Cory Lown. A Transaction Log Analysis of NCSU's Faceted Navigation OPAC. A Master's Paper for the M.S. in L.S degree. April, 2008. 66 Pages. Advisor: Brad Hemminger

Since 2006 a small number of libraries have implemented faceted navigation on their online catalogs. This number seems likely to increase. Aside from small differences in some of the details of these interfaces, the implementations are remarkably similar, especially among academic libraries. A number of assumptions have gone into the design of the faceted catalogs. However, there is little empirical evidence available to suggest how people use text searching in combination with faceted navigation in a library catalog. This study reports the results of log analysis of the NCSU Endeca interface for searches conducted between January and April 2007. Findings from this exploratory study will be useful for designing user studies to answer additional questions about faceted navigation systems.

Headings:

Online catalogs / Evaluation

End-user searching / Evaluation

Online Catalogs

North Carolina State University / Libraries

Classification / Systems / Faceted

Endeca Technologies Inc.

A TRANSACTION LOG ANALYSIS OF NCSU'S FACETED NAVIGATION OPAC

by
Cory Lown

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2008

Approved by

_____

Brad Hemminger

Table of Contents

Introduction[1]

Faceted navigation is emerging as the latest trend for search and navigation on library Online Public Access Catalogs (OPACs). Traditional OPACs present to users one or more text boxes that enable searches against particular Machine-Readable Cataloging (MARC) record fields. Keyword, subject heading, author, and title are typical fields available for search. Users can form complex queries by combining Boolean operators across multiple search fields. Text based search works well for known-item queries (such as title), but not so well for browsing to discover useful material (Mat-Hassen, 917).

Faceted navigation premises that search and discovery are enhanced when the metadata (facets) that describe records are exposed as part of an interactive interface, allowing users to drill down to desired results. In a library setting facets may be divided into general categories such as: subject-topic, author, genre, format, location, subject-era, and other metadata groups available within the MARC record. Within each of these groupings users see a list of orthogonal categories that may be combined to form complex, Boolean "AND" queries without foreknowledge of how holdings are cataloged or an understanding of Boolean querying techniques.

Faceted search interfaces have emerged over the past five years on e-commerce websites such as Home Depot[2] and PC Connection[3]. In early 2006 libraries began to add

---

[1] Parts of the following section are modified from Cory Lown's research proposal that was written for INLS 780.
[2] See http://www.homedepot.com/
[3] See http://www.pcconnection.com/

faceted navigation to their OPACs. NCSU[4], McMaster University[5], and FCLA[6] are current examples.

Because faceted navigation catalogs are new, there is very little data or literature available to suggest that these catalogs actually improve the user experience over traditional OPACs that offer only text searching. Through transaction log analysis this study aims to begin to reveal the way users interact with faceted navigation systems by combining text and facet searching.

Literature Review

1.1 OPACS Are Hard to Use

Information seeking is an interactive and iterative process (Borgman, 568). This has implications for the way researchers approach studying and evaluating information retrieval tools and interfaces. Over the past twenty years, researchers have shifted from focusing on the product of a search input to the process of the search itself (Borgman, 571). Additionally, there has been a shift in focus from outcome measures of success (the precision and recall of a system's response to a particular query), to measures that focus on the search process and user perception (Borgman, 580).

There has also been a change in how researchers define the typical OPAC user. Traditionally, OPAC users were considered to be expert searchers, such as librarians. Library users would convey an information need to a trained librarian who would form a complex query to return relevant material to the user (Borgman, 568). The librarian approaches the catalog with rich knowledge of the collection, as well as an understanding

[4] http://www.lib.ncsu.edu/catalog/
[5] http://library.mcmaster.ca/
[6] http://catalog.fcla.edu/ux.jsp

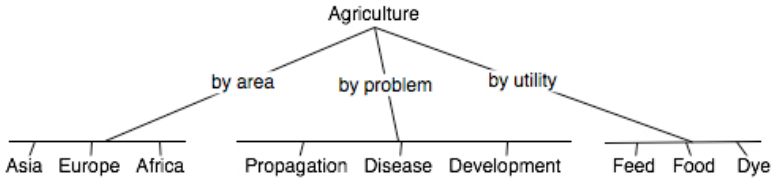
of cataloging standards. This specialized knowledge enables librarians to translate an information need into a query that will return relevant results. Presently, however, users expect to be able to search the catalog and explore library resources on their own. This is especially true now that many OPAC users are also frequent users of the Internet and search engines such as Google. Despite users' familiarity with search interfaces, most remain search novices with little incentive to become experts (Novotny, 529).

The literature abounds in problems users encounter as they use traditional OPAC systems. The most common problem is search failure, typically defined as a query that returns zero results (Yu, 169). Other problems include typographical errors and misspellings, use of uncontrolled vocabulary terms that do not match the controlled vocabulary of the system, incorrect use of search fields, and searches for items that are not in the catalog (Yu, 169). From their experiences using Internet search engines, users have come to expect to see the results of their searches ranked by relevance, even for poorly formatted queries. It is challenging for an OPAC system to meet the needs of expert as well as novice searchers.

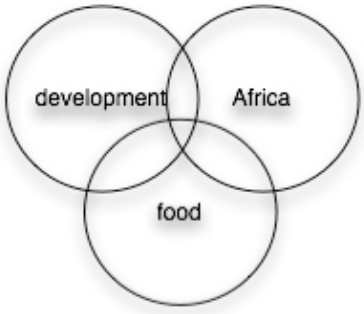
1.2 What are Facets?

Although facets are a popular topic (2008), as a concept they are not new. The most famous treatment of facets is in Ranganathan's "A Descriptive Account of Colon Classification" in which he describes the theory, utility, and practice of faceted classification as a means to organize human knowledge. Fundamentally, faceted classification enables items to be classified in multiple ways. This is in contrast to traditional taxonomies, where objects occupy a single location within a pre-determined hierarchy.

A clear illustration of this is from Ranganthan's own book, where he describes how one might use faceted analysis to describe the main group "Agriculture." Under agriculture there are several different characteristics that might be useful for describing an item. In Ranganathan's example one might classify items under "Agriculture" by area (e.g. Asia, Europe, Africa, etc.), by problem (e.g. Propagation, Disease, Development, etc.), or by utility (Feed, Food, Dye, etc.). Here is a reproduction of the illustration on page 30 of Ranganathan's book:



The utility of a system such as this is that one can locate items by identifying and combining multiple characteristics: "I want all items on Agriculture that relate to the area of Africa and have to do with the problem of development and food." This is particularly powerful in online retrieval as one can easily create complex set queries. The items that satisfy this set query are represented by the intersection of all three facets as depicted in a Venn diagram.



Colon Classification and Decimal Classification are complex when put to use in physical space, as one must decide where to place an item on the shelf. This limitation of faceted classification is overcome in virtual spaces such as the Web, however, as items can easily occupy any number of virtual locations. The capacity of faceted classification to organize items across multiple characteristics enables it to thrive in virtual environments.

1.3 Facets at NCSU

In a move to advance the OPAC beyond its current state, North Carolina State University (NCSU) looked to Endeca to provide search and navigation features on their OPAC. The Endeca platform is used by e-commerce websites such as Home Depot, Barnes and Noble, and PC Connection, among many others, to provide search and browse functionality to help customers find products. There are two hallmarks of the Endeca platform that seek to address problems users encounter in information retrieval. The first is relevance ranked results. Much like Google, the system attempts to place records most likely to be relevant to the user's query ahead of those that are less relevant. This differs from many OPACs, which rank retrieved records alphabetically or by date added to the system. The second feature is that the system provides, in the form of links, metadata associated with records in the system. The user is only shown metadata (facets) that apply to records currently in the result set. Because facets are always relevant to the result set, the user will never retrieve zero results by clicking on a facet[7].

The NCSU OPAC, which is based on the Endeca platform, makes use of the metadata available in the MARC records to display faceted navigation options to the user.

[7] It is still possible to return 0 results when executing a text search.

After the release of the catalog, the library completed basic assessment of their new system for finding library resources. The full write-up of their assessment can be found in Antelman's 2006 paper, "Toward a Twenty-First Century Library Catalog."

The team at NCSU used three methods to assess the performance of the Endeca faceted navigation catalog. They performed a log analysis of two months' worth of server logs and compared usage patterns between the old Web2 catalog and the new Endeca catalog. The logs revealed that authority searching decreased by 45% in the new catalog and keyword searching increased by 230%. They note that this change may have to do with the default behavior of the search box, which was changed from title to keyword. They also found that 55% are keyword searches, 30% include some refinement by using the facets, and 15% are browse-only searches.

NCSU also evaluated the relevance of search results returned by the old and new catalogs. One of the authors of the Antelman paper ran 100 topical queries in the new and old catalogs and coded the relevance of the results to the topic. Findings from this test revealed that 40% of the top results in the old catalog were relevant to the topic searched, while 68% of the top results in the new catalog were deemed relevant. This is a 70% improvement in performance.

The NCSU team also conducted exploratory usability studies to compare the old and new catalogs. They recruited ten undergraduate students at NCSU. Five were given a set of tasks on the old interface, and the other five were given the same tasks on the new Endeca interface. They measured task success, duration, and difficulty.

Except for one task, users completed their search tasks more quickly using the Endeca interface. The Antelman paper notes that, "the largest improvement is in the

increased percentage of tasks that are completed easily in Endeca and the nearly equivalent decrease in the percentage of tasks that were rated as hard to complete." It is noteworthy, however, that while failed tasks decreased with the Endeca catalog, many users still failed to complete their search task (22% failure rate for Endeca, 34% for the traditional OPAC). Also noteworthy is that users still encountered difficulty choosing the correct text field. They tended to choose keyword-subject (which searches only LCSH), over keyword-anywhere (which searches across all data fields) (Antelman, 135). The study notes that all participants who used the Endeca interface understood that the facets could be used to narrow results, but only three used the facets. None of them understood the LC Classification facets that appear above the result in the interface.

1.4 Transaction Log Analysis[8]

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. Transaction logs record a variety of information, but may include such information as: the date and time, the user's IP address, the Web page requested, any string the user entered into a search box, the parameters the user assigned to the search, the URL generated as a result of the query submission, and the number of items returned by the query, among other information.

Transaction log analysis (TLA) is the examination of transaction logs as a means to better understand an IR system, its users, and the interaction between user and system. In his 1993 examination of the history of transaction log analysis, Thomas Peters

[8] This section of the paper is modified from Cory Lown's final paper for INLS 500.

identifies a 1967 study by Meister and Sullivan as one of the earliest examples of TLA. His overview of TLA divides the history of the practice into three major phases. In the 1960s and 1970s, which represent the first phase, transaction logs were used primarily to evaluate and understand system performance. The focus shifted in the late 1970s and the mid-1980s to the examination of the search behavior of users and how systems were being used. The third phase, which Peters identifies as the mid-1980s to the early 1990s, was notable for the diverse use of transaction logs and replications of previous studies. Since the burgeoning of the Web from the early 1990s through the present (2008), TLA has gained importance as one means of practicing Web analytics, which attempts to measure the behavior of users as they interact with websites and information systems. This is of interest to libraries with online public access catalogs (OPACs), e-commerce businesses, and nearly any organization that cares about whether users can effectively interact with their websites. TLA can provide information to aid decision-making in system improvement and user instruction. In general, TLA has expanded from a focus on system monitoring to include the study of human computer interaction.

One major benefit of transaction log analysis, and the reason why attention has been paid to it, is that transaction logs are usually generated as a matter of course by information systems and servers. The question is how to best make use of transaction logs. Another related benefit is that the persistence of log files makes it possible to study and track a system and its users over a long period of time (Peters, 1993). Transaction logs also overcome one obstacle of most human computer interaction studies by providing researchers with unobtrusive observations, as most users are either unaware of or not concerned with the fact that the server records their actions. Additionally, the data

in search logs contain information about all users of a system, rather than the much smaller samples typical of observations or questionnaires (Hert, 1997).

However, like many methodologies, TLA has a number of limitations. Transaction logs provide the researcher with no information about what a user felt or thought while they were interacting with the system. There is also little way to know, other than inferring from their search terms and general behavior, what it was they were looking for or what information problem they were trying to solve. Search logs also do not provide an easy way to identify individual users, as an IP address might belong to a lab computer shared by many users. It is also difficult to determine what constitutes a discrete search session. Each researcher must decide how much time of inactivity denotes a new session, and whether a search session can span multiple days or take place from more than one computer. Determining successful or unsuccessful use of the system may also be difficult. Users cannot be asked whether they found what they wanted. Additionally, the logs examined in this paper do not record whether a user clicked on a returned record to examine it more closely, an action that might indicate a measure of success of the search. Furthermore, the amount of data in a transaction log can be overwhelming and difficult to process. For the month of January alone, the NCSU transaction log in plain text format was over 70MB and contained 218,083 lines of text, each representing a separate transaction. Consequently, it is generally necessary to analyze logs programmatically. Despite these limitations, TLA provides useful clues about user interactions with the faceted navigation OPAC system. Many researchers have combined transaction log analysis with other forms of observation, and a future study of the NCSU OPAC might combine TLA with user observation.

Methodology

1.5 Overview and Purpose

The purpose of this study is to process and analyze the transaction logs generated by all human interaction with the NCSU faceted OPAC from January through April of 2007. A variety of methods will be used to make the data contained in the logs reveal how users are combining text and facet searching. Some analysis will be done manually, which will provide a basis for developing automated tools for a complete analysis of four months of data. The manual analysis will also provide a baseline for verifying the accuracy of the automated tools. For automating the log analysis, a series of perl scripts will be written to parse, annotate, and code the information in the logs. Finally, the statistical package SAS will be used for running analyses on the processed log data.

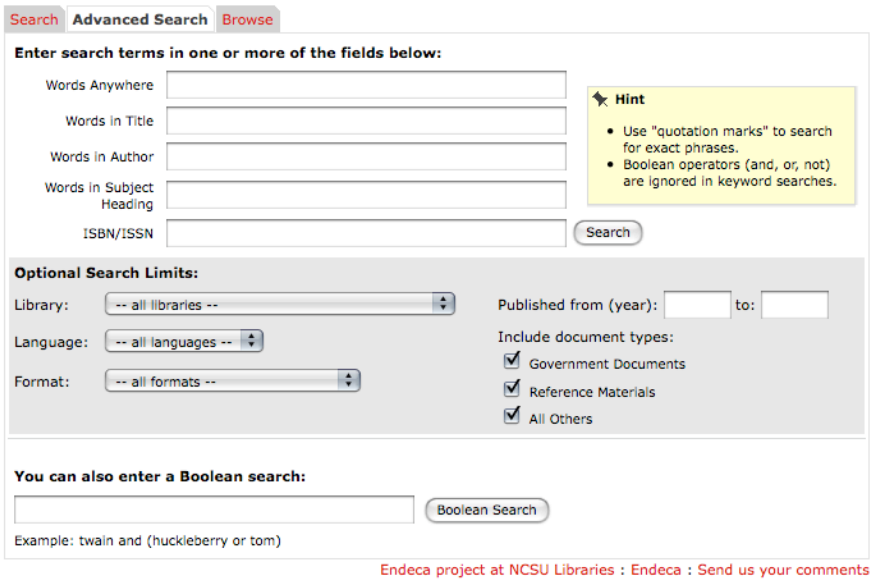
1.6 The NCSU Endeca OPAC Interface

Users who navigate to the NCSU libraries catalog through the main NSCU home page arrive at a search page that should appear familiar to anyone who has used a web-based OPAC.

