Comparative Analysis of User Searching in Domain-Specific and Domain-Independent Digital Libraries

Oksana L. Zavalina  
College of Information  
University of North Texas  
Oksana.Zavalina@unt.edu

Elena V. Vassilieva  
College of Information  
University of North Texas  
Elena.Vassilieva@unt.edu

Abstract

This note reports preliminary results of a comparative transaction log analysis study of user searching in two digital libraries in the United States — a large-scale digital library in the domain of US history and a domain-independent federal-level library that aggregates digital collections regardless of their subject scope. This investigation reveals similarities in user search behavior among the two types of digital libraries with regard to rate of collection-level search use and application of search limits, most often occurring search categories, etc. At the same time, notable differences are observed in the rate of fielded search and phrase search use, average search query length and frequencies, and distribution of some search categories. This study provides empirical data to support digital library developers’ decision making regarding audience-based information organization in large-scale digital libraries.

Keywords: digital libraries, domain-specific digital libraries, domain-independent digital libraries, information search, search queries, search categories, transaction log analysis

Introduction

The United States funding agencies have actively supported digitization of collections of information objects — documents, letters, photographs, learning objects, etc. — of high cultural, historical, and educational value. For example, the Institute of Museum and Library Services (IMLS) has awarded grants to fund creation of over 600 digital collections. The National Science Foundation has funded over 300 digital collections (Skog, McCourt, & Gorman, 2009).

Large-scale digital libraries provide centralized access to hundreds or even thousands of such digital collections and millions of information objects. A number of large-scale digital libraries of varying subject focus and scope exist. One of the largest national-level digital libraries — IMLS Digital Collections and Content (IMLSDCC http://imlsdcc.grainger.uiuc.edu) with 650 digital collections — is aimed at a general audience and covers a wide range of subject areas and disciplines. IMLSDCC is representative of the most widely occurring type of a digital library — a domain-independent digital library. Some other national-level large-scale digital libraries are domain-specific, i.e., are created for distinct audiences. For example, Opening History (http://imlsdcc.grainger.uiuc.edu/history) with over 1500 digital collections is developed for researchers and educators in the domain of United States history.

To make sure that large-scale digital libraries successfully meet information needs of their intended user communities — either broad as in case with IMLSDCC or specific as in case with Opening History — the design of their respective discovery and access systems should be informed by user tasks such as finding, identifying, selecting, and obtaining information (IFLA, 2008).

People engage in information search to satisfy their information needs. Searching is one of the two major types of interactions between users and discovery and access systems should be informed by user tasks such as library catalogs, databases, search engines, or digital libraries (Wilson, 2000). Searching is expressed through search queries — sets of one or more symbols (e.g., words, phrases, etc.) used to instruct a discovery and access system to locate potentially relevant information.
Transaction log analysis — “the study of electronically recorded interactions” between discovery and access systems and “the persons who search for the information found in those systems” (Peters, 1993, p. 41) — is one of the methods actively used for unobtrusive observation of user interaction with information systems, in particular, user searching. For example, Jansen, Spink, and Pedersen (2004) compared search query length and Boolean usage rates in catalogs and search engines; Beitzel and colleagues (2007), Jansen et al. (2007), Koshman and colleagues (2006), and Spink et al. (2002) categorized web search queries into groups such as people, places, and things. However, studies of user searching have largely focused on web search engines or online library catalogs while ignoring large-scale digital libraries that have been created with the goal to support information needs of researchers and educators in various domains.

To improve user interaction with large-scale digital libraries and facilitate access to information, needs and information searching of their users should be taken into consideration in systems design and development. Different groups of users may interact with such libraries differently based on their varying information needs. This may require different policies regarding organization of information in domain-specific and domain-independent large-scale digital libraries. However, systematic investigation into the user searching in the large-scale digital libraries, let along comparative analysis, is in its infancy. Several studies (e.g., Agosti et al., 2007; Khoo et al., 2008; Pan, 2003) have analyzed transaction logs of large-scale digital libraries but they focused mainly on selection of search options and quantitative characteristics such as search query frequency, length etc. Only two studies (Zavalina, 2007; 2011) undertook detailed content analysis of user search queries in a large-scale digital library. However, user search queries in the different types of digital libraries aimed at different user communities have not previously been compared.

