Using Transaction Logs to Better Understand User Search Session Patterns in an Image-based Digital Library*

이미지 기반 디지털 도서관에서 이용자 검색 패턴의 효과적 이해를 위한 트랜잭션 로그 데이터 분석

Hye-Jung Han**
Soohyung Joo***
Dietmar Wolfram****

ABSTRACT

Server transaction logs containing complete click-through data from a digital library of primarily image-based documents were analyzed to better understand user search session behavior. One month of data was analyzed using descriptive statistics and network analysis methods. The findings reveal iterative search behaviors centered on result views and evaluation and topical areas of focus for the search sessions. The study is novel in its combined analytical techniques and use of click-through data for image collections.

초 록

본 연구는 이미지 기반 디지털 도서관의 이용자 검색 패턴을 효과적으로 분석하기 위해 이용자 검색 로그 데이터를 분석하였다. 기술 통계와 네트워크 분석 방법을 사용하여 한 달간 수집한 트랜잭션 로그 데이터를 분석하였다. 연구 결과는 이용자들이 특정 주제 내에서 검색 결과 보기와 이미지 아이템 평가를 반복적으로 수행하고 있음을 밝혀내었다. 본 연구는 이미지 자료 검색의 로그 분석을 위해 복합적 데이터 분석 방법을 이용하였다는 점에 의의가 있다.

Keywords: Transaction log analysis, Image retrieval, User search behavior
트랜잭션 로그 분석, 이미지 검색, 이용자 검색 행태

---

* This study was presented at the joint conference of Korean Biblia Society for Library and Information Science and the University of Wisconsin-Milwaukee in 2013.
** School of Information Studies, University of Wisconsin-Milwaukee(hanh@uwm.edu) (first author)
*** School of Information Studies, University of Wisconsin-Milwaukee(sjoo@uwm.edu) (co-author)
**** School of Information Studies, University of Wisconsin-Milwaukee(dwolfram@uwm.edu) (corresponding author)
논문접수일자 : 2014년 2월 9일 논문심사일자 : 2014년 2월 24일 게재확정일자 : 2014년 3월 16일
한국비블리아학회지, 25(1): 19-37, 2014. 〔http://dx.doi.org/10.14699/kbiblia.2014.25.1.019〕

1. Introduction

The study of user interactions with information retrieval systems has been examined from different perspectives, including direct observation and transaction logs. The growing availability of server-side transaction logs has provided researchers with a rich set of data that can objectively record the actions of system users, whether for online public access catalogs, bibliographic database systems, web search engines, or digital libraries. Studies of user interactions with public search engines have been carried out since the mid-1990s. Jansen (2006) has reported that stored data in transaction logs of Web search engines, intranets and websites can offer valuable insights about the information searching process of online searchers. A number of researchers (Jansen, Spink, and Saracevic 2000; Spink, Wolfram, Jansen, and Saracevic 2001; Jansen and Spink 2006; Zhang, Wolfram, and Wang 2009) have conducted transaction log query analysis of various websites. These studies have revealed that interactions with public search services such as search engines are generally brief with little browsing of results and modification of findings (Spink, Wolfram, Jansen, and Saracevic 2001; Jansen, Spink, and Koshman 2007; Markey 2007; Jansen, Booth, and Spink 2009). These studies have also focused primarily on query-related data and not on the full click-through data of transaction logs, i.e., actions related to all aspects of the information seeking process, including search results and related actions taken by users. The present research is prompted by the desire to determine if the findings regarding user interactions for public search services are similar for digital libraries (DLs) where, unlike general search engines, users will rely on specialized document collections, such as image collections, for specific purposes. Similarly, does the nature of the collection (e.g., images) also influence how users search?

The research questions addressed by the current study include:

1) What types of search actions do users engage in most frequently when searching image-based collections?
2) How do users engage in search actions during search sessions as reflected by an action transition network?
3) What are the frequently observed queries and terms in searching the studied image-based collections?
4) Are there differences in query patterns between queries submitted internal to the digital collections and those originating from outside the collections, such as search engines?

This study is novel in that it investigates users' search actions, especially transitions in those actions, in an image-based digital library. In research questions 1 and 2, we intend to understand which types of user actions occur most or least frequently and identify frequent transition patterns between those actions during the search process in the context of an image-based digital library. By analyzing un-

obtrusive transition logs, we are able to quantitatively identify the most and least frequently applied user actions. Furthermore, we are able to identify the most common transition patterns in actions that enable us to better understand users' engagement in search processes. The understanding of users' search actions is important to model users' search behavior in an image-based digital library, and to generate insights into more efficient and effective system design. Research questions 3 and 4 are prompted by the desire to determine if there are specific areas of the collections that are of particular interest to searchers and whether users are more likely to engage in more natural language-based searches or if users rely on metadata content from the digital collections. Given that the content of the digital collections is also externally accessible through search engines, a comparison of queries submitted internally through the digital collections interface and those external to the system may reveal different patterns arising from access to the digital collections search features that would not be available through a search engine interface.

2. Literature Review

To date, user interaction studies involving DLs have been carried out using direct search observation (Xie and Cool 2009; Joo and Xie 2012) and resource usage logs (Wolfram and Xie 2002). Transaction log analysis (TLA) studies of DLs also have been undertaken, but have focused on systems that are primarily text-based or mixed media (Jones, Cunningham, McNab, and Boddie 2000; Bollen and Luce 2002; Jamali, Nicholas and Huntington 2005). Khoo et al. (2008) discussed the use of web metrics to analyze various user actions in digital libraries including session length and page views. They pointed out the difficulty in identifying firm inferences of users' intentions made from web metrics using transaction logs. In other words, web metrics record users' behavior rather than the thought processes and intentions. The authors also noted the difficulty in distinguishing separate sessions from the same IP address.