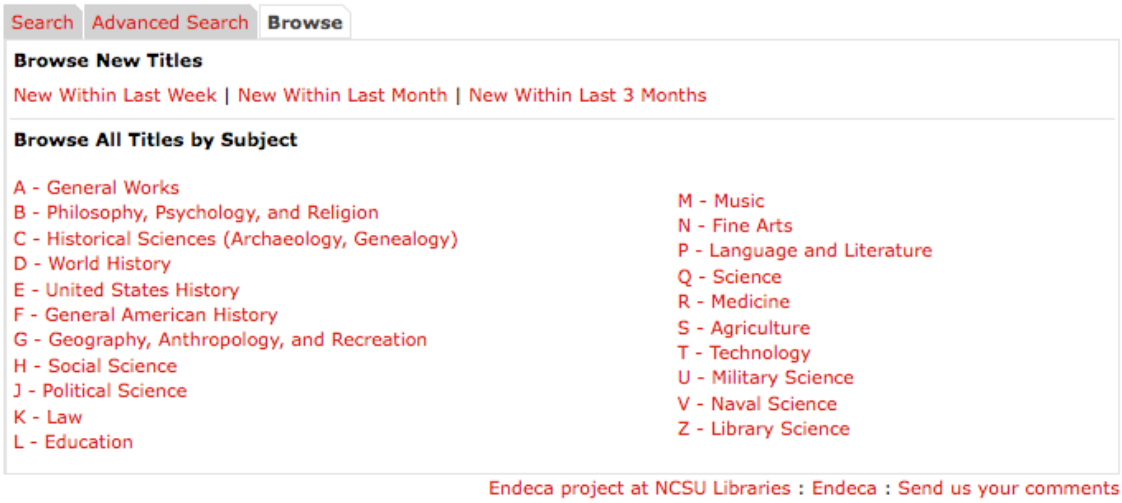
The NCSU Catalog search page.

There are three search boxes on the Catalog page. The first one labeled "QuickSearch" searches across the NCSU library website, not just the catalog. The second search box on the page with the heading "Search for words:" searches the Endeca based catalog. It gives the user the option of searching "Anywhere," "in Title," "in Author," "in Subject Heading," or "ISBN/ISSN." The search box below that one, which is labeled, "Search begins with…," takes the user to the old catalog system (not Endeca-based). The tabs across the top of the search box region of the web page give the user the option to switch to different search modes. "Advanced Search" looks much like many advanced search pages offered by other OPAC systems. The browse page displays Call Number Range

facets, giving the user the option of entering into the Endeca based catalog by first refining by Call Number Range.



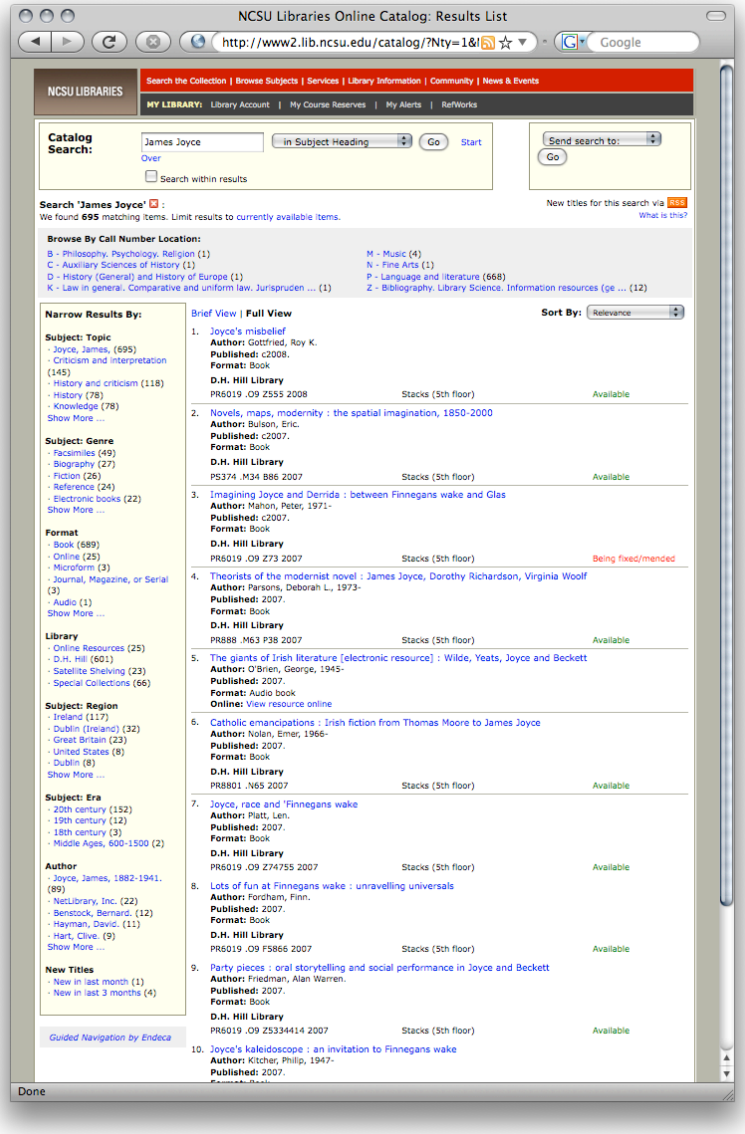
Options available on the Advanced Search Page.



Call Number Range facets available from the Browse page.

Whichever of these options the user chooses to use to begin searching, the next page the user sees is quite different from most OPACs. The page of results for the user's initial

query is accompanied by options for refining the current search. To illustrate, a search "Anywhere" on the phrase "James Joyce" yields the following page of results.



Results page for "James Joyce" search.

The search term appears below the search box as "Search 'James Joyce'," with a red "X" icon. Clicking the "X" removes that search parameter. Below the list of current search parameters is a horizontal grey rectangle labeled, "Browse By Call Number." This box contains a list of Call Number Ranges that apply to the current result set. That is, at least

one item in the results list will be found within each of displayed call number ranges. It is important to note that if the user were to execute a text search from the results page, all previous text search parameters and any facets would be removed, effectively restarting the search. Along the left of the results page, oriented vertically, is another rectangle with an off-white background. This box contains the rest of the facets available for use by the user to refine the current result set. In this case as well, each of the displayed facets applies to at least one of the items in the results list. Clicking on a facet to refine results will never lead to 0 results, because the facets are always relevant to the current set. The facet groups available along the left of the page are: "Subject: topic," "Subject: genre," "Format," "Library," "Subject: Region," "Subject: Era," "Author," and "New Titles." Each of these facets and facet groups is derived from data stored in the MARC record. The Subject Headings, however, have been atomized and displayed to the user as facets. It is notable that Subject Headings were never intended to be used in a faceted navigation interface.

1.7 Capturing Usage Information

Endeca includes built-in reporting software that tracks term and facet usage. The built-in reporting software is convenient because it automatically reports usage of the catalog. However, it lacks the ability to track a series of actions from a user, and so has no sense of tracking statistics by sessions. Because the reporting software is stateless, it gives a broad picture of how the Endeca navigation and search is being used. However, there is no information about what a typical series of actions from a user might look like.

The Endeca reporting software is not the only repository for information about user activity on the OPAC. There are two sets of server logs that also record information

about the activity of users. One is the dgraph log, which records requests received by the Endeca server that were sent from the application server. This log has the advantage of including all information about each request encoded in a URL. It also includes requests to view particular records. The problem with this log is that it records the IP address of the Web application server rather than the IP address of the end user. The Web server log does record the IP address of the end user, and also includes all the details of each request (the search term and any facets included in a search request). However, the Web server log does not record any requests to view particular records.

Four months of activity recorded in the Web application server between January and April 2007 were used for the log analysis study. This enabled an analysis of catalog use from a session perspective, as this was the only log that recorded the necessary information. The tradeoff is that this set of logs does not include information about detailed record requests.

The Web application server records each request received by the server. Server logs are reasonably standardized, with some variability depending on the configuration of the server. Each line of the log includes the following information: host, date and time, request, status code, bytes sent, referrer, and user agent. For instance, the first entry in the January 2007 Web application server log looks like this (on a single line):

> www2_search-access.log.1169596800: 24.124.228.163 www2.lib.ncsu.edu - [24/Jan/2007:00:03:04 -0500] "GET /catalog/?N=0&Nty=1&Ntk=Author&view=full&Ntt=pettersmann HTTP/1.1" 200 23888 "http://www2.lib.ncsu.edu/catalog/?N=206417" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.9) Gecko/20061206 Firefox/1.5.0.9"

Host	www2_search-access.log.1169596800: 24.124.228.163 www2.lib.ncsu.ed
Date and time	[24/Jan/2007:00:03:04 -0500]
Request	"GET /catalog/?N=0&Nty=1&Ntk=Author&view=full&Ntt=pettersmann HTTP/1.1"
Status code	200
Bytes sent	23888
Referring URL	"http://www2.lib.ncsu.edu/catalog/?N=206417"
User agent	"Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.0.9) Gecko/20061206 Firefox/1.5.0.9"

Single line of the server log parsed into components.

Of primary interest is the IP address of the remote user (in this case, 24.124.228.163), the date and time stamp ([24/Jan/2007:00:03:04 -0500]), the URL component of the request (/catalog/?N=0&Nty=1&Ntk=Author&view=full&Ntt=pettersmann), and the referrer (http://www2.lib.ncsu.edu/catalog/?N=206417).

The request also contains URL parameters that encode the user's search request. "N" refers to "Navigation" and sets the unique ID of any dimension (facet value) selected by the user. If "N=0", as in the case of the example, there are no facet refinements as part of the search request. "No" stands for "Record Offset", and is a pagination method. For instance, if "No=20" the system will return records beginning at the 21st record. If "No" is not defined in the request, the system will return records starting at the 1st record. "Ntk" stands for "Record Search Key" and defines the record field that will be searched for any text strings defined in the query. In this case, "Ntk=Author", which means the system will search for the term "pettersman" in the "Author" field. Multiple search keys can be combined with a pipe operator, "|". "Ntt" stands for "Record Search Terms" and holds the string of text, if any, that the user has entered as part of her search. In the

example "Ntt" is set to "pettersman", so the system would use that text string as a search parameter. Multiple terms are combined with the plus operator, "+". Other parameters include: "sort", which enables the user to override the default sort order; "view", which alters the format of the results between a longer listing and a shorter listing with less information; and "Ne", which indicates a request to expose additional facets within a particular group. The complete list is truncated to conserve space.

Although requests are recorded in the log sequentially, and not sorted by the IP address of the request, one can track requests from a single client by examining requests from the same IP address. This can be done manually using a text editor and reordering the lines of the log file by host and then by date. Viewed as a series, the URLs of each request from a single IP address can be used to recreate a set of actions from a single session.

Though a series of requests may come from a single IP address, there is no guarantee that this series of requests comes from a single user. There is also no way to know for certain how many discrete information searches are contained within a series of requests. A user could, in a series of requests, be actively seeking to satisfy multiple information needs. Because the logs are so far removed from the person making the requests, it is impossible to know for sure how each request is linked. Additionally, web browsers cache information so that the user can revisit pages without making requests from the server. A user may use the browser's back button to return to a previous search state; depending on their browser's configuration this action would not be recorded in the server log because no request was made by the browser.

For the purposes of this study, a "session" is defined as a series of requests from a single IP address with not more than 30 minutes passing between individual requests. If more than 30 minutes passes between requests from that IP address, the next request from that address marks the beginning of a separate session. Thirty minutes was chosen as a cut off time after surveying the literature. There is no consensus as to what constitutes the most effective cut off time for distinguishing sessions. In general, researchers have chosen times between 5 and 60 minutes. Thirty minutes was chosen for this study because it falls near the center of the range of times used in studies that employ TLA. For further discussion of this issue see Silverstein (1999), Göker (2000), Hert (1997), Marchionini (2002), Mat-Hassan (2005) and Chau (2005). This study makes no attempt to track a single user across sessions for several reasons. First, lab computers have many users. Additionally, many people share their personal computers with others. Finally, IP addresses are not static; service providers may assign a different IP address to the same computer each time it connects to the Internet.

"Dwell time" refers to the difference in time between two sequential requests within a single session. It is impossible to know what a user was doing between requests. They might have been interrupted by a phone call, talking with a friend, or any number of other activities aside from their interaction with the OPAC. In the aggregate, however, knowing the dwell time is still useful. First, it is useful as a means to divide sessions as described previously. Also, because of the large number of sessions analyzed, excessive dwell time because of extraneous activity should not have a large effect on the general trends reported.

An "action" refers to a user's interaction with the system. There are a finite number of things that the user can manipulate and do when interacting with the system. Additionally, in most cases, a request represents a single interaction. This is because the page reloads each time the user sets another search parameter or reorganizes the results on the page.

Related to "actions" are the codes used to indicate generic categories of requests that users can make of the OPAC. The codes relate to broad categories of actions, not the particular facet or facet group the user chooses, but the mere act of clicking on a facet; not the precise search term or field the user utilized, but the act of running a text search. Because the current search state and all its parameters are encoded (or, in some cases, implied by absence) in the URL, one can determine the action the user took to transition from one state to another by comparing what has changed in the URL between sequential actions within a session. For instance, given the following two sequential URL requests one can determine what action the user took to arrive at the second state from the first state:

First state: /catalog/?N=0&Nty=1&view=full&Ntk=Keyword&Ntt=photography
Second state: /catalog/?view=full&Ntt=photography&Ntk=Keyword&N=201015&Nty=1

First State	**Second State**
N=0	N=201015
Nty=1	Nty=1
view=full	view=full
Ntk=Keyword	Ntk=Keyword
Ntt=photography	Ntt=photography

Comparing search states to determine user action.

All parameters of the search are the same except for the "N=" value, which in the first state is "N=0" and in the second state is "N=201015". "201015" is the unique ID assigned

to a particular facet. In this case, "201015" is the ID of the facet "A – General Works," which appears within the facet group "Browse by Call Number Location." By comparing sequential search states within this session, it can be determined that the user chose the "A-General Works" facet to go from the first search state to the second search state. This user action is coded as "Facet Search," because that is the only action the user could have taken to go from the first search state to the second search state.

There are 12 possible categories of actions a user can take when interacting with the Endeca based OPAC that can be determined by comparing differences in URL parameters between sequential search states. The following table outlines the codes, what they mean, and the logical rules used to determine them.

Coded Action	Explanation of user action	Logical rules
Text_Search	First appearance of a search string within a session	Ntt has a value. Value is different from previous Ntt value. Ntt value has not appeared previously in the same session. If multiple values have changed, Ntt change takes precedence.
Facet_Search	Refines the result set by selecting a facet	N has a value other than "0". N value is different from previous N value, or N has an additional value that was not present in the previous N value. If multiple values have changed, takes precedence over all but a change in Ntt value.
Beg_Text_Facet_Search	Begins a search with a facet and a text string	First step in a session. N has a value other than "0" AND Ntt has a value.
Beg_Full_Set	Begins a search by selecting search without entering a search string; this returns all possible results	First step in a session. N has a value of "0" AND Ntt is not present in the URL or has a value of "-".
Refresh	No change in search state; suspect user reloaded/refreshed the page	Not the first step in a session. Present and previous URL are exactly the same.
Switch_Field	Switches field searched	If none of the previous conditions are met, Ntk has a value, and that value is different from the previous Ntk value.
Next_Page	Views a different page of results	If none of the previous conditions are met, No has a value and that value is different from the previous No value.
Sort	Changes the sorting of the results	If none of the previous conditions are met, sort has a value, and that value is different from the previous sort value.
Switch_View	Switches record view from brief to full or vice versa	If none of the previous conditions are met, view has a value, and that value is different from the previous view value.
Previous_Term	Searches on a term again	Ntt has a value. Ntt value. Ntt value has appeared previously in the same session.
Remove_Facet	Removes a facet from the search (either by clicking the "X" in the interface, or by clicking the "back" button in the browser)	Determined by counting the number of concatenated facet ids in N value. If present count is less than previous count, then this condition applies
Expand_Facet	Expands a facet group by clicking "Show More …"	None of the other conditions are met, Ne has a value, and that value is different from previous Ne value.