The study, preliminary results of which are reported in this note, addresses this gap through comparative analysis of user searching in domain-independent and domain-specific digital library.

Methods

This study analyzed transaction log data from a representative domain-independent large-scale digital library (IMLSDCC) and a representative domain-specific digital library (Opening History, hereafter referred to as OH). At the time of analysis, both digital libraries were hosted on the same server by the University of Illinois at Urbana-Champaign and, with exception of a color scheme, had identical user interfaces, which, among other options, included a rarely found option for collection-level search — a search where the user can search for entire digital collections as opposed to individual items. This similarity of interfaces makes the two digital libraries uniquely positioned as perfect candidates for comparative analysis of user searching as any differences in searching cannot be attributed to the differences in search interface.

This study aimed to answer the following research question: How does user searching in digital libraries that have wide subject scope and serve general audience (i.e., domain-independent) compare to user searching in digital libraries with distinct subject foci such as US history (i.e., domain-specific)? In particular,

• in search query length and search query frequency?
• in the distribution of search options: advanced search, collection-level search, etc.?
• in the distribution of search categories, such as personal and geographic names, dates, etc.?

One complete year of search log data, collected by Google Analytics application, for each of the two target digital libraries was used in this analysis. Non-empty search queries were grouped with identical queries. This resulted in a total of 5,917 unique search queries: 2,715 in OH and 3,202 in IMLSDCC. This sample size allows for generalizations with 95% confidence level and 5% margin of error.

Both quantitative and qualitative characteristics of search queries were assessed. Query length was measured as the number of words in a query, query frequency — as the number of times a query appears in the sample (cf., Spink et al., 2001). Unique search queries were analyzed qualitatively. They were categorized into ten search categories, including seven derived from the Functional Requirements

---

1 On August 1, 2012, the interfaces of two digital libraries were merged into one, under the IMLSDCC name.
for Bibliographic Records (IFLA, 1998; 2008) (work, person, corporate body, concept, object, event, and place), one from Functional Requirements for Authority Data (IFLA, 2009) (family), and two from the study by Zavalina (2007): class of persons, and ethnic group. Polysemic user search queries and most phrase queries were assigned to multiple categories. Whenever possible, search queries formulated in foreign languages were translated and categorized into appropriate search categories. This approach allowed to meaningfully categorize more than 95% of the queries in the sample. Approximately 4.71% of unique search queries in two digital libraries (4.27% in the Opening History and 5.09% in the IMLSDCC) could not be categorized into any of the search categories. These unknown category searches were excluded from further analysis.

A detailed coding manual had been developed to support coding activity; the manual included definitions and examples for each of the coding categories, along with other guidelines. In the coding process, the two authors or this note worked independently of each other and applied the same coding instructions to the same subsets of the units of analysis. The first author coded all 2,715 unique search queries in the OH sample; half of them were also coded by the second author. The second author also coded all 3,202 of IMLSDCC unique search queries; half of them were coded by the first author of this note.

To establish the reliability of the coding measures, there is a need to assess the amount of agreement among the coders. An intercoder reliability agreement coefficient of .90 or greater is considered acceptable to all and one of .80 or higher is acceptable to most situations. In this research, a strong intercoder reliability — 99.40% or Cohen’s Kappa of .976 for the OH dataset and 98.92% or Cohen’s Kappa of .954 for the IMLSDCC dataset — was observed between the two coders.

