BEYOND TEXT QUERYING AND RANKED LIST:
FACETED SEARCH IN LIBRARY CATALOGS

Xi Niu <xiniu@email.unc.edu>
University of North Carolina at Chapel Hill (UNC)

BACKGROUND

Information technologies today are experiencing greater use than at any other time in their history, and more importantly, by average people other than scientists. As information becomes more and more available, various search technologies are in demand to facilitate access to information. A search system today should go beyond the traditional query-response and ranked list paradigm to incorporate more human search behaviors, such as filtering, browsing, and exploring, in addition to simple lookup. As described by Belkin’s (1980) concept of Anomalous State of Knowledge (ASK), text querying is not always the best solution to express users’ information need. Similarly, a ranked list of results with the assumption of “the best first” is not without shortcomings. Information seekers often express a desire for a user interface that organizes search results into meaningful groups, in order to better understand and utilize the results (Hearst, 2006).

Faceted search plays a key role in the demand for new forms of search interfaces. It provides a categorized view of the search results, as an alternative to best-first search. The underlying theory for faceted search is the facet theory, a classic theory of knowledge representation developed in the 1930’s by Ranganathan, an Indian mathematician and librarian. In recent years, faceted search has grown to be a well-accepted approach. It has been applied as a standard technique in commercial Web sites for many years (Breeding, 2007). Since 2006, a small number of academic libraries have implemented faceted search on their online catalogs. Among them are North Carolina State University Library, the McMaster University Library, the State University Libraries of Florida, the OCLC’s WorldCat, etc. Since the adoption of faceted search in these academic libraries, faceted library catalogs have gained popularity in many academic or public libraries. Faceted search has become one of the features used to describe the “next-generation” catalogs.

In this study, I will investigate the effectiveness of faceted library catalogs, explore how people use facets combined with text search in library catalogs, and discuss implications for future library catalog design and more broadly, search system design.
PROBLEM STATEMENT AND RESEARCH QUESTIONS

For a long time, online catalogs have been criticized as being hard to use. According to some researchers, library catalog designs do not incorporate sufficient understanding of search behaviors. It is believed that modern online catalogs have been able to do a good enough job of tackling the problem of what library scientists called known-item search, in which the user knows what documents are searched for or at least certain aspects about the documents. In contrast, comparably mature tools for exploratory tasks, where the information need and the target documents may not even be well established, are not well developed. Searchers always have problems when dealing with these more complex exploratory search tasks. Common problems are: 1) choosing suited terms to represent their information need, 2) reformulating or re-scoping their search after the initial search, and 3) being overwhelmed by long result lists or zero returns.

Then faceted library catalogs has emerged with the potential benefit of solving those problems. From the catalog developers’ perspective, faceted catalog may be helpful in at least three ways: assist users in issuing a query by exposing them to available facets; guide users to refine the initial search by drilling down a specific facet; and help users discover the hidden items by flatten out the depth of the original result list.

However, it remains an empirical research question whether library patrons are able to take advantage of this facet feature and whether this feature really helps them locate what they need in library.

This study is to investigate if faceted catalog provides more effective information-seeking support, especially more support of exploratory tasks to users than text querying and best-first search. In commercial settings, the advantages of faceted search over conventional search have been known more and more. Yet little is known about the effectiveness of faceted search applied in library catalogs. Even less is known about the effect of different facet implementations on people’s interacting with the catalogs. These issues are encapsulated in the following guiding questions:

RQ1: Whether faceted interfaces help improve the search with library catalogs?

This research question concerns the effectiveness of faceted library catalogs. In order to address this overarching question, two research questions have been developed:

RQ1.1: Does the faceted search help improve the search performance?

RQ1.2: Does the faceted search help improve the user satisfaction?
The manipulated UNC library search interface, for which the facet feature has been removed, will serve as the baseline. The regular UNC faceted search interface will be the experimental interface. Subjects will be recruited to experience both interfaces. Comparison will be made across the two interfaces according to search performance and user satisfaction. Search performance is defined as search accuracy, search duration, and other descriptive measures. User satisfaction is an affective measure reflecting users’ feelings and perceptions when searching with the faceted search.