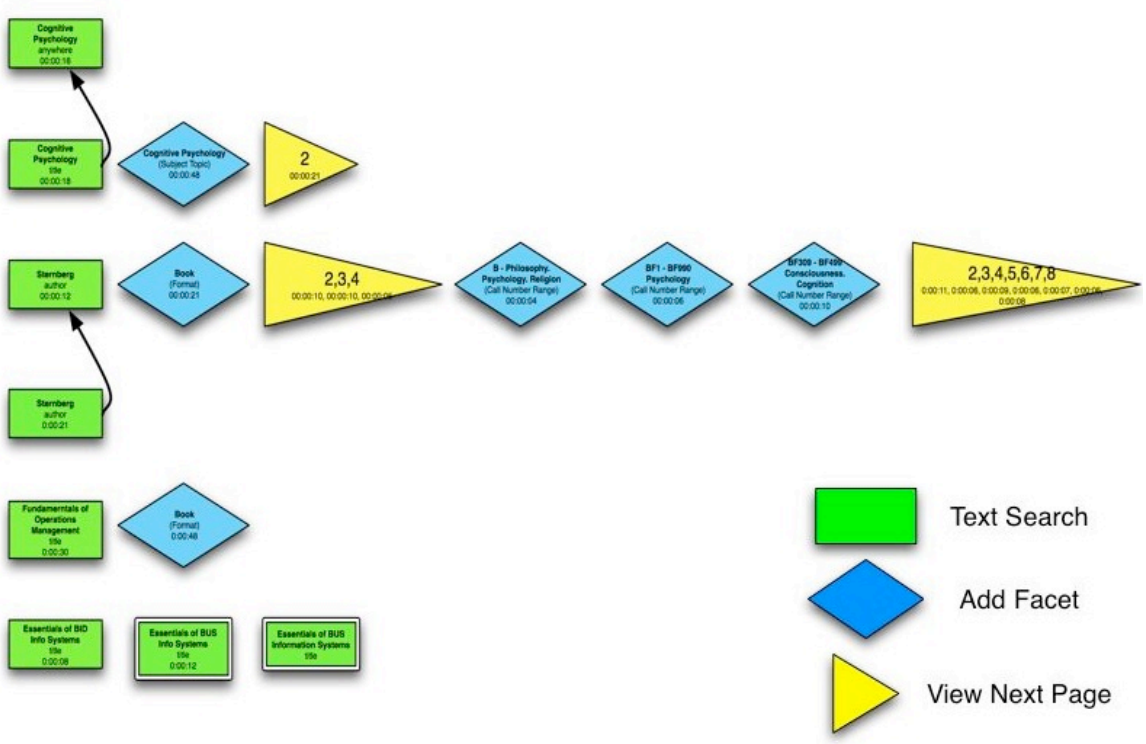
By dividing the record of each request in the log into separate sessions, and further, coding each action within each session, one gains an overview of the paths users follow as they interact with the OPAC.

1.8 Manual Analysis of Logs as Pilot

As a pilot exercise to analyze the effectiveness of the coding scheme, twenty sessions were extracted from the logs and coded by hand. This was done in two ways. First, differences between the URL encoded search parameters within each sequential request were examined. Additionally, each request in each session was entered into a browser. This effectively recreated each step of the user's search session in the browser. Analyzing the log files by hand is beneficial in a number of ways. It sets a baseline and means to check any automated analytical techniques that are developed later. Also, by recreating a number of sessions by hand one gets a much better sense of what the coded actions mean and how they affect what the user sees within the browser from action to action throughout a session. Twenty sessions were chosen randomly from the logs for analysis by hand. Manual analysis is, however, extraordinarily time consuming. Valuable as an initial exercise to gain insight into the data, manual analysis is impractical for analysis of more than a very small proportion of all the data stored in the log. An automated means to process and analyze the full data is necessary.

During this pilot phase of the study a method was developed for producing a graphical representation of action sequences within a session. Shapes and colors represent the different actions taken by the user. The following is an example of one session that

was examined as part of the manual log analysis and then transformed into a graphical representation:



Graphical representation of a session. Each shape represents an action by the user. Read from top to bottom, left to right.

1.9 Automated Log Analysis

The server log for January through April of 2007 recorded 938,848 requests, far too many to analyze by hand. Some form of automated analysis is necessary to process such a large number of requests. Server logs lend themselves to automated analysis because each record in the log records the same set of information. The challenge is correctly parsing the plain text logs into meaningful pieces of data. For this task Perl was chosen, with its strong implementation of regular expressions, to parse and process the logs. Regular expressions provide syntax for constructing complex sets of pattern

matching rules. With Perl, these matched patterns can be stored in variables, processed, transformed, compared, or any number of other operations available in the scripting language. This provides a powerful set of tools for extracting the rich information contained in the plain text log files.

Although the log processing could have been accomplished in a single script, four scripts were written to accomplish the entire process[9]. This enabled some degree of error checking between each phase of the processing. Also, as more was learned from the logs, additional processing was necessary and it was expedient to write additional scripts to accomplish this processing rather than rewriting the original scripts. The following is a summary of the processing tasks completed by each script.

The first script (see step_01.pl[10]) uses regular expressions to parse each meaningful section of data from each line into its own variable. The time date field is converted to Unix Epoch time to facilitate comparison operations. This script also extracts from the request URL specific parameters of the query. This includes any search term, text field, facet, view, page, sort, or facet group expansion that applied to the request. Some of these parameters may have carried over from a previous request, and some may represent default behavior of the catalog. Determining and coding the action the user took to generate each request will be handled in a later script. The first script also categorizes the referring page. The referring page is the page that generated the request. Referring pages were categorized as follows:

---

[9] The scripts used in this study were adapted from scripts described in Callendar's "Perl for Website Management."

[10] http://ils.unc.edu/neoref/lown_mp/step_01.pdf

Referring Page Category	URLs or Sites
Default Tab	http://www.lib.ncsu.edu/catalog/index.html
Advanced Tab	http://www.lib.ncsu.edu/catalog/advanced.html
Browse Tab	http://www.lib.ncsu.edu/catalog/browse.html
WWW2 Page	http://www2.lib.ncsu.edu/catalog/
Search Collection	http://www.lib.ncsu.edu/searchcollection/
Browse Subjects	http://www.lib.ncsu.edu/browsesubjects/
Main Library Page	http://www.lib.ncsu.edu/
Other Library Pages	http://www.lib.ncsu.edu/… (or other NCSU URL not caputured by rules above)
Library Main Search Box	http://www.lib.ncsu.edu/search/
Record Page	http://catalog.lib/ncsu.edu/web2/tramp2.exe
No Referring Page	-
External Search Tool	Wikipeda, WorldCat, OCLC FirstSearch, Google, Yahoo, Amazaon, Scirus, Ask
Other External Site	Anything not captured by rules above

List of referring pages and corresponding URLs.

The second script (see step_02.pl[11]) takes the output from the first script and does the work of tracking sessions. Recall that the log entries are stored in the order received and that HTTP is stateless. Although the server has no ability to track a series of requests from a single user, information stored in the log can be used to construct this data. The script works by storing the IP address and time stamp from each record in the log in an array. At each line the script checks whether that IP address has been stored previously. If it has been stored previously, the script calculates how much time has passed between this record and the previous record from the same IP address. If the amount of time passed between this record and the previous record is greater than 1800 seconds (30 minutes), then the script records this as a new session, gives it a unique session id, and marks this record as the first in the session. If the time passed between this record and the previous record is less than 1800 seconds, then it stores this record with the session ID of

[11] http://ils.unc.edu/neoref/lown_mp/step_02.pdf

the previous record from that IP address and records this as the nth step in the visit by adding 1 to the previous record's visit step. If the IP address has not been seen before, the record is automatically counted as a new session, given a unique session id, and counted as the first step in the visit. After this script is completed, every record in the log has been assigned a session id, which identifies the group of records to which the record belongs, and also a visit step, which identifies the sequence of that record within the session. Additionally, the script stores the time elapsed between each sequential action within each session.

The third script (step_03.pl[12]) takes the output of the second script and codes each action within each session with one of the codes defined previously. Coding is accomplished by storing each URL parameter in an array. The values stored in the array for the current record can be compared with the values stored in the array for the previous record within the same session. Following the rules established previously in the paper, the script assigns action codes by identifying the differences in states between requests.

The fourth script (step_04.pl[13]) was added later, when particular analyses required that the code from the previous step be stored with each request to facilitate analyzing transitions from one state to another. This script is relatively simple; it processes the output of the third script in such a way that each record also contains a record of the action code that was assigned to the previous request within the same session. The output from the fourth script is stored in a mySQL database.

[12] http://ils.unc.edu/neoref/lown_mp/step_03.pdf
[13] http://ils.unc.edu/neoref/lown_mp/step_04.pdf

1.10 Data Cleansing

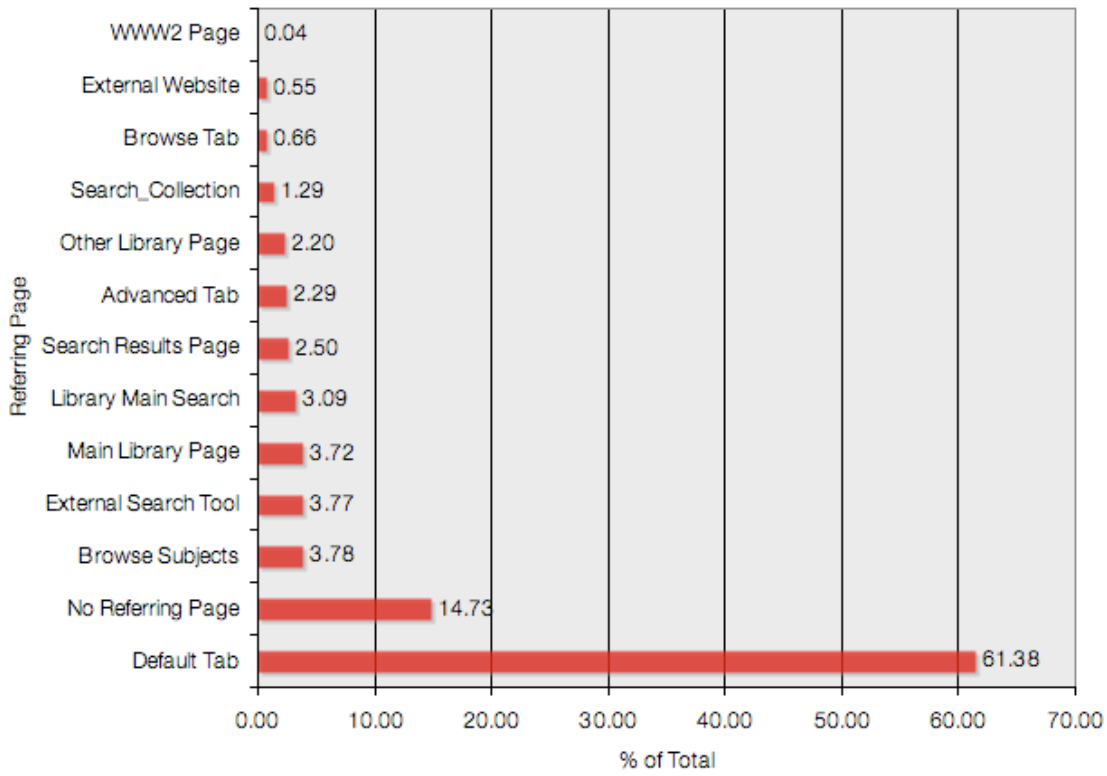
For the analysis of the data in the logs, it was important to be sure that the data collected from the logs reflected actions of real people interacting with the system. The log files, however, include any request received by the server. This can include web crawlers, automated processes that make requests of servers. The presence of web crawler activity in the logs poses problems for the research questions about how humans interact with the faceted navigation system. Consequently, several methods were employed to eliminate as much automated crawler activity from the logs as possible.

While it is impossible to eliminate all non-human requests from the log, there are a number of ways to minimize their presence by applying logical rules. A study by Jansen found that eliminating sessions with over 100 actions is an effective way to cull automated requests from logs. Although this method is imperfect, it minimizes the number of sessions from humans eliminated from the logs, and also removes most automated requests. Additionally, NCSU reported the IP address of a bot that had been actively crawling the site. All requests from this IP address were excluded. Surely some automated requests are still present in the final data set, and some human requests were removed. However, in aggregate, these methods eliminate as much non-human activity as possible from the data. After removing automated crawler activity from the logs, 130,482 sessions were left for analysis.

Results

1.11 The Search Begins

Let's start where the users start, at their entry point into the catalog. The entry point is determined by looking to the referring page of the first request (not the URL of the request, but the page that was used to produce the request). Sixty-one percent of sessions begin from the Default search page. This is the page one arrives at by default if one follows links to the catalog from the library home page. The Default tab provides two search boxes, as well as the ability to constrain the search to specific fields by selecting options from the drop down menu. Only the first text box takes one to the results page of the Endeca based catalog. On the same page, there are tabs across the interface that will take the searcher to the Advanced Search page or the Browse Subjects page. Neither option is a frequent starting place for searchers. Fewer than 4% of searchers begin their search from the Browse Subject page, and just over 2% of searchers begin from the Advanced Search page. There are a variety of other places for users to start their search, summarized in the table below. Most users start somewhere within the NCSU library Web pages, although some are referred to the catalog from external Web sites or search tools. See the discussion in the methodology section for a detailed explanation of the page names and their associated URLs.



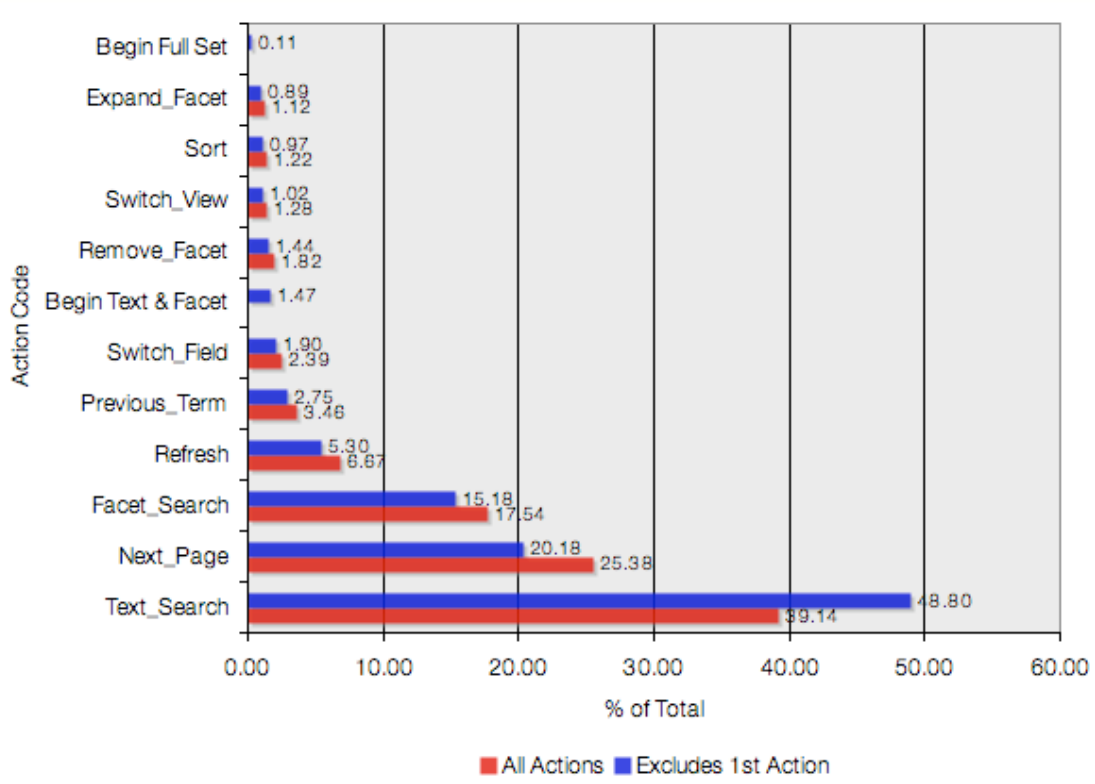
Frequency that sessions begin from a particular page.