Findings and Discussion

Search Query Lengths and Frequencies

In the OH, search queries had an average of 2.32 and a median of 2 words per query (Table 1). The average length of IMLSDCC search queries was lower (1.94 words per query) while the median was the same. The median search query frequency was identical (1) in the two digital libraries. However, the mean was higher in OH (1.91) than in the IMLSDCC (1.60).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Query Length and Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
</tr>
<tr>
<td>OH search queries:</td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>2.32</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.91</td>
</tr>
<tr>
<td>IMLSDCC search queries:</td>
<td></td>
</tr>
<tr>
<td>Length</td>
<td>1.94</td>
</tr>
<tr>
<td>Frequency</td>
<td>1.60</td>
</tr>
</tbody>
</table>

Search Options

Most of the search queries in both digital libraries were basic keyword searches. The use of one or more advanced search options was observed in slightly under 15% of OH search queries and in almost 20% of IMLSDCC search queries overall.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Distribution of Advanced Search Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced search</td>
<td>OH (% of queries)</td>
</tr>
<tr>
<td>search limit</td>
<td>12.58%</td>
</tr>
<tr>
<td>fielded search</td>
<td>0.61%</td>
</tr>
<tr>
<td>phrase search</td>
<td>1.32%</td>
</tr>
</tbody>
</table>
As part of advanced search, both digital libraries provide options to limit collection-level search results by the type of objects in digital collections (for example, to retrieve only collections that contain photographs), and to limit item-level search results to a specific digital collection or a group of collections. Over 12% of search queries in OH and 13.37% in the IMLSDCC included one or more of these search limits (Table 2).

Fielded search option — advanced search where the user is allowed to search either by author or by title and subject words — is available for item-level search in both digital libraries. Fielded search was observed in 0.61% of OH and 5.00% in the IMLSDCC search queries. Use of another advanced search feature — phrase search with quotes — was observed in 1.32% of search queries in OH and 1.75% in the IMLSDCC.

In both digital libraries, collection-level search was used more often than item-level search option, although the latter is less prominently located on the page. In the OH, collection-level search accounted for 52.52% of all unique search queries; in IMLSDCC this proportion was even higher (57.21%).

Search Categories

In this subsection, results are presented as percentages of “known queries” — the queries that the authors of this paper were able to categorize into one or more search categories. There were a total of 2,599 such queries in OH sample, and 3,039 in IMLSDCC sample. As shown in Figure 1, the top two user search categories in the domain-specific OH were place (e.g., “Wyoming”) with 34% of queries and object (e.g., “quilt”) with 31% of queries. It is worth noting that these categories belong to FRBR Group 3 of entities, or subject entities. Another Group 3 search category — concept (e.g., “public housing”) — was the fourth most often occurring search category (17%). However, the fourth FRBR Group 3 subject entity — event (e.g., “1935 meat strike”) — was observed much less than the other three (10%). The FRBR Group 2 search categories — person (e.g., “James Henry Blake”) and corporate body (e.g. “United Fruit Company”) — were observed in 26% and 14% of queries respectively. The work search category (e.g., “Crafte of conjureynge and howe to rule the fferye spiritts of ye”) was observed in 9% of queries, while the class of persons (e.g., “pilot”) and ethnic group (e.g., “Navajo”) search categories occurred in 8% and 5% of queries respectively. Finally, family search category (e.g., “Wright brothers”) was observed in only 0.43% of unique search queries.

Similarly, the top two user search categories in the domain-independent IMLSDCC were place (32% of queries) and object (31%). Concept was the third most frequently occurring search category in IMLSDCC with 27% of search queries, considerably more than in OH. The event category occurred much less often than other three FRBR Group 3 search categories and somewhat less often than in OH (7%).
Person and corporate body categories were also found less often than in OH: in 19% and 9% of search queries. The work search category was observed in the same proportion of queries as in OH (9%). Class of persons and ethnic group categories occurred in IMLSDCC considerably less frequently than in OH: in 5% and 2% of queries respectively. Finally, family category was observed in only 0.3% of queries, which is somewhat lower percentage than in OH.