The second research questions involve how people use facets. The question is addressed below:

**RQ 2: How facets are used in the library catalogs?**

UNC faceted catalog interface and Phoenix Public Library (PPL) catalog interface are chosen as two cases of the two mainstream facet designs. The significant difference between the two is the support of faceted browsing. Phoenix has the facet hierarchies available from the starting point of a search. Users are able to reach the result set by browsing the hierarchies without entering any text query. In contrast, the facets on UNC interface do not have hierarchies for browsing and are not available on the opening page. Another difference is about metadata. Generally speaking, the UNC Library is highly dependent on the MARC records and Library of Congress Subject Headings (LCSHs) to generate metadata. In contrast, the PPL has more varied sources of metadata and subject headings (including Book Industry Standards and Communications headings, also called BISAC subject headings). The commonalities are that both catalogs have leading high quality facet implementations and they both implemented the faceted search with Endeca\(^1\). Under this overarching question, three research questions have been raised:

- **RQ 2.1: Does the use of facets differ in two faceted catalogs? If so, in what ways?**
- **RQ 2.2: Beyond facet use, what is the typical search process in two library settings?**
- **RQ 2.3: Can we establish the relationship between users’ query formulation tactics and the task types?**
- **RQ 2.4: How are facets used in conceptual high-level search tactics?**

To answer these questions, subjects will be recruited to experience both catalogs. Due to the different information collections and different facet implementations, direct quantitative comparisons across the two

---

\(^1\) Endeca is a software company that provides faceted search platform
interfaces are hard to make. Rather, the result will be compared more “descriptively” and “qualitatively”.

RQ 2.1 deals with how, when, and why facets are used in two library settings. Some hypotheses are proposed. For example, is facet usage higher when faceted browsing is supported? When both browsing and refining are supported by facets, are people more likely to use facets for browsing?

RQ 2.2 is about the whole search process, which can be described quantitatively by investigating the session lengths, action distributions, clusters of users, transition probabilities, and qualitatively by search sequence patterns.

RQ 2.3 concerns the relationship between query formulation tactics and task types. Two types of search tasks are proposed: close-ended and open-ended search. Descriptions of these two types are in the methodology section. The query formulation/reformulation tactic is a general term referring to the methods and strategies adopted by searchers for issuing a query for a search system. The tactics include query submission methods (through text or facet), tabs chosen (simple search or advanced search, book collection or DVD collection, etc), query length, and the number of query submissions during search sessions. Would searchers issue shorter queries, click more facets, and conduct more search iterations for open-ended questions? In addition to the discrete search activity level, query formulation tactics will also be investigated at the query sequence level by analyzing the sequential query formulation patterns through case studies. Is specified reformulation, where a user persists in specifying previous queries, more common for close-ended tasks? In the contrast, is parallel reformulation, where a user modifies the queries from one aspect to another, frequently seen in open-ended questions? Is the interface design a confounding factor of task type in affecting searchers’ interaction style?

RQ 2.4 addresses the impact of facets on the whole search tactics beyond just query formulation tactics. This question particularly investigates the search tactics enabled by facets. Past literature (e.g., Kules, 2006) and preliminary transaction log analysis have suggested that searchers may adapt their search tactics when using facets. They may use facets as a way of organizing their search order, an overview, a backup plan, or completely ignore them. They may also explore deeper or use shorter queries for later refining. This study will identify the search tactics and strategies made possible by the two facet implementations. Do searchers rely on facet hierarchies to guide their search when the hierarchies are available to them? Do
searchers explicitly use facets by watching them without clicking them through, to make sure they are at a right place?

SIGNIFICANCE AND CONTRIBUTION

This study is not only about the “whether” question, but more about the “how” question, which aims at understanding how people use facets in library catalogs. An insight into people’s search behavior with facets will have implications for OPAC (Online Public Access Catalog) developers to develop more responsive systems for different behaviors and different patrons. Future OPACs might separate the “probing search” from the “quick search” as a way of separating the paths for known-item search and subject search. Future academic OPAC design may take advantage of some features that have already been implemented by public OPACs, such as adopting varied sources of metadata and subject headings. The commercial bookstore flavor design of the Phoenix Library interface, includes BISAC headings, slide show of popular book covers, book reviews of popular sources, and N.Y Times book recommendations, proves to increase the catalog visits. It may provide ideas for academic libraries to better serve their patrons.

Many OPAC users in the libraries, especially in academic libraries, are likely to be the Web search engine and commercial website users. They will bring their mental models and search experiences to OPACs (Young & Yu, 2004). The result of this study will give thoughts to libraries that need to benefit from the web search engine instead of being overwhelmed by it. The findings will also have implications for developing collaborative relationship between library services and commercial web search engines.