It is worth noting that nearly 15% of searches start with a request for which there is no referring page recorded by the server. There is no ready explanation for this, and it may affect the statistics presented here, since 15% of starting points are unknown. Very few searches begin from the faceted interface, so it is unclear how users would choose to begin their search if they were presented with this interface from the beginning of their search. The page from which they start constrains their options for beginning a search. For instance, the Default tab, where over 61% of users begin their search, allows only a text search. No facets are presented as options on that page.

1.12 General Session Statistics

Overall, sessions are short, although there is wide variation in session length and number of actions per session, which affects how the statistics should be interpreted. The median session consists of 2 requests and lasts about 45 seconds, with 22 seconds passing between each request. The means are higher because of the distribution of the data. On average, sessions consist of 5 actions and last 6.5 minutes, with 1.5 minutes passing between each request. The median provides a better picture of central tendency in this case, because it compensates for the extreme outliers present in the data that draw the averages higher.

The codes assigned to each request provide a means to show quantitatively the kinds of requests users make of the system. The chart on the next page summarizes the frequency of occurrence of each coded action.

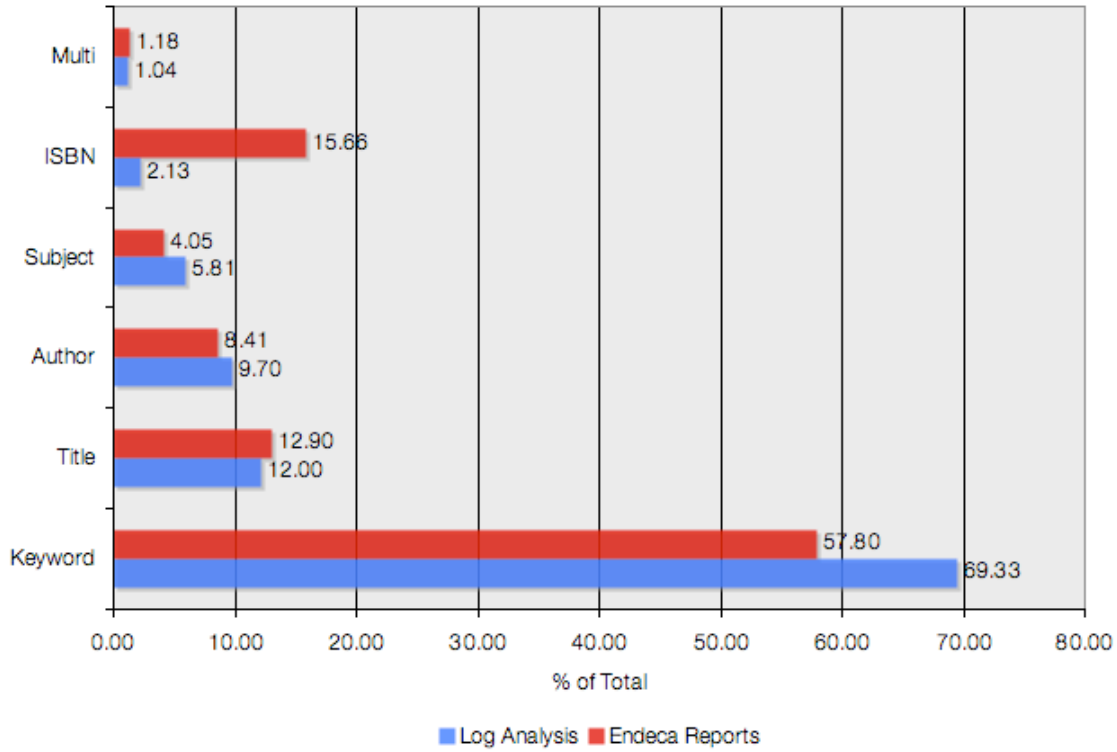


Shows overall frequency of particular actions codes. Red bars show all actions. Blue bars exclude the first action from all sessions to compensate for entrance pages to the catalog that offer only a text search option.

The blue bars exclude the first request from each session. (Sessions with only a single step would be excluded altogether.) The red bars include all requests. It is important to look at both numbers because the most common starting point of searches (the Default tab) does not allow any other action but text searching, and so it biases the results. It is unclear whether users would use text searching less and facet searching more if they were given the option from the outset, but this chart does indicate that text searching decreases after the first request in a session while facet searching increases. In general, most requests are text searches, accounting for about 39-49% of all requests. Page views are also frequent, at about 20-25%, and facet searching accounts for 15-18% of all requests. These three actions (text searching, page views, and facet searching), account for most of

the activity on the OPAC. Each of the other action codes account for fewer than 5% of requests.

Text searching is the most common action on the site. When searching on a text string, users have the option of choosing which field they wish to search against. The default setting is Keyword, which searches across all available fields. The following chart summarizes the frequency of use of each text field option.



Shows relative usage of different search fields for text searching. Results from Endeca reporting tool are shown in red. Log analysis results are shown in blue.

Results from the log analysis are shown in blue, while results for the same period from the Endeca reporting system are shown in red. The differences illustrate the value of removing web crawler activity from the logs. By removing web crawler activity one gains a more accurate representation of text field usage. Keyword searching actually

occurs more often than indicated by the Endeca reports (69% versus 58%), while ISBN searching occurs less often than indicated by the Endeca reports (2% versus 16%).

Users search differently depending on which text field they are using. In a 2000 study of Web searching behavior, Jansen found that searchers use an average of 2.21 terms per query when using a Web search engine. The results from the log analysis of the OPAC are similar.

Text field	**Average # of terms**
Title	3.01
Multiple Fields	2.87
Keyword	2.51
Subject	1.97
Author	1.90
ISBN	1.09

Average terms used depending on text field selection.

When using the default (keyword searching), searchers use an average of 2.51 terms in their search. The title field is used with the most number of terms (3.01), while ISBN searching uses a small number (1.09). The results make sense intuitively. Titles, in general, comprise several terms, while authors frequently have two terms. For keyword searching the significance is that users are utilizing the search box on the OPAC in much the same way they approach the search box on a Web search engine; searchers use a small number of terms in their query.

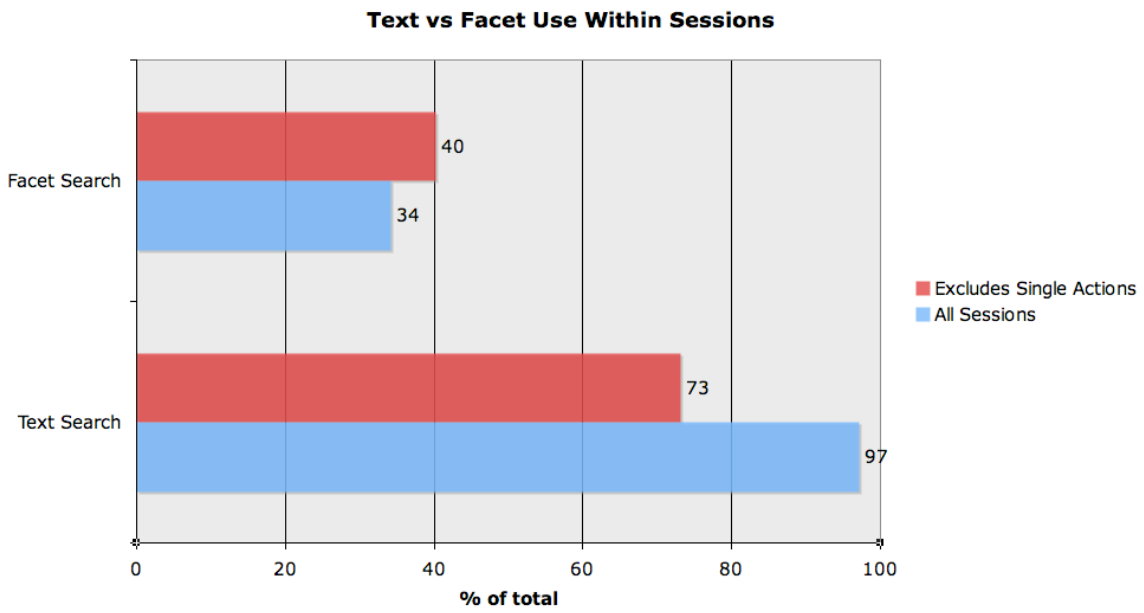
When users add facets to their query to narrow their search, how many facets do they use? Among requests that include at least one facet, 58% include just a single facet, 23% include 2 facets, 12% include 3 facets, and 6% include 4 facets. Very few requests include more than 4 facets, presumably because four facets would significantly reduce

the result set; there would not be much value in adding additional facets. On average, among requests that include at least one facet, there are 1.66 facets per request.

1.13 Facet and Text Searching

The real novelty of the Endeca based OPAC is the addition of facets as a means to narrow results. It is the combination of text searching and facets that give users great control over searching and the results of their search. How do they use each of these functions? How do they use them in combination? The statistics generated from the log analysis are at least a starting point for understanding users' search behavior in this mixed search and faceted refinement environment.

As described previously, most requests are text searches, accounting for about 39-49% of all requests. Facet searching accounts for 15-18% of all requests. There is another way of looking at this data, though, and that is counting sessions where facet searching or text searching appear at least once. This produces a somewhat different view.



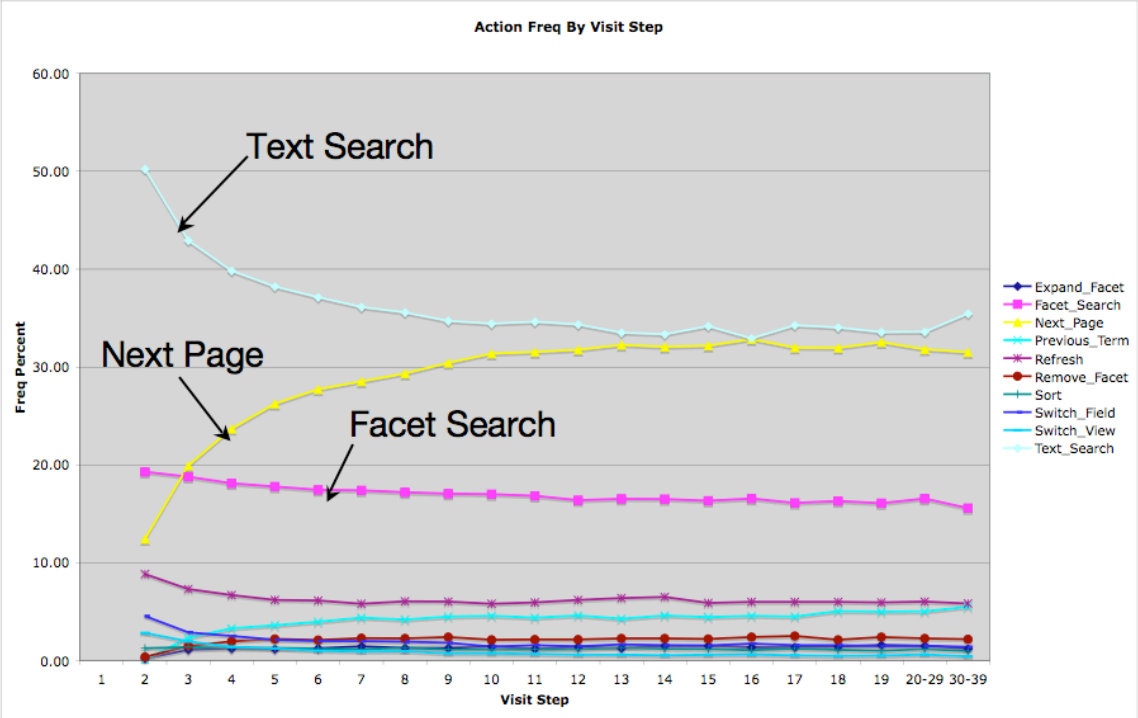
Shows occurrences of facet and text searching within sessions. Red bars exclude sessions with only a single action. Blue bars account for all sessions.

Among all sessions, 34% (44,278) include at least one facet search. Likewise, 97% (127,437) include at least one text search. Looking at it this way, about a third of all sessions make use of facet refinements, while nearly all sessions make use of text searching. This suggests that text searching is the primary mode of interaction with the catalog, while facet searching is supplemental, or only useful (or understood) by a smaller subset of users. However, when sessions that contain only a single action are excluded, the picture changes. In this case, 40% of all sessions include a facet search, and 73% of searches include a text search. This might occur because most users start from an interface that only allows text searching, drawing the total number of sessions that include text searching higher.

What does all this mean? It could be that the majority of information needs users have as they approach the catalog are best satisfied by text searching, and that facet

searching satisfies less common information seeking tasks. It is also possible that text searching is the search paradigm people are most used to, and so that is the strategy they tend to use when they approach the catalog.

It is also possible from the processed log data to look at the use of facet and text searching over the course of a session.

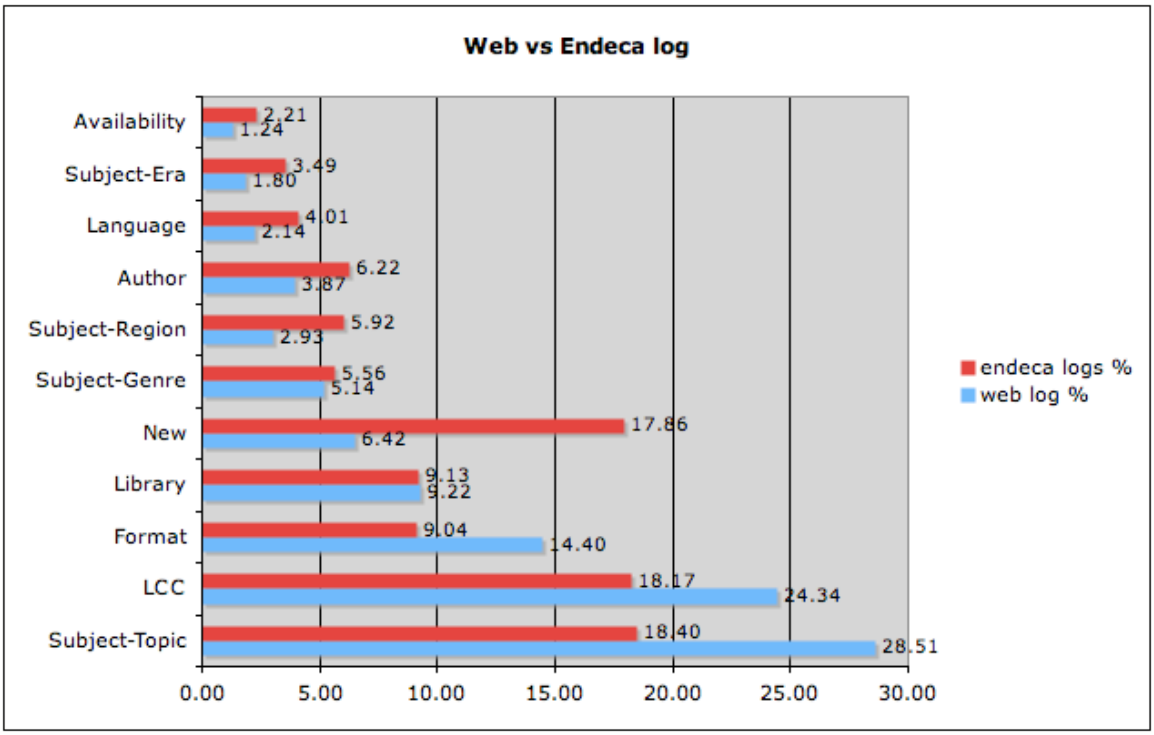


Shows relative frequency of actions over time.