Discussion of Findings

Preliminary findings of this exploratory comparative study of user searching in domain-specific and domain-independent large-scale digital libraries reveal several notable differences and some similarities as well. Among the two digital libraries with identical interfaces, in domain-independent digital library advanced search was used somewhat more often. While the rate of use of search limits was comparable, the use of phrase search differed substantially, with higher numbers for domain-independent digital library, and the use of fielded advanced search was significantly higher in domain-independent digital library. The proportion of collection-level searching was high for both digital libraries but also somewhat higher in domain-independent digital library.

Overall, the level of advanced searching observed in both domain-specific and domain-independent digital libraries is high compared with the findings of studies of user searching on the web (e.g., Spink & Jansen, 2004) or in online databases (e.g., Nicholas et al., 2009). This may indicate higher proportion of domain expert users in large-scale digital libraries, as a number of user studies (e.g., Hembrooke et al., 2005; Wildemuth, 2003; Zhang, Anghelescu, & Yuan, 2005) report that selection of advanced search options increases with increase in the user domain knowledge. However, the actual level of advanced searching in IMLSDCC and OH is still significantly lower than the level of preference (81%) for advanced search that was observed in a survey of the users of a similar large-scale digital library (European Digital Library) conducted by Agosti and colleagues (2007).

Both average length and average frequency of search queries were substantially higher in domain-specific digital library. The authors of this note are aware of only one previously published study (Jones et al., 2000) that measured search query lengths in a comparable, large-scale digital library environment. Current study's findings for domain-specific digital library query lengths are consistent with Jones et al. results, but domain-independent digital library search queries are found to be shorter.

Results of this study mostly agree with Bates’ (1996) taxonomy of key search query types. Search queries in both domain-specific OH and domain-independent IMLSDCC most often include person, place, and concept search categories, which correspond to Bates’ query types: names of individuals, geographical names, and discipline terms. However, event category, which corresponds to Bates’ key search query type of chronological terms, is found significantly less often in both digital libraries. This study has also revealed additional key search query category — object.

The same 4 search categories — place, object, person, and concept — occur more often than others in both domain-specific and domain-independent digital libraries. However, 5 search categories (person, corporate body, event, class of persons, and ethnic group), are used substantially more often in domain-specific digital library while one search category (concept) is used significantly more often in domain-independent digital library.

Finally, this study revealed the overall prevalence of subject searching in large-scale digital libraries. Three out of four most frequently occurring categories in both types of libraries — place, object, and concept — match Group 3 (subject) entities in the model of Functional Requirements for Bibliographic Records (IFLA, 1998; 2008) and therefore are in fact subject search categories.

Conclusion

This study provides empirical data to support digital library developers’ decisions regarding information organization in large-scale digital libraries. For example, based on the findings of this study, it would be wise for developers of both digital libraries in the domain of US history and domain-independent libraries to prioritize recording places, objects, persons, and concepts in their metadata records. In metadata creation for digital libraries in the domain of US history, special attention should be given to documenting corporate and personal names, dates, classes of persons and ethnic groups. At the same
time, developers of domain-independent digital libraries need to ensure recording concepts, places, and ethnic groups in metadata, and providing advanced search options, especially for fielded search.

Overall, user experience can be improved if large-scale digital libraries — both domain-specific and domain-independent — supply an option to limit search results by geographical area, which is suggested by the high proportion of place searching observed in this study. Last but not least, the prevalence of subject searching among the users of both domain-independent and domain-specific digital libraries, which was observed in this study, suggests that provision of the subject-based advanced search option should be prioritized in the design of large-scale digital libraries, regardless of domain.

While accurately and unobtrusively capturing objective data on the actions of digital library users, transaction log analysis is unable to provide data on users' motivations and reasoning behind these actions. Triangulation of transaction log analysis results with results obtained through other methods, for example through think-aloud protocol observations, would overcome this limitation and help develop a fuller picture of user searching in domain-specific and domain-independent large-scale digital libraries.

References


