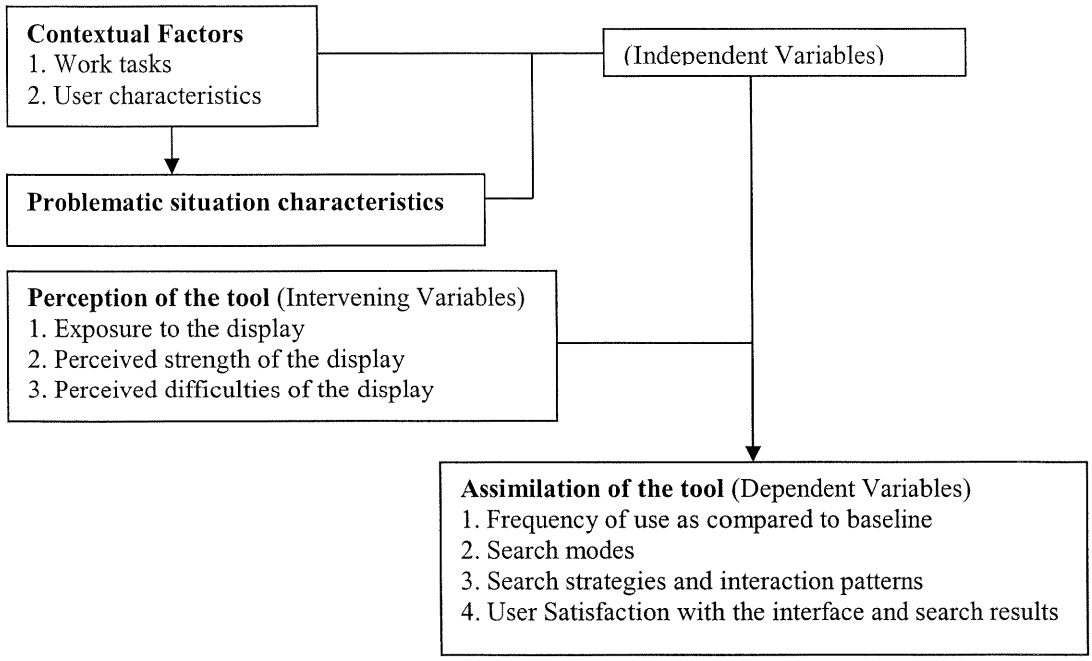
Variable group	Variable label	Methods of Collection
User Characteristics	Primary Occupation; Specialty; Age; Gender; Past Search Experience with PubMed, MEDLINE and MeSH (frequency, purposes, self-assessed expertise).	Entry Questionnaire
Work Tasks	Research, (monitor latest development, forming hypothesis etc.) Teaching (preparing teaching materials, answer student question, etc.), Primary care (Information for patients, monitor latest development etc.)	Pre-search questionnaire
Problematic Situations	Type of search tasks (background information on a familiar topic area, background information on an unfamiliar topic area, comprehensive exposition of a previous familiar area, comprehensive exposition of a previous unfamiliar area, and a specific fact); Search goals (look for known articles, look for unknown articles for a specific question, and browse without a specific question in mind); Familiarity with the topic; Complexity of the problem; Specificity of the problem; Thoroughness of the results desired;	Pre-search questionnaire
Tools and Strategies used	Search modes ("string search using simple query box," "string search using the classification display" and "scanning and selecting using the classification display," as well as "combination of string search and the classification display"); Search characteristics (interfaces chosen, number of query/facets used, duration, number of submissions, pages viewed/seen, use of free-text, use of fields in "limits"); Query characteristics (root facet, free-text or MeSH terms).	Log files
Perception of the display	Reasons for not using the display; Usefulness of the display; Easiness of finding terms in the MeSH display, Confidence in using the display; Familiarity with the display; Perceived advantages of the interface for this search (by showing the authorized terms to represent the concepts I already have in mind, by reminding me of concepts I can use, by showing the scope and structure of the database, and by helping me manage the size of the returned hits)	Post-search Questionnaire
User Satisfaction	Satisfaction with the display; Satisfaction with the search results	Post-search Questionnaire

To test the effectiveness of the classification display in query elicitation, comparisons will be made between the query terms the participants actually use in each search session and the search concepts they have in mind before the actual search. As an ancillary question to the effectiveness of query elicitation, we will also investigate whether the participants grow to internalize the faceted structure. This will be done by looking into, as the study progresses, whether there is an increase in either the number of the pre-search concepts or the facets these concepts derive from. The increase in the number of the pre-search concepts will suggest that the participants have internalized the faceted categories during their interactions with the display.