The chart begins at the second step in a session. The codes for the first step are different from codes used later in the session, making comparisons between the first step and later steps impossible. It is notable that even at the second step of a session text searching is dominant, accounting for 50% of all requests while facet searching accounts for 20% of requests. Next page requests account for just over 10%. Over the course of a session, facet searching decreases slightly, stabilizing at about 17% by the 6th or 7th step within a

session. Next page views increase sharply, rising to about 32% by the 11th step of a session.

The facets available to the searcher to refine result sets are divided into categories. To recall from the explanation of the interface earlier in the paper, LCC Subject Heading facets appear horizontally, above the result list. The rest of the facets appear vertically, to the left of the list of results. The Endeca reporting software tracks usage of the different facet groups; however, it includes web crawler activity. In the chart below, the results generated from the log analysis are shown in blue. The Endeca-generated numbers are shown in red.



Shows relative usage of different facet groupings. Statistics from the Endeca reporting system are shown in red. Statistics generated from this log analysis study are shown in blue.

Subject-Topic and LCC are the two most frequently used facet groups, accounting for 24% and 29% of facet usage, respectively. It is significant to note that these are the facets

that are displayed at the top of the results page. They nearly always appear above the fold of the results page; the user would not have to scroll to see these, even on a relatively small screen. Format (14%) and Location (9%) are the next most frequently utilized facet groups. Availability (1%), Subject-Era (2%), and Language (2%), Author (4%), Genre (5%), and New (6%) are used least frequently.

The logs also make it simple to calculate the time passed between two actions within the same session, simply by calculating the difference between the time stamps on the two sequential requests. Although it is impossible to know for certain what other tasks users might have been involved in during the time between requests, because of the large number of samples, it is possible to make general claims about dwell times between tasks. This assumes that finding the averages across a large number of samples will minimize any outliers, where dwell times were increased by activities unrelated to interacting with the catalog. The average time between actions, if the first action[14] is a text search and the second action is also a text search, is 33 seconds. If the first action is a text search and the second action is a facet search, the average dwell time shrinks to 14 seconds. This is illustrated in the following chart.

[14] "First" and "second" do not, in this case, refer to the first and second action within a session, but the first and second action in any sequence of two actions. This is an important distinction.

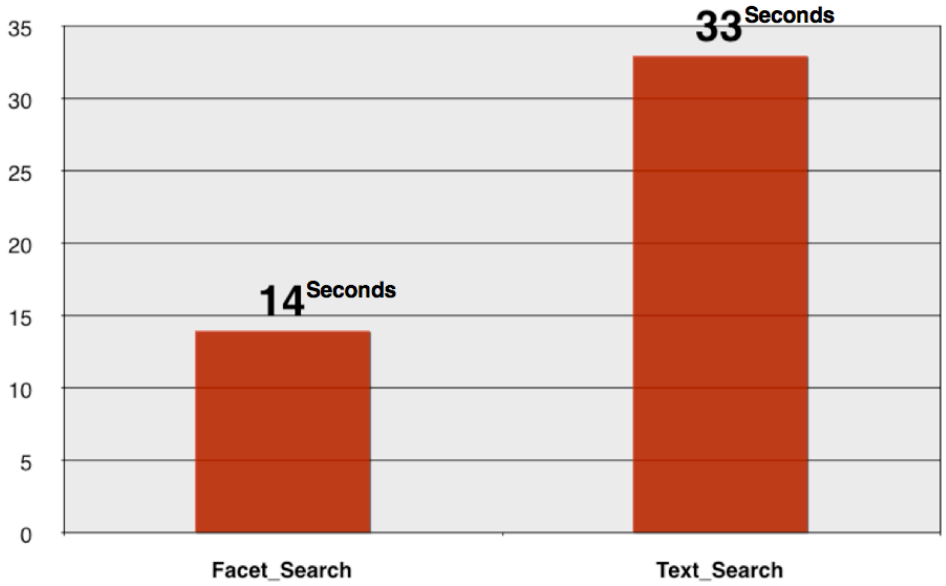
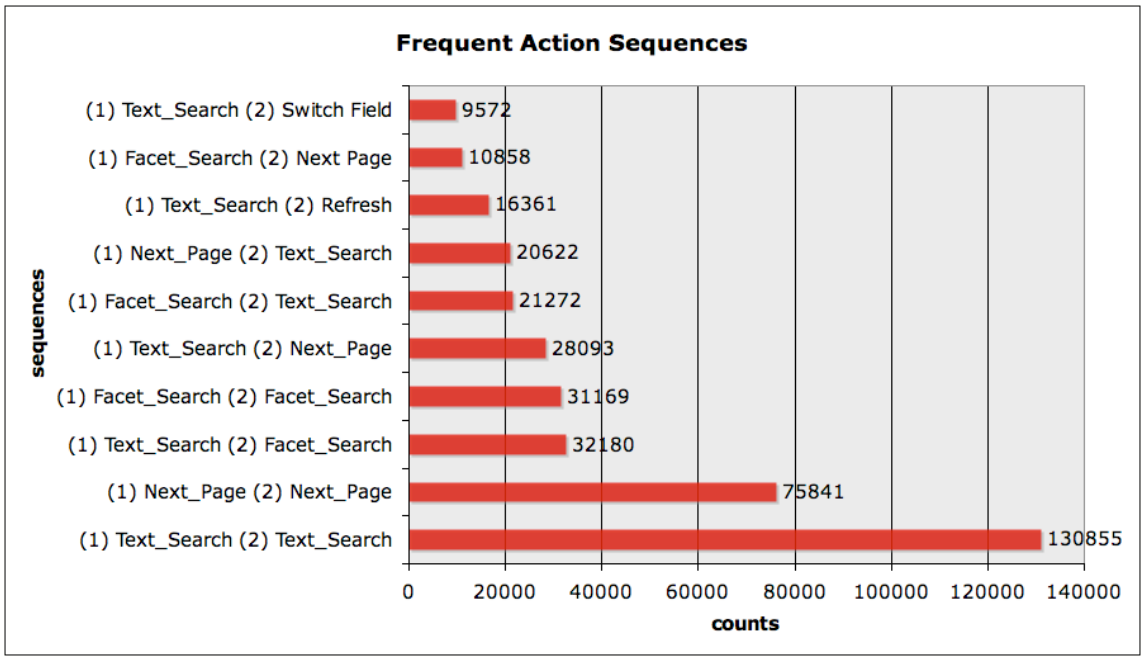


Chart shows transition from text searching to facet searching/text searching. Transition from facet searching to facet searching/text searching is nearly the same (13 seconds, 34 seconds).

These dwell times hold for the opposite case as well. If the first action is a facet search and the second action is a text search, the average dwell time is 34 seconds. Likewise, if the first action is a facet search and the second action is also a facet search, the dwell time is 13 seconds. This suggests, though does not prove, that facet searching is less expensive than text searching. It makes sense that it takes less time to recognize the utility of a particular facet and click on it than it does to formulate a query, type out that query, and execute it by clicking the search button.

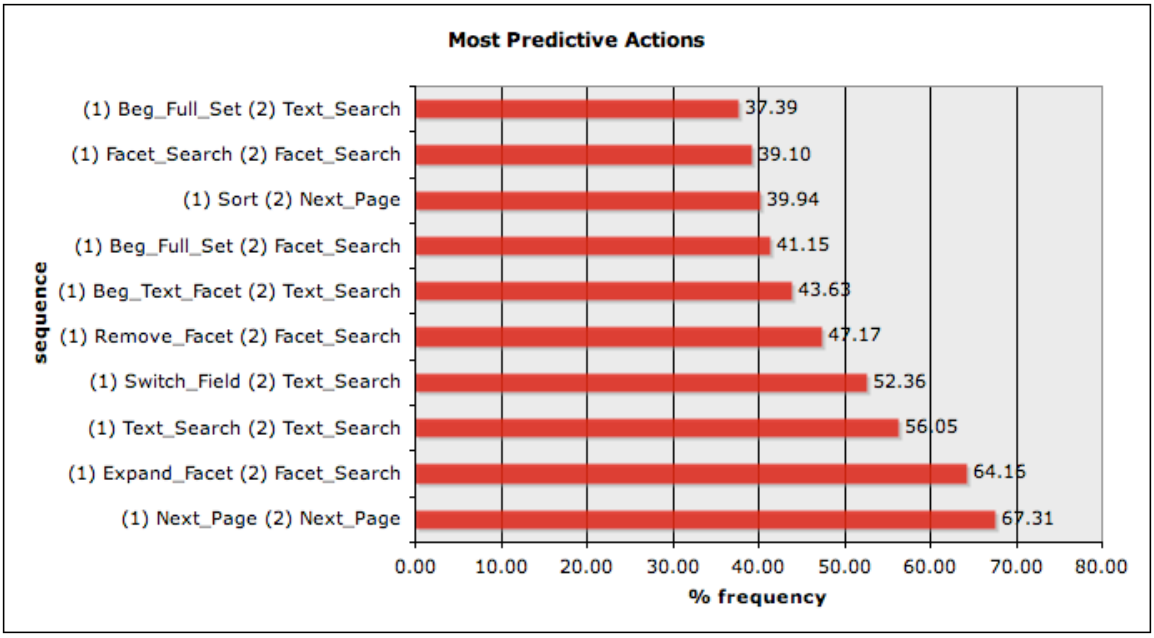
The processed logs also reveal frequent sequences of actions. The chart below shows the top ten most frequent sequences. The most common is text search to text search, which makes sense, since text searching is the most common action on the site. The next page to next page sequence is the second most frequent. The other actions occur less frequently, but generally show that users switch frequently between text searching

and facet searching. They also tend to perform several of the same actions in a sequence, two text searches in sequence, or, less frequently, two facet searches in a row.



Shows most frequent sequences of actions.

Rather than looking at frequency alone, one can calculate how likely it is that a user will transition from one state to another given all possible options. For instance, if the user has just decided to view another page of results (Next_Page), they will, 67% of time, follow that action with another page view. Additionally, if a user expands a facet group (Expand_Facet) they will, 64% of the time, follow this action with a facet search. Other predictable sequences are: a text search followed by another text search (56% probability), switching the text field followed by a text search (52% probability), and removing a facet followed by adding a facet (47%). The top 10 most predictable sequences are illustrated in the following chart.



Shows actions that are most predictable from prior action.

A few of these are easily explained, since there is a logical order in which certain tasks must be completed. In the case of the transition from expand facet to facet search, one must expand a facet grouping to see the full list of facets within that group. If one takes the trouble to expand a facet group, it seems logical that one has the intention of choosing a facet if there is one available that would help refine the current result set, hence the high predictability of this sequence.

Discussion

One of the striking findings from this study is that well over half of the sessions begin from the same place. Over 60% of sessions start from the Default search page, which presents to the user a simple text box with a drop down menu that can be used to specify different search fields. The default search field is "keyword" and it is not surprising that most users just type in a term or two and press search. This raises a

number of questions. First, should the user see one interface for beginning a search, and then another interface for viewing results and refining their search? The approach taken by NCSU, and many other faceted navigation systems, seems to be a compromise. This approach gives users a familiar Google-like search box to begin their searches, and then provides the more cluttered, but arguably more powerful faceted interface, later. There are some compelling reasons to take this approach. Informally talking with users reveals that they do find the faceted interface to be overwhelming at first glance. Additionally, many facet groupings do not make much sense until the result set is reduced somewhat.

However, it is possible that users would choose to interact with the catalog differently at the beginning of their search if they were faced with the faceted interface from the start. The start of a search is not an insignificant thing, as the log analysis shows that most sessions comprise just two actions, and many sessions comprise just a single action. Catalogs need to be designed from the perspective that most interactions are short. So, a key question then becomes, how can we design catalogs that can be used effectively for short interactions? Related to this is the question of whether there may be more than one logical user group that would be best served by multiple interfaces. For instance, most faceted catalogs display a single text search box interface that display facets only on the results page. A modified version of an advanced search page could be designed to expose a more complex interface. The advanced search page could expose the library classification system, allowing users to drill down the classification scheme, as well as providing traditional searching against individual metadata fields.

Despite the bias toward text searching that the interface introduces, the statistics produced from the log analysis do suggest that even without this bias, searchers tend to

use text searching primarily and facet searching less frequently. Even excluding the first step across sessions to eliminate the bias of the single search box on the default catalog page, 73% of sessions include text searching while 40% include facet searching. The reasons for this preference are not apparent from the log analysis, but there are a number of possible explanations. Text searching might, in fact, be the most effective search strategy for the most common information seeking tasks. For known-item searching, especially, it makes intuitive sense that users would prefer to type the title or other known parameter rather than try to locate the item within a faceted classification scheme. Facets tend to lend themselves more for exploration and discovery than for known-item searching. It may be that known-item searching is the most common task undertaken on the OPAC, which would explain the higher frequency of text searching. This does not mean the facets are unimportant, however, as they might provide enhanced searching for less common, but no less important, item discovery tasks.

It is also not surprising that the most commonly used facet groups, Subject-Topic (29%) and Call Number Range (24%), are the groups that appear most prominently in the interface. Under normal circumstances, users would not have to scroll to see either of these facet groups. The discovery and use of other facet groups requires scrolling for many typical screen resolutions. Do people use these two facet groups most frequently because they are the most useful, or because they are the most visible? Future research will have to address this question. It is worth noting, however, that the two most commonly used facet groups, while potentially being quite useful, pose challenges to users. Call Number Range and Subject-Topic both have the appearance of being subject related. They attempt to describe what items are about. However, Call Number Range is

based on the physical location of the item in the library, while Subject Topic is an atomized version of LC Subject Headings. Only the most savvy library users will understand the differences between these two prominently displayed facets. Additionally, LC Subject Headings were never intended to be used in a faceted search environment. The headings represent different levels of granularity, share many overlapping terms, and pose many challenges to users in a faceted environment. As an example, a keyword search for James Joyce also returns the following Subject-Topic facets (among others): English fiction, English literature, Fiction, American fiction, American literature, In literature, Literature, Literature Modern, Politics and literature. It is unclear whether these headings overlap, and even, in many cases, what the semantic difference is between items (English literature and English fiction, for instance). This is an area ripe for further research, and in fact, OCLC's FAST project is an attempt to adapt Library of Congress cataloging data for effective use in an online environment[15].

It is noteworthy that the two next most commonly used facet groups, Format (14%) and Library (9%) appear well below the fold on most common screen resolutions. The distinguishing feature of these facets is the ease with which they are understood. Most library users have no trouble distinguishing the difference between, "Book" and "Video and DVDs," two facets that appear within Format. It is clear what choosing one or the other will do: eliminate everything but books, or eliminate everything but videos and DVDs. Library is equally clear; the facet will limit the result to items available within the selected library or collection. These are both physical qualities of an item. "What it is" and "where it is" are characteristics that lend themselves readily to faceted systems,

[15] See http://www.oclc.org/research/projects/fast/

because their meaning is generally unambiguous. Other facets, especially, those related to subjects and topics, are more ambiguous and challenging both for users and designers of catalog systems.