Subject headings and classification numbers have been notoriously difficult to understand for average users. Knutson (1991) suggested that inadequate subject access is one of the reasons why many items in large academic libraries are hardly, if ever, checked out. More subject search has been replaced by keyword search since then. One potential benefit of faceted library catalogs is to make subject headings more friendly and accessible to users, by decoupling and repackaging these subject headings and exposing them to users. This study will pay special attention on use of the subject facet to investigate if users are able to take advantage of this benefit. In addition, since Phoenix library adopts two types of subject headings—LCSH and BISAC, the comparison between the two subject heading systems will have implications for catalogers to better tailor the subject headings for users.
The faceted search itself is a case where a classic theory can be used in a newer context. Faceted search is an application of the classic facet theory in an online, digital environment, although the specific definition of facets has been tailored to better fit the newer context. It demonstrates the importance of theories and concepts in the field of information science. Many technological innovations and revolutions are in fact the applications of the fundamental concepts and theories in the field.

Another contribution of this work is the methodological contribution. This research will demonstrate an innovative method for analyzing transaction logs of modern, faceted search and browse OPACs. The innovative methods include user session identification rules, differently grained coding schemes, and automated data processing scripts. In addition, the visualization technique named VUTL (Visualization for Understanding Transaction Logs), supports exploration and analysis of large amounts of log data. The framework for automatically converting transaction log data into visualizations, and the open software tool can be used for other libraries.

METHODOLOGY

This is an empirical research under the broad research area of information seeking behavior with a special focus on library catalogs as the search environment. Combinational use of user experiment and transaction log analysis (TLA) is proposed. While TLA is a non-intrusive, inexpensive way of collecting large amount of data from a great number of users, it fails to capture any information about the “context” in which the event occurred. The contextual information includes users’ demographics, motivations, satisfactions, etc. The user experiment will complement the limitation of TLA by providing those missing information. In addition, the user experiment gives researchers more control about the search process. On the other hand, the limitation of the user experiment, such as the limited amount of data collected, artificial problematic scenarios, can be complemented by TLA. The logs are the supplemental data for the user experiment, not only because of the large amount of data and information richness, but also in that logs can capture users’ behaviors in a naturalistic setting in addition to a lab environment.

User Experiment

Experiment Design

There are tradeoffs between pure experimental (between-subject) and repeated measure (within-subject) user studies. In this work, I conduct two studies using similar experimental conditions, a within-subject
study with 24 participants and a between-subject study with 12 participants. In the first experiment, the learning effect across the two interfaces is not major and the burden for each subject to experience two interfaces is reasonable. Thus within-subject design is preferred. We manipulate the UNC interface by removing the facet features as baseline. The UNC regular interface with the facet part is the experimental interface. Each subject is assigned two types of search tasks with both search interfaces. The orders of the interfaces and the tasks are controlled by Latin square design to reduce the order bias. The search performance and user satisfaction with the experimental interface will be compared with those in the baseline interface. In the second experiment, the learning effect is unwanted and reuse of the results from the first experiment is desirable, so between-subject design is chosen. Each subject will be recruited to experience the PPL search interface with two kinds of search tasks. Then the results will be compared to the 12 subjects’ searches randomly picked out from the first experiment. Latin square design is used to counterbalance the order effect. The independent factors of this experiment are the interfaces and task types, whereas the outcomes are facet use and search tactics.

Along with the experiment are two questionnaires (a post-search questionnaire and a post-interface questionnaire) and a semi-structured interview. All these instruments are to collect users’ feelings, perceptions, their likes and dislikes about the catalogs.

Tasks

Past literature of OPAC studies suggest that people primarily conduct two types of searches with OPACs. One is the known-item search where the user wants to find a specific item with information such as author, title, and publication year. In contrast, another type of search—subject search, conducted on a topic using either a keyword or a Subject Heading, is also frequently conducted by users. Known-item search and subject search are also called close-ended and open-ended questions because the former has a definite answer while the latter has more open-ended answers. Accordingly, in this research, two types of tasks are proposed: close-ended tasks and open-ended tasks. Open-ended tasks are of special interest to this study since the faceted search is designed to provide effective support to exploratory tasks. In this study, open-ended tasks are further broken down into two sub-types based on two possible search paths enabled by the faceted feature: facet searching, where the task is best solved by text queries combined with some facets; and facet browsing, where the task is best solved by browsing facets without any text query. Each type
(close-ended, open-ended facet searching, and open-ended facet browsing) of search tasks has two levels of
complexity based on the completeness of the information provided to the searchers. Pre-defining task
difficulty is a way to control the difficulty of the tasks presented to searchers. Searchers, however, may not
agree with the difficulty definition. Therefore, searchers’ perceived difficulty will also be collected through
the post-search questionnaires to compare with the pre-defined task difficulty.