Correlation Tests will be performed between the problematic situation variables and perceived advantages of the classification display for today's search, both are measured on a 1-7 scale. Other quantitative information about each search session such as number of terms used, number of facets chosen, number of modification of search will be also be analyzed along with variables that represent the participant's problematic situations.

In summary, the study sets out to investigate the adoption and usage patterns of the faceted display, which will be the dependent variables, in a naturalistic setting. A group of contextual factors, which compose mainly of characteristics of searchers' problematic situations, will be introduced as independent variables. The relation between problematic situations and strategies used is hypothesized to be mediated by searchers' perceived strengths and difficulties when using the display (See figure 1 for a graphic representation of the relationships among variables).

Figure 1.



Reference

Allen, R. B. (1995). Retrieval from Facets Spaces. *Electronic Publishing* 8(2&3): 247-257.

Anderson, James D. (2002). "Effective Display of Browsable Classification on the WWW and other Hypertext Media." In Jens-Erik Mai, Clare Beghtol, Jonathan Furner, and Barbara Kwasnik (Eds.) *Proceedings of the 13th ASIS&T SIG/CR Classification Research Workshop* (pp.110-123). Silver Spring, MD: Information Today.

Anderson, James. D. (1990). Ad Hoc, User-Determined Classification Displays Based on Faceted Indexing. In Susanne Humphrey and Barbara Kwasnik (Eds.), *Proceedings of the 1st ASIS SIG /CR Classification Research Workshop* (pp. 95-100). Silver Spring, MD: Information Today.

Bates, M. J. (2002). After the Dot-Bomb: Getting Web Information Retrieval Right This Time. *First Monday* [On-line serial], **7(7)**. Available http://www.firstmonday.org/issues/issue7_7/bates/index.html

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

Belkin, N. J., Marchetti, P.G. and Cool, C. (1993). BRAQUE: Design of an interface to support user interaction information retrieval. *Information Processing and Management* 29, 325-344.

Broughton, V. (2001). Faceted classification as a basis for knowledge organization in a digital environment: the Bliss Bibliographic Classification and the creation of multi-dimensional knowledge structures. *New Review of Hypermedia and Multimedia*, (7), 67-102.

Byström, K. and Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management*, 31(2), 191-213.

Chen, Hsinchun, Yim, Tak, Fye, David and Schatz Bruce (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science* 46 (3), 175-193.

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

Drabenstott, Karen M. (2001). Web search strategy development. *Online* 25(4) (Jul/Aug),18-27.

Ellis, David, Ana Vasconcelos (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., S. C. Hirtle (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.

Hearst, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ping Yee. (2002). Finding the flow in Web site search. *Communications of the ACM 45*(9), 42-49.

Ingwersen, Peter and Irene Wormell (1992). Ranganathan in the perspective of advanced information retrieval. *Libri* (42),184-201.

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 (ASIST '02)*, Philadelphia, PA, 316-325.

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

Nardi, Bonnie A. (Ed.) (1996). *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Cambridge, MA, 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.

Star, S. L. (1995). Grounded Classification: Grounded Theory and Faceted Classification. *Library Trends: How classification work: problems and challenges in an electronic age* 47(2), 218-232.

Svenonius, Elaine. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, Massachusetts, The MIT Press.

Svenonius, Elaine. (1992). Ranganathan and Classification Science. *Libri* 42(3), 176-183.

Vakkari, Pertti. (2001a). A Theory of The Task-Based Information Retrieval Process: A Summary and Generalisation of A longitudinal Study. *Journal of Documentation* 57(1), 44-60.

Vakkari, Pertti (2001b). 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.

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

Vickery, B. C. (1960). *Faceted classification: a guide to construction and use of special schemes*. London: Aslib

Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes*. Cambridge, Massachusetts, Harvard University Press.

Schedule of Completion January 2004 to March 2005

Activity	J	F	M	A	M	J	J	A	S	O	N	D	J	F	M
Develop and refine the interface	—	—	—	—	—	—	—	—							
Develop and refine the instruments	—	—	—	—	—	—	—	—							
Pilot study							—	—	—						
Recruit participants								—	—						
Collect data									—	—	—				
Analyze data											—	—	—		
Report the results													—	—	—

Budget Justification

Interface development $ 750

A total of 50 hours are needed for the programming of the interface, paid at the rate of $15 per hour

Compensation for the participants $ 1250

A total of 25 participants (including the participants in the pilot study) will be recruited and each paid $50 for their efforts.

Total expenses $ 2,000

*Currently no other sources to support the research