Text searching and facet searching complement one another in the OPAC, although future research is necessary to determine the specific circumstances under which one or the other (or both) is an effective search strategy. What we can say from this study is that searchers spend about twice as much time before running a text search than before choosing a facet. It is likely that this difference is seen because it takes more time for users to formulate, type, and then execute a text search than it does for them to notice and select a facet. If this is the case then facet searching is less expensive in cognitive load and time than text searching. Although more research is necessary to prove this, if true, it makes sense to design the OPAC to encourage users to make use of facets when they would be as effective as or more effective than text searching.

Examining sequences of actions within the logs reveals that users switch readily between different kinds of activities. Although the most common sequences involve two or more of same action (text searching and page views being the most common), users also switch frequently between facet searching and text searching. When exploring ways to improve the OPAC interface, it should be taken into account that it must be simple for users to combine different search strategies and actions. Although text searching appears to be the primary activity, and should be supported through the interface, facet searching and viewing additional pages of records must also be easily accomplished.

Conclusions and Future Research

Despite challenges posed by faceted navigation OPACs to users, designers, and cataloging standards, such systems show much promise. Although there are problems with current implementations, catalog users do make use of facets when they search. It is a paradigm they are used to encountering in other online environments, such as e-commerce. Much work remains, however. Cataloging practices will need to be reexamined, or adapted to work better in a faceted environment. Though frustrating, it is probably a great virtue that facets tend to reveal flaws in metadata that otherwise would remain hidden. Studies are needed to determine how users combine text and facet searching, and under what circumstances one or the other or both are most useful and effective. Additionally, studies are needed to determine which facet groupings to display, and which to display most prominently on the interface. The most challenging question posed by faceted navigation systems is how to adapt or change cataloging standards for effective and intuitive use in an online environment for the novice, while not inhibiting professional librarians.

More broadly, the role of the OPAC is an open question. The OPAC is just one of many tools available in and beyond the library for locating information. Users can use Google and Amazon to locate items, and then open the OPAC to see whether the items are in the library's holdings. Better understanding and support for the way users really search and want to search are necessary. Additionally, library users are not just looking for books or other physical items located within the library walls, but also PDFs of journal articles and even eBooks, items not owned by the library, but to which the library provides access. These resources are often not well integrated into the library's web

presence. Users are bounced from the library pages and the OPAC to subscription databases with a wide variety of interfaces and capabilities. Simplifying and integrating access to all library resources, supporting the way users actually search for information, and providing better search tools with easy to use interfaces are challenging problems, to which there are few easy answers. Designing a better OPAC is a step in the right direction, but it is just one piece of what should be the question motivating all information professionals: how do we get the right information to the right user at the right time?

References

Antelman, K., et al. (2006) "Toward a Twenty-First Century Library Catalog." *Information Technology and Libraries*. 25(3) 128-139.

Borgman, C. (1996) "Rethinking Online Monitoring Methods for Information Retrieval Systems: From Search Product to Search Process." *Journal of American Society for Information Science*. 47(7) 568-583.

Borgman, C. (1996). "Why Are Online Catalogs Still Hard to Use?" *Journal of the American Society for Information Science*, 47(7), 493-503.

Callender, J. (2001) "Perl for Website Management." O'Reilly Media, Inc.

Chau, M., Fang, X., & Sheng, O. R. L. (2005). Analysis of the Query Logs of a Web Site Search Engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363-1376.

Göker, A., & He, D. (2000). Analysing Web Search Logs to Determine Session Boundaries for User-Oriented Learning. *Proceedings of Adaptive Hypermedia and Adaptive Web-Based Systems*, 319-322.

Hert, C. A., & Marchionini, G. (1997). "Seeking Statistical Information In Federal Websites: Users, Tasks, Strategies, and Design Recommendations." *Final Report to the Bureau of Labor Statistics*.

Jansen, B. J. (2006). "Search Log Analysis: What Is It; What's Been Done; How to Do It." *Library and Information Science Research*, 28(3), 407-432.

Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). "Defining a Session On Web Search Engines." *Journal of the American Society for Information Science and Technology*, 58(6), 862-871.

Mat-Hassan, M., et al. (2005) "Associating Search and Navigation Behavior through Log Analysis." *Journal of the American Society for Information Science and Technology.* 56(9), 913-934.

Marchionini, G. (2002). "Co-Evolution of User and Organizational Interfaces: A Longitudinal Case Study Of WWW Dissemination of National Statistics." *Journal of the American Society for Information Science and Technology*, 53(14), 1192-1209.

Novotny, E. (2004). "I Don't Think, I Click: A Protocol Analysis Study of Use of a Library Online Catalog in the Internet Age." College & Research Libraries, 65(6), 525–37.

Peters, T. A. (1993). "Transaction Log Analysis." *Library Hi Tech*, 11(2), 37-106.

Ranganathan, S.R. (1967). "A Descriptive Account of the Colon Classification." New York: Asia Publishing House.

Silverstein, C., Henzinger, M. R., Marais, H., & Moricz, M. (1999). "Analysis of a Very Large Web Search Engine Query Log." *SIGIR Forum*, 33(1), 6-12.

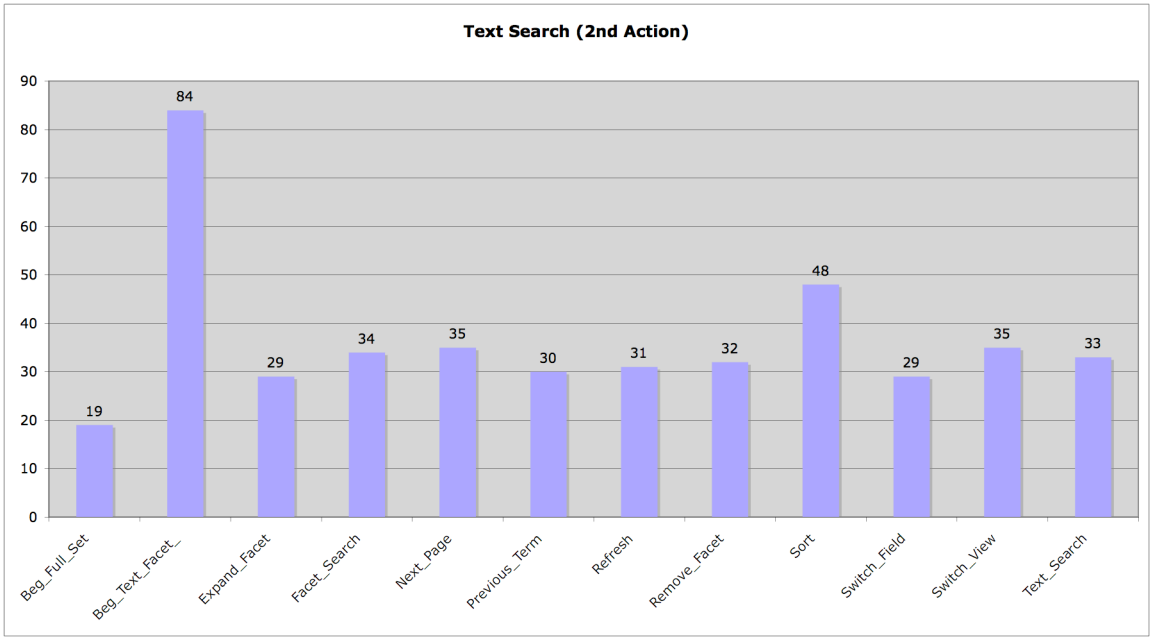
Yu, H. & Young, M. (2004) "The Impact of Web Search Engines on Subject Searching in OPAC." *Information Technology and Libraries.* 23(4), 168-180.

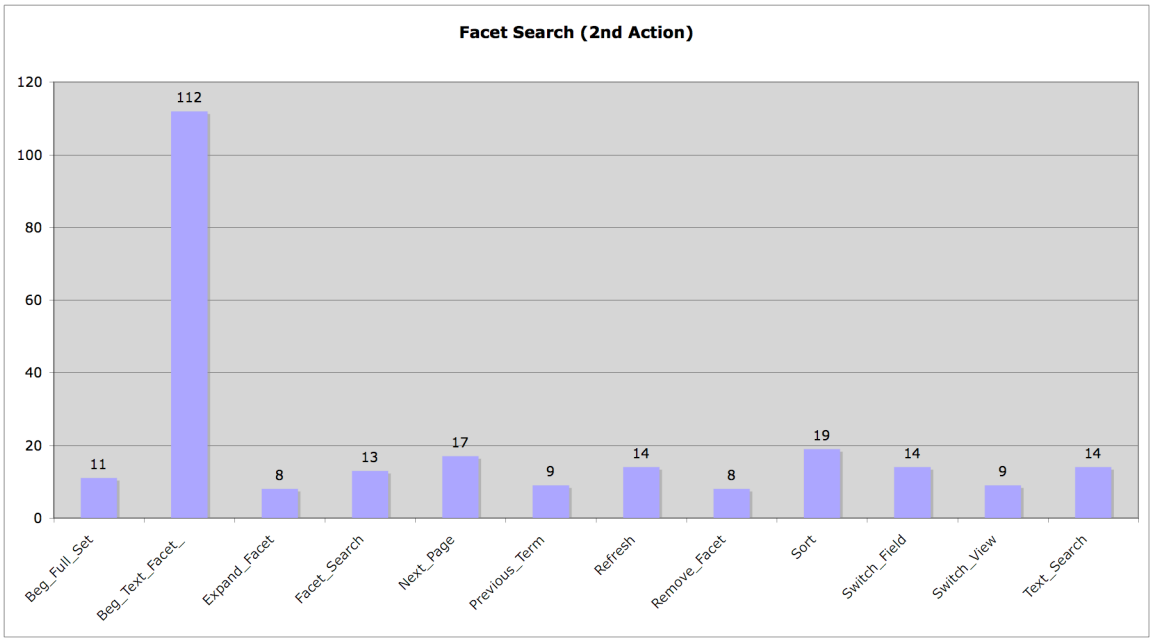
Appendix A – Additional Statistics

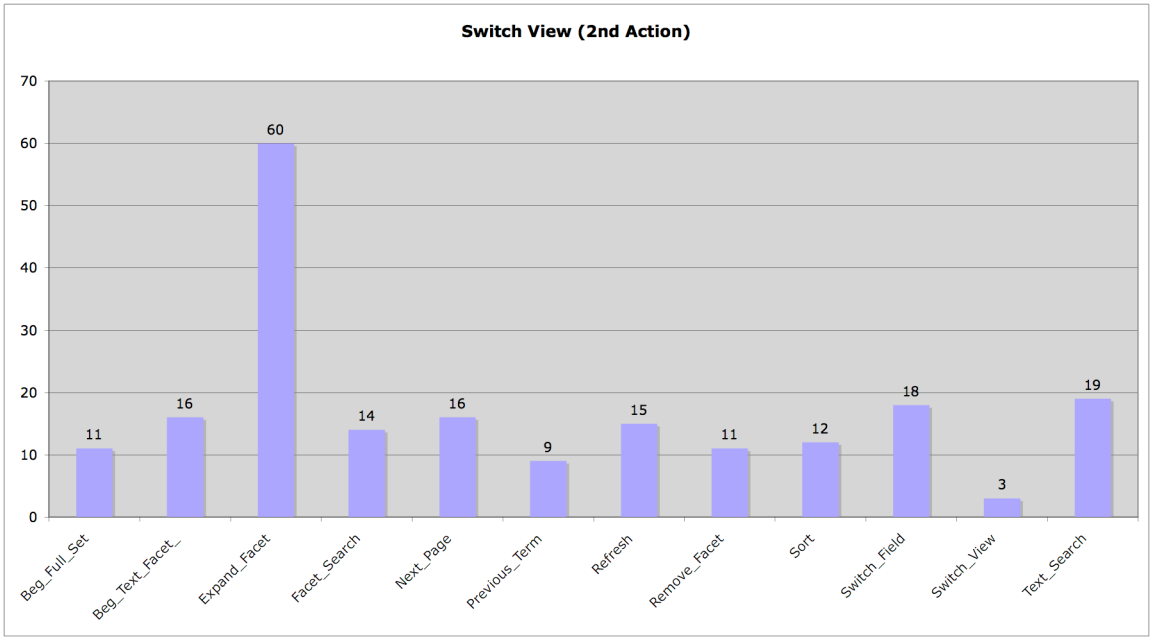
A strength and weakness of log analysis is the large amount of data it produces. The following statistics supplement those discussed in the body of this paper. I have chosen to report this information to provide a fuller picture of the information in the logs. However, it is unclear how the information is significant, or what it might mean. With the addition of user studies, the following may take on new significance.