Subjects

In both experiments, the subjects’ prior searching experiences need to be controlled. They are expected to
have similar experiences with the UNC catalog and the PPL catalog. Based on the above reason and also
taking convenience into consideration, the study will recruit freshmen across the UNC campus as
experiment subjects. Freshmen are assumed to have similar searching experiences and have similar levels
of familiarity to the UNC catalog. They also represent novice users and have limited experience and
therefore possible bias with either version of the library interfaces.

Transaction Log Analysis

When users interact with information retrieval (IR) systems they leave evidence of their actions in the
server’s transaction log file. The transaction log is one tangible artifact of the many digital footprints people
leave as they interact with systems. Through this footprint, better understanding will be gained about the
interactions that have occurred during a searching episode between the searchers and the faceted catalogs.

Two huge log data sets have been collected for this study: 15,237,808 records of UNC catalog server logs
(5.0G, ranging from 12/16/2008 to 4/16/2011); and 5,237,808 records of Phoenix Public Library server
logs (1.3G, ranging from 10/1/2010 to 3/31/2011).

A series of Perl scripts were written to conduct data cleansing, data parsing, and data coding work. Several
technical challenges are met during the processing of the logs. Huge data size is one of them. Scalability
makes a difference in data processing. In addition to the huge data size, another challenge is the
identification of meaningful search sessions. In this paper, we define a “session” as series of consecutive
requests from the same IP address with no periods of inaction greater than 30 minutes. Another technical
challenge is to code log entries as particular individual actions that are part of the search process. Due to
the large amount of the data, the coding work has to be done automatically rather than manually. A Perl
script is used to infer the action for each line of logs by comparing its URL and referrer.
Compared to the technical challenges mentioned above, what challenges us more is to interpret the logs and to understand the search behavior represented by the logs. Are we able to discern the search clusters and patterns through mining the log data? Are we able to identify the underlying search tactics adopted by searchers from those identified clusters and patterns? Do searchers seek information differently in a naturalistic setting from a lab setting? It’s even more challenging to translate the understanding of the search behavior to the implications for future faceted OPAC designs.

**Deep Analysis on Transaction Logs**

*Visualization of Transaction Logs*

This research proposes a visualization technique, named VUTL (Visualization for Understanding Transaction Logs). The technique, based on the flowchart model, utilizes standard Web programming tools (XML, HTML, JavaScript) to automatically convert standard library transaction logs into the graphical representations that can be visualized in web browsers. The resulting visualizations support exploration and analysis of large amounts of log data produced by library catalog interfaces, so that individual and group behaviors can be studied. This technique will be applied to logs from both UNC Library and the Phoenix Public Library. The details of the visualization can be found in Niu and Hemminger (2010). One visualization example is available at http://ils.unc.edu/bmh/pubs/SearchLogAnalysis_UNC2010/instructions.html.

*Clustering*

In order to investigate whether the search sessions derived from transaction logs naturally segregate into groups, a clustering method is proposed to detecting groups of homogeneous sessions based on session characters. It is believed that dividing users into groups with common features aids in the study of the search process and usage patterns. In this study, the session character is the distribution of actions within the session. To reduce the computational cost and avoid subjectively predefining the number of clusters, a hybrid approach of hierarchical and non-hierarchical methods is employed (Chen & Cooper, 2001).

*Sequential Event Analysis*

In order to investigate the search action sequence in the “temporal context”, Markov models are applied to describe the sequence of moves performed by searchers in a session. Furthermore, an order test (based on the chi-square goodness-of fit, Anderson and Goodman (1963)) is conducted to indicate the statistical
significance of the zero-, first-, second-, or higher order expressions of the process. However, the Markov action sequences identified in preliminary studies do not provide enough information about the whole search process and whole action patterns since they are just short segments extracted from the search process. Longer Markov sequences, on the other hand, not only are computationally expensive, but also fail the statistical significance test. The problems of Markov model lead to another complement analyzing method for sequential actions: Maximal Repeating Patterns (MRP). According to Siochi and Ehrich (1991), an MRP is “a repeating pattern that is as long as possible, or is an independently occurring substring of a longer pattern” (Siochi & Ehrich, 1991, p.316). According to this definition the algorithm systematically identifies those sequences of events that occur repeatedly within the data set. When using MRP, a balance between the sequence length and the sequence frequency should be made.