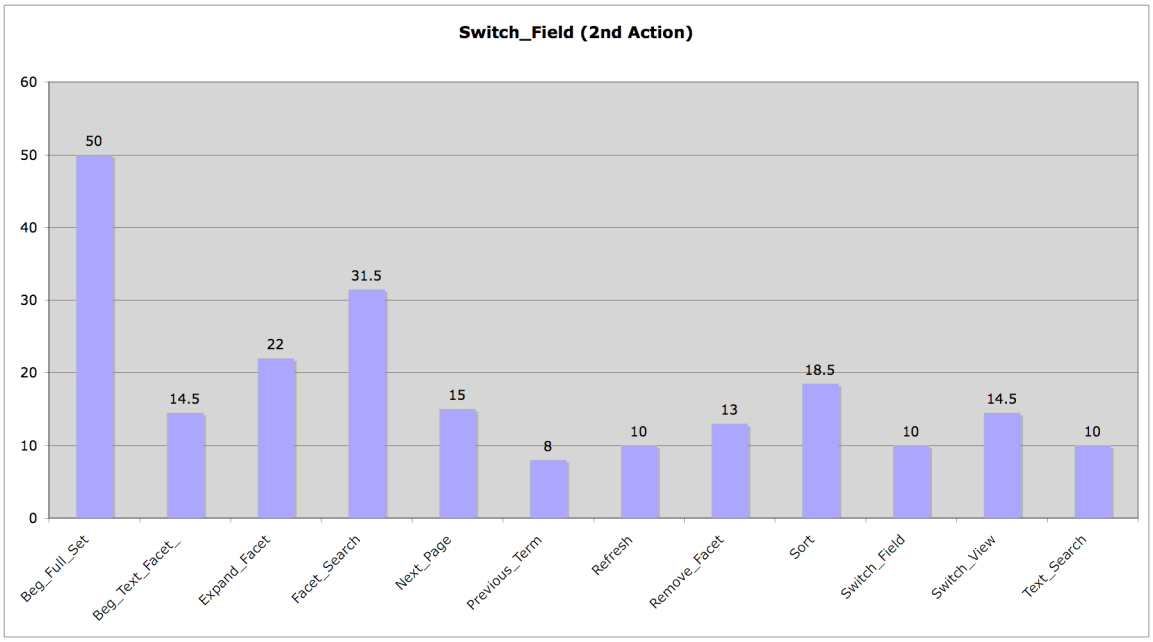
1.14 Dwell Times

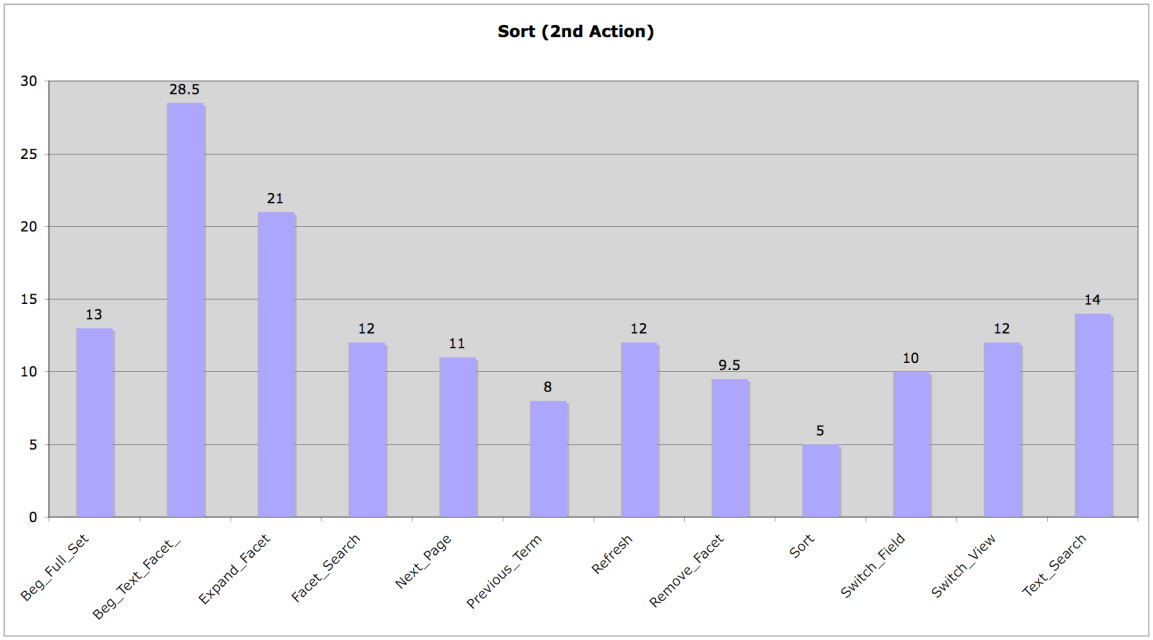
The following series of charts show the median dwell time between the actions on the x-axis and the action that serves as the title of the chart (2nd action). Of note, the first chart labeled "Text Search (2nd Action)" shows that in general, no matter what the previous action, text searching requires about 20-30 seconds to execute. The second chart, labeled "Facet Search (2nd Action)," shows that in general facet searching takes between 10 and 20 seconds. This supports the hypothesis discussed in the paper that text searching is more expensive than facet searching. The rest of the charts are included for comparison.

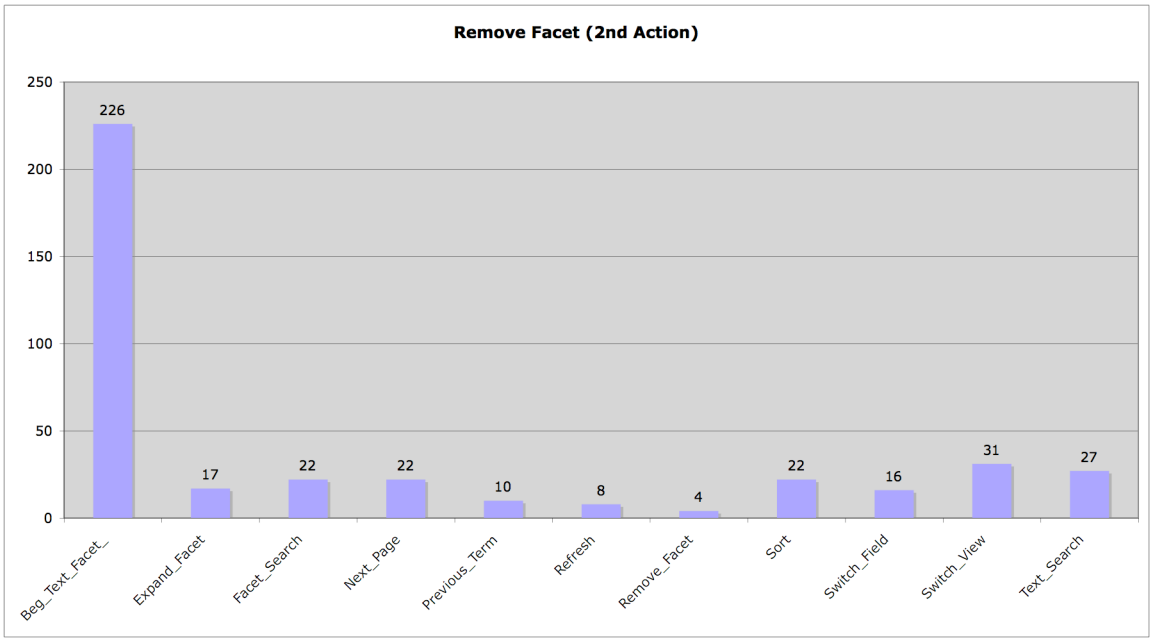

Text Search (2nd Action)
90
80
70
60
50
40
30
20
10
0
19
84
29
34
35
30
31
32
48
29
35
33
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search

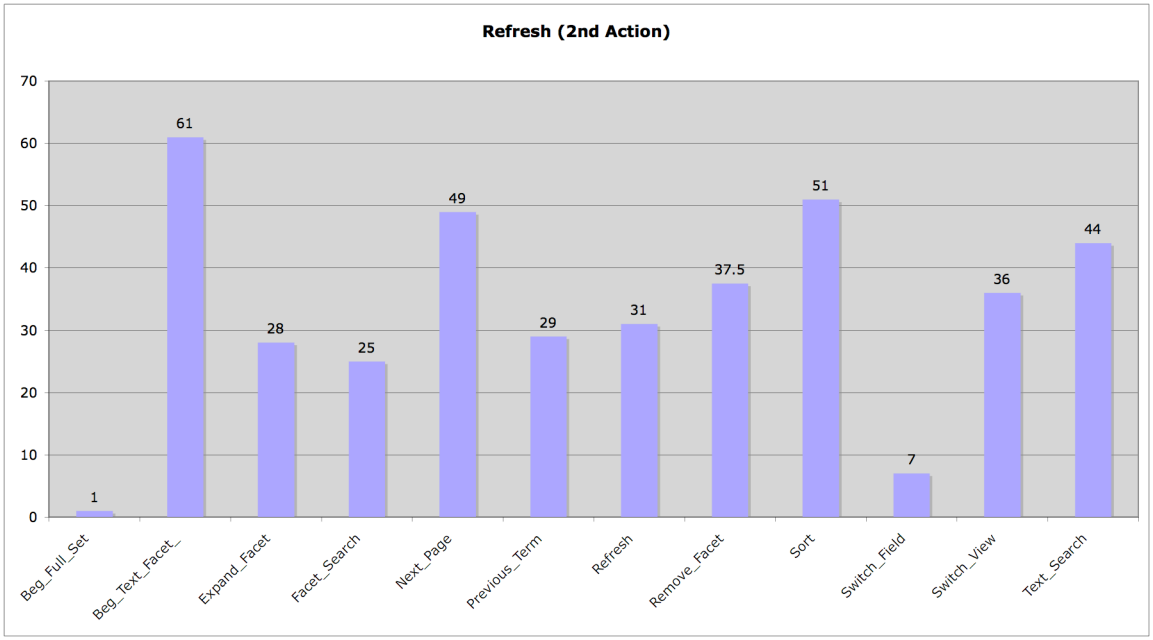

Facet Search (2nd Action)
120
100
80
60
40
20
0
11
112
8
13
17
9
14
8
19
14
9
14
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search

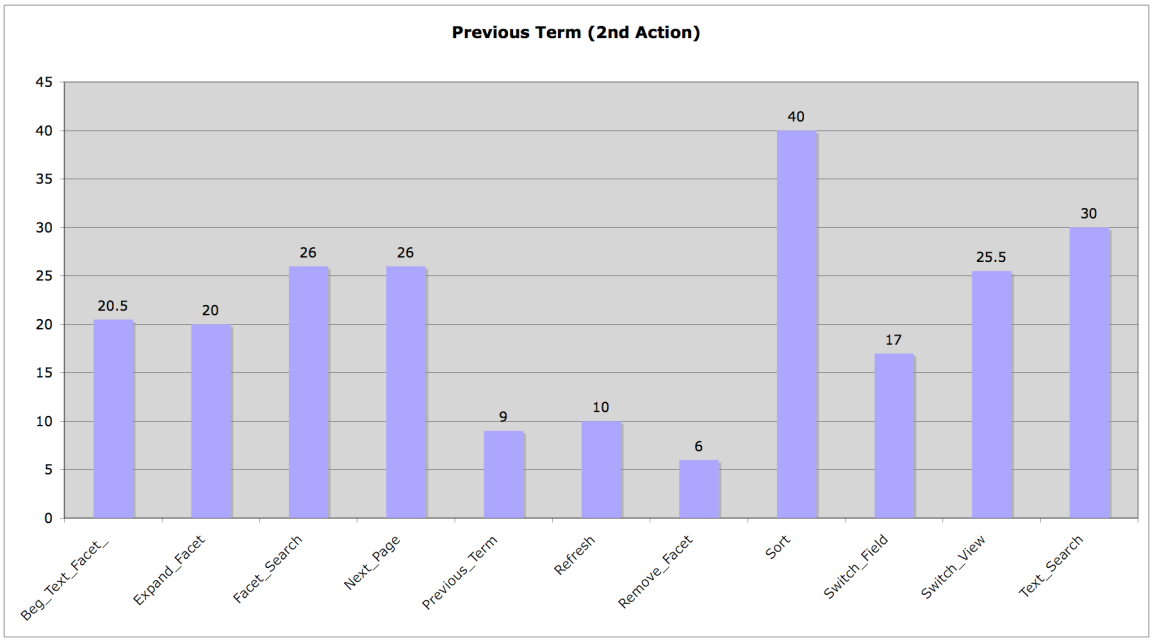

Switch View (2nd Action)
70
60
50
40
30
20
10
0
11
16
60
14
16
9
15
11
12
18
3
19
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search

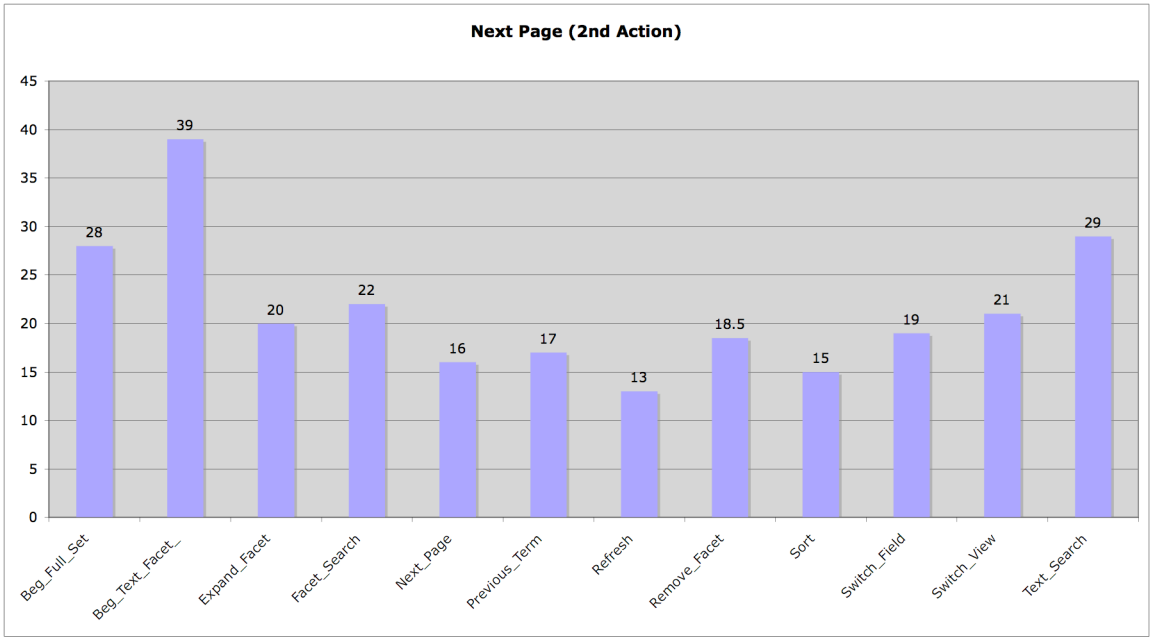

Switch_Field (2nd Action)
60
50
40
30
20
10
0
50
14.5
22
31.5
15
8
10
13
18.5
10
14.5
10
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search

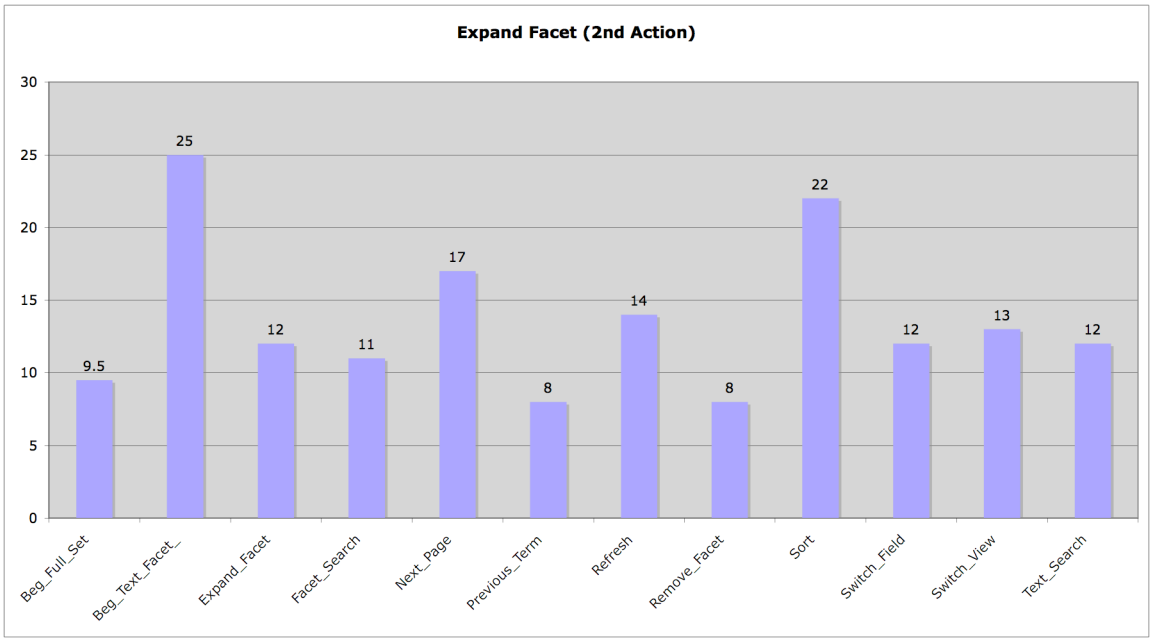

Sort (2nd Action)
30
25
20
15
10
5
0
13
28.5
21
12
11
8
12
9.5
5
10
12
14
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search


Remove Facet (2nd Action)
250
200
150
100
50
0
226
17
22
22
10
8
4
22
16
31
27
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search


Refresh (2nd Action)
70
60
50
40
30
20
10
0
1
61
28
25
49
29
31
37.5
51
7
36
44
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search

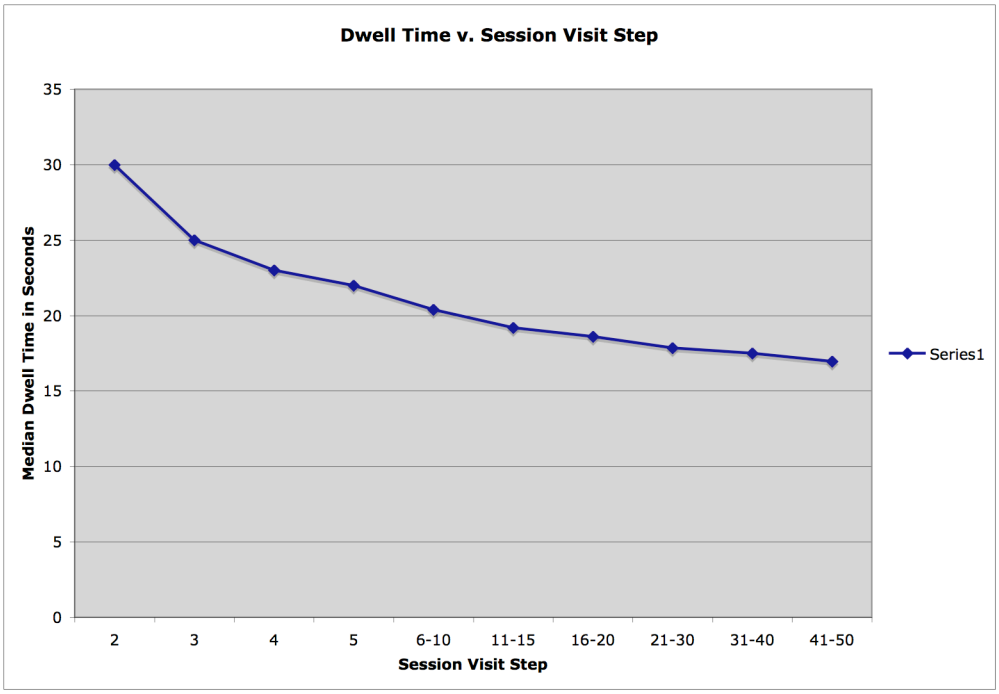

Previous Term (2nd Action)
45
40
35
30
25
20
15
10
5
0
20.5
20
26
26
9
10
6
40
17
25.5
30
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search


Next Page (2nd Action)
45
40
35
30
25
20
15
10
5
0
28
39
20
22
16
17
13
18.5
15
19
21
29
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search


Expand Facet (2nd Action)
30
25
20
15
10
5
0
9.5
25
12
11
17
8
14
8
22
12
13
12
Beg_Full_Set
Beg_Text_Facet_
Expand_Facet
Facet_Search
Next_Page
Previous_Term
Refresh
Remove_Facet
Sort
Switch_Field
Switch_View
Text_Search

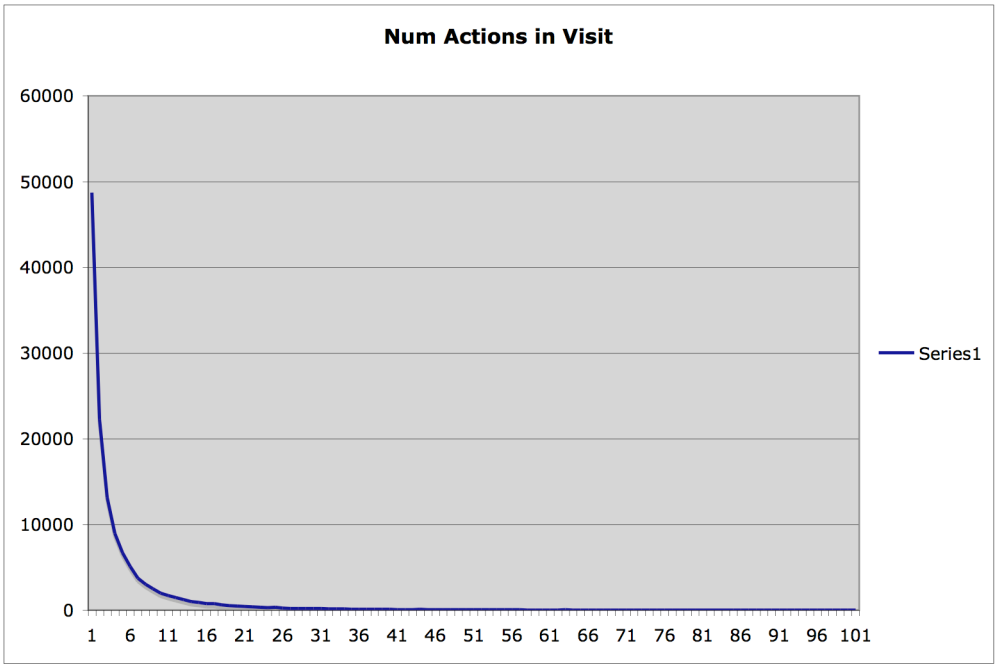
1.15 Dwell Time by Visit Step

This chart shows median dwell time by visit step within a session. In general, dwell time decreases over the course of a session, starting from about 30 seconds and leveling off to about 17 seconds by the 16th step.



1.16 Histogram of Actions per Visit

The following chart shows the number of sessions that contain a given number of actions. Most sessions are short, lasting not more than 6 to 10 steps. Although, some sessions last much longer, there are very few sessions that reach 101 actions (the artificial maximum allowed in this analysis).



1.17 Top 20 Most Frequently Used Facets By Group

The following table shows the 20 most frequently used facets within each facet group. There are only 13 options for the Library facet group, so all 13 are shown. For comparison, Format includes 50 possible facets, Language includes 44 possible facets, Library of Congress Call Numbers has 899 facets, Subject-Era has 268, Subject-Genre has 286, Subject-Region has 440, and Subject-Topic has 6323. Not all of these are visible on the interface at one time. Only facets that are relevant to the current set are displayed.

Dimension ID	Facet Group	Specific Facet	Count
4294952695	Author	Shakespeare, William, 1564-1616.	20
4294960224	Author	Bloom, Harold.	18
4294964396	Author	Society of Photo-optical Instrumentation Engineers.	15
4294936489	Author	Faculty Publication Collection (North Carolina State University)	15
4294567023	Author	Rowling, J. K.	13
4294781389	Author	Darwin, Charles, 1809-1882.	12
4294885107	Author	Institute of Textile Technology (Charlottesville, Va.)	12
4294880833	Author	Twain, Mark, 1835-1910.	11
4294934972	Author	Adler, Alfred, 1870-1937.	10
4294964045	Author	Mozart, Wolfgang Amadeus, 1756-1791.	10
4294956010	Author	Institute of Electrical and Electronics Engineers.	10
4294855097	Author	Criterion Collection (Firm)	10
4294952720	Author	Geological Survey (U.S.)	10
4294860034	Author	United States. Congress. Senate.	9
4294966990	Author	Kotler, Philip.	9
4293646480	Author	Recorded Books, Inc.	9
4294429545	Author	IEEE Xplore (Online service)	9
4294916878	Author	Lewis, C. S. (Clive Staples), 1898-1963.	9
4294948926	Author	King, Martin Luther, Jr., 1929-1968.	8
4294268597	Author	Kleinrock, Leonard.	8

206437	Format	Book	3602
206432	Format	Online	3310
206439	Format	Journal, Magazine, or Serial	1977
206431	Format	Videos and DVDs	1472
206429	Format	NCSU Thesis/Dissertation	534
206434	Format	Software and Multimedia	387
200044	Format	Book	320
206438	Format	eBook	247
206430	Format	Audio	238
200046	Format	eBook	237
206433	Format	Microform	194
200077	Format	Video DVD	110
200052	Format	Electronic journal	78
200049	Format	Journal or Magazine	77
200061	Format	Map	48
200074	Format	Electronic resource	44
200088	Format	Musical score	43
206435	Format	Electronic journal	40
200047	Format	Thesis	38
206426	Format	Audio book	37
200672	Language	English	608
200685	Language	French	151
200696	Language	German	119
200936	Language	Spanish	118
200745	Language	Italian	44
200630	Language	Chinese	34
200722	Language	Hindi	28
200897	Language	Russian	27
200786	Language	Latin	24
200986	Language	Urdu	18
200747	Language	Japanese	17
200882	Language	Polish	10
200665	Language	Dutch	9
200573	Language	Arabic	8
200569	Language	Algonquian (Other)	7
200709	Language	Greek, Ancient (to 1453)	6
200884	Language	Portuguese	5
200773	Language	Korean	5
200831	Language	Multiple languages	5
200983	Language	Ukrainian	3
200008	Library	D.H. Hill	1909
200011	Library	Design	1649
200012	Library	Online Resources	1595
200009	Library	Textiles	881
200013	Library	Special Collections	384
200007	Library	Natural Resources	362
200014	Library	Satellite Shelving	304

200010	Library	Veterinary Medicine	229
200016	Library	CED Media Center	193
200015	Library	AACCRR	93
206425	Library	Mathematics Work. Coll.	37
206421	Library	Off-site Shelving	30
206465	Library	Prague Institute	4
205324	Library of Congress Class	Q - Science	2627
202477	Library of Congress Class	H - Social sciences	1925
204514	Library of Congress Class	P - Language and literature	1713
205872	Library of Congress Class	T - Technology.	1536
201047	Library of Congress Class	B - Philosophy. Psychology. Religion	1438
201959	Library of Congress Class	E - History: America	1041
204342	Library of Congress Class	N - Fine Arts	1021
201493	Library of Congress Class	D - History (General) and History of Europe	1015
203862	Library of Congress Class	L - Education	893
205454	Library of Congress Class	R - Medicine	746
205673	Library of Congress Class	S - Agriculture	625
202270	Library of Congress Class	G - Geography. Anthropology. Recreation	593
201015	Library of Congress Class	A - General Works	531
202862	Library of Congress Class	J - Political Science	518
204152	Library of Congress Class	M - Music	510
202114	Library of Congress Class	F - America: local history	378
206305	Library of Congress Class	Z - Bibliography. Library Science. Information resources (general)	355
205329	Library of Congress Class	QA1 - QA939 Mathematics	320
206102	Library of Congress Class	U - Military science (General)	250
205355	Library of Congress Class	QC1 - QC999 Physics	248
4294967002	Subject: Era	20th century	444
4294967144	Subject: Era	19th century	188

4294966650	Subject: Era	Civil War, 1861-1865	91
4294967107	Subject: Era	18th century	85
4294965080	Subject: Era	Colonial period, ca. 1600-1775	70
4294966661	Subject: Era	17th century	53
4294961465	Subject: Era	Revolution, 1775-1783	47
4294872605	Subject: Era	21st century	29
4294966790	Subject: Era	-1945	27
4294966662	Subject: Era	16th century	20
4294966664	Subject: Era	Early modern, 1500-1700	19
4294966804	Subject: Era	Medieval, 500-1500	18
4294963769	Subject: Era	1933-1945	17
4294506954	Subject: Era	-2001	12
4294959019	Subject: Era	1918-1945	11
4294964260	Subject: Era	Middle Ages, 600-1500	10
4294964308	Subject: Era	-1980	10
4294953513	Subject: Era	1775-1865	10
4294966692	Subject: Era	To 1500	8
4294963994	Subject: Era	-1960	8
200023	Subject: Genre	Fiction	848
4294967075	Subject: Genre	Biography	507
4294966538	Subject: Genre	Handbooks, manuals, etc	450
206428	Subject: Genre	Primary Sources	417
4294967063	Subject: Genre	Congresses	210
4294962939	Subject: Genre	Feature films	185
4294966898	Subject: Genre	Case studies	166
4294965685	Subject: Genre	Statistics	153
4294966855	Subject: Genre	Dictionaries	102
4294967058	Subject: Genre	Poetry	85
4294966293	Subject: Genre	Pictorial works	76
4294965991	Subject: Genre	Bibliography	75
4294967089	Subject: Genre	Maps	73
4294966284	Subject: Genre	Drama	64
4294966872	Subject: Genre	Encyclopedias	63
4294965205	Subject: Genre	Early works to 1800	60
4294960266	Subject: Genre	Personal narratives	58
4294967269	Subject: Genre	Juvenile literature	54
4294965320	Subject: Genre	Guidebooks	48
4294952098	Subject: Genre	Documentary films	47
4294967215	Subject: Region	United States	564
4294966769	Subject: Region	North Carolina	196
4294966738	Subject: Region	Great Britain	160
4294966913	Subject: Region	India	89
4294967131	Subject: Region	England	68
4294967109	Subject: Region	Europe	59
4294964936	Subject: Region	Italy	57
4294964564	Subject: Region	France	53
4294966573	Subject: Region	Latin America	48

4294967126	Subject: Region	Rome	41
4294967262	Subject: Region	China	38
4294966676	Subject: Region	Germany	37
4294966750	Subject: Region	Southern States	36
4294963098	Subject: Region	Spain	33
4294966310	Subject: Region	Japan	29
4294965669	Subject: Region	Mexico	27
4294963878	Subject: Region	North America	26
4294966381	Subject: Region	Developing countries	25
4294963039	Subject: Region	Massachusetts	23
4294966183	Subject: Region	Africa	23
4294967172	Subject: Topic	History	1892
4294967157	Subject: Topic	History and criticism	397
4294966755	Subject: Topic	Criticism and interpretation	299
4294966864	Subject: Topic	Politics and government	250
4294967080	Subject: Topic	Social aspects	234
4294966802	Subject: Topic	Social conditions	193
4294966582	Subject: Topic	Mathematics	192
4294966296	Subject: Topic	Architecture	188
4294966174	Subject: Topic	Design and construction	168
4294967214	Subject: Topic	Management	166
4294966373	Subject: Topic	Environmental aspects	163
4294966732	Subject: Topic	Education	158
4294966020	Subject: Topic	Materials	147
4294965432	Subject: Topic	Mathematical models	140
4294967170	Subject: Topic	African Americans	126
4294966326	Subject: Topic	Women	126
4294967217	Subject: Topic	United States	121
4294967121	Subject: Topic	Social life and customs	119
4294967176	Subject: Topic	World War, 1939-1945	112
4294966520	Subject: Topic	Study and teaching	112

1.18 Endeca Generated Reports

This chart is a summary of some of the data available in the Endeca generated reports from the same time period as the logs studied in this report, January through April 2007.

	January	February	March	April	Total	Percent
Requests with One Dimension Value	25,829	22,485	35,610	28,433	112,357	36.91
Requests with Two Dimension Values	18,947	9,912	30,587	14,646	74,092	24.34
Requests with Three Dimension Values	14,795	2,919	53,057	6,031	76,802	25.23
Requests with Four+ Dimension Values	936	985	38,031	1,204	41,156	13.52
				TOTAL	304,407	
	January	**February**	**March**	**April**	**Total**	**Percent**
Search-Only Requests	112,741	115,931	113,842	114,142	456,656	48.59
Navigation-Only Requests	11,311	9,475	116,893	16,659	154,338	16.42
Search-Then-Navigate Requests	51,161	28,532	43,021	35,891	158,605	16.88
Root Requests	297	221	216	319	1,053	0.11
Record Requests	1	0	0	104	105	0.01
Other Requests	42,672	37,440	47,151	41,703	168,966	17.98
				TOTAL	939,723	
Top Search Keys						
Search Key	**January**	**February**	**March**	**April**	**Total**	**Percent**
Keyword	60,554	68,367	64,476	70,263	263,660	57.80
ISBN	19,645	16,967	18,804	16,018	71,434	15.66
Title	16,270	16,477	12,845	13,267	58,859	12.90
Author	10,478	8,663	10,262	8,953	38,356	8.41
Subject	3,955	4,017	6,131	4,379	18,482	4.05
Title|Author	1,225	847	793	733	3,598	0.79
Keyword|Author	287	269	185	198	939	0.21
Keyword|Title	105	117	107	140	469	0.10
Author|Subject	42	28	71	64	205	0.04
Keyword|Subject	38	52	52	28	170	0.04
				TOTAL	456,172	