CONCLUSION

The anticipated outcomes from this research—a set of rich stories about how library patrons use facets to search the library catalog, categorized via interfaces and task types—will provide an illustration of the landscape of faceted searching and browsing activities, as well as an insight into complex practices of faceted catalog design that may be affected by many artificial decisions, local requirements, and other forces. This research will illustrate several features and factors of this landscape: problematic situations (close- or open-ended), search environments (lab or naturalistic), interfaces (UNC or Phoenix), search actions, search sequences, search tactics, search results (performance), and affective feelings. Understanding of these features and factors will be translated to implications for faceted OPAC design, and hopefully extrapolated to other OPACs or search systems.

This dissertation also intends to serve as a novel application of deep transaction log analysis to this domain. Analyses will focus on both the discrete search activity level and the search process level. A hybrid use of user experiment and transaction log analysis is proposed. The two research methods complement each other in that one is for a controlled lab environment with limited amount of data collected; the other is for a naturalistic setting with a large amount of data obtained. It is hoped that this approach will provide richness and depth while remaining practical for other OPAC studies or information retrieval system studies.
APPENDIX A: REFERENCES


APPENDIX B: SCHEDULE OF COMPLETION

The proposed time table for the dissertation is listed in the table below. The goal is to finish drafting the dissertation (without the user experiment results) on Aug 20 and to finish the user experiment on Oct 1. The rough draft ready to sent to my dissertation committee should be due on Nov 1.

Dissertation Schedule

<table>
<thead>
<tr>
<th>Task</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot study done</td>
<td>Aug 1, 2011</td>
</tr>
<tr>
<td>Dissertation preliminary draft (without the user experiment results, only with the transaction log analysis results) done</td>
<td>Aug 20, 2011</td>
</tr>
<tr>
<td>Start collecting data for the experiment</td>
<td>Aug 25, 2011</td>
</tr>
<tr>
<td>Experiment done</td>
<td>Oct 1, 2011</td>
</tr>
<tr>
<td>Draft dissertation send to committee</td>
<td>Nov 1, 2011</td>
</tr>
<tr>
<td>Defense</td>
<td>Nov or Dec, 2011</td>
</tr>
</tbody>
</table>
APPENDIX C: BUDGET JUSTIFICATION AND ADDITIONAL FORMS OF SUPPORT

For the duration of this study, I will be registered as a full-time student and will be supported by Dr. Brad Hemminger as a Graduate Research Assistant (50% FTE), a graduate student service appointment that is accompanied by a stipend and tuition waiver. For the specific activities described in this proposal, funds are requested to obtain the equipment necessary for the user experiment.

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject compensation(^1)</td>
<td>$30 \times 36 = $1080</td>
</tr>
<tr>
<td>Camtasia license(^2)</td>
<td>$179</td>
</tr>
<tr>
<td>Transcription kit (compatible with Microsoft Windows 7)</td>
<td>$200</td>
</tr>
<tr>
<td>Digital audio recorder</td>
<td>$40</td>
</tr>
<tr>
<td>Other hardware and software for data analysis</td>
<td>$0</td>
</tr>
<tr>
<td>Incidental expenses (photocopies, printing, workstation access, batteries, etc.)</td>
<td>$0</td>
</tr>
</tbody>
</table>

Total requested from the Thomson ISI Doctoral Dissertation Proposal Scholarship program $1499

\(^1\) Each subject will spend approximately 2 hours in the experiment. An honorarium of $30 will be given to them as compensation of the time devoted to the experiment. There are totally 36 subjects needed for this dissertation.

\(^2\) Camtasia is a software application for recording the computer screen. It is a standard tool for user experience studies. The education pricing for the license is $179.

DISSEMINATION ADVISOR

Brad Hemminger, Associate Professor
School of Information and Library Science
University of North Carolina at Chapel Hill
100 Manning Hall, 216 Lenoir Drive
Chapel Hill, NC 27599-3360
Phone: 919-966-2998
Email: bmh@ils.unc.edu