Similarly, earlier studies of image or multimedia searching using public search engines focused largely on query-related activities surrounding image collections (Goodrum and Spink 2001; Choi and Rasmussen 2003; Jörgensen and Jörgensen 2005; Tjondronegoro, Spink, and Jansen 2009; Huurnink, Hollink, van den Heuvel, and de Rijke 2010) or click-through data for results assessment (Smith, Brien, and Ashman 2012). Effective searching for images presents challenges not found in text-based retrieval. The search process and actions undertaken over the course of a session (i.e., a series of actions undertaken to fulfill an information need) are not well understood. Moreover, studies of image retrieval on different information retrieval systems investigated users' query patterns. Choi and Rasmussen (2003) found that users' queries for images in digital archives focused on searches for names of events, actions or conditions, individuals and place names. Jansen (2008) identified users' image query structure

on the Web and provided attributes to classify Web image queries, while Pu (2008) examined the characteristics of users' failed queries for images on the Web. The results revealed that users showed successful image retrieval on the Web with short queries while users' failed searches for images were associated with longer queries. Finally, Choi, and Hsieh-Yee (2010) conducted a query analysis for images using an OPAC. The authors found that the Boolean operator "AND" was the most commonly used feature in the queries and that the most frequently occurring initial queries were composed of two terms.

Although previous studies of DL interactions have relied on user observation and screen capture software, the amount of data collected from this approach is more limited than the types of data that can be collected using click-through data automatically collected by Web servers. The present study sheds new light on user interactions with image-based DL collections by examining user session behavior on a larger scale than is feasible with direct observation. Each user engages in search behaviors that are reflected in recorded actions. When combined with other users' data, a more complete picture emerges of overall user behavior that reveals patterns which may help to inform how digital libraries are designed, and particularly how to better facilitate browsing.

3. Method

The University of Wisconsin-Milwaukee libraries house a digital library consisting of approximately 30 digitized visual collections. The collections comprise more than 54,000 photographic images and maps. In particular, the selected digital library has specialized collections of regional archives that address Milwaukee history, Milwaukee neighborhoods, and University of Wisconsin-Milwaukee-related archives. In addition, the digital library houses collections of the American Geographical Society Library, so there are a large number of maps or geographical items in the collections. The selected digital library has been developed using CONTENTdm software (http://www.contentdm.org/), which is a content management tool for archiving, managing, and servicing digitized items. CONTENTdm basically provides a search function, including basic search, advanced search, and faceted search, and supports different viewing options. Also, the UWM digital collections provide well organized metadata based on the Dublin Core metadata standard. Item view pages present images or videos with metadata fields to help users evaluate items. The metadata fields provide different attributes of a digitized item, including title, notes, date of photograph, photographer, subject terms, regional information, item size, rights, and other information. The primary entry point to the DL consists of several HTML pages that permit users to browse the collections or to submit queries to search the collection metadata. A screenshot taken from the DL appears in Figure 1.

〈Figure 1〉 Screenshot of the UWM Libraries Digital Collections

3.1 Session Identification and Analysis

Most user actions were carried out as PHP requests. PHP is a scripting language for dynamic and interactive Web pages. The presented analysis is based on one month of data collected in August of 2012. The transaction log files consist of all requests made by users and system actions taken. Each transaction log entry consists of an encrypted Internet Protocol (IP) address, time, date, HTTP command with PHP actions, and originating Uniform Resource Locator (URL). In addition, transaction records include users' query input and selection of topic categories through the HTTP "Post" data transition method. Data were first cleaned to remove system-initiated action requests that were not user-related. Unique page requests from outside the DL were also removed. These were assumed to represent direct links to specific image files within the collections and were not representative of search sessions.

One challenge for TLA in environments that do not require a login is the identification of session boundaries, i.e., when a search session begins and ends. IP addresses are usually used to identify actions

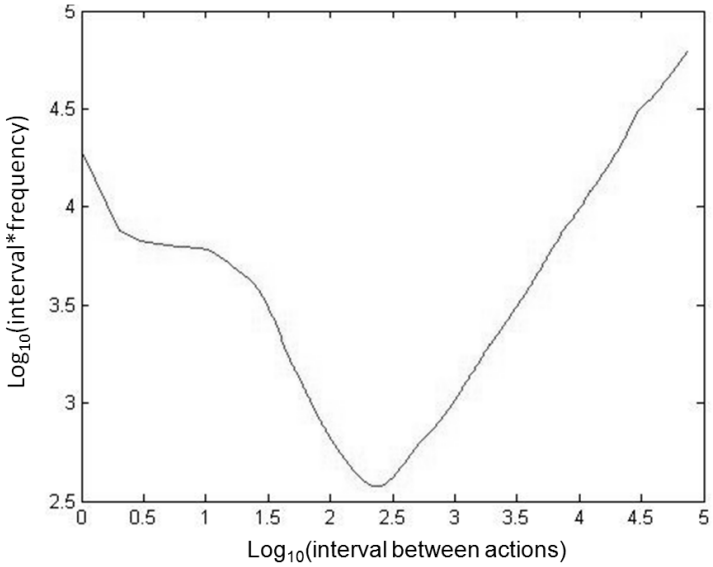
associated with a given session. This is also an issue in public DLs as noted by Khoo et al. (2008). For the present study, IP addresses were used as one indicator to delineate sessions. Actions taken on different days for the same IP address were also used to delimit sessions. The challenge with differentiating session boundaries arises when actions are associated with the same IP address on the same day, but in reality represent different sessions. It is possible that IP addresses may be shared by different computers, so that actions associated with one IP address may represent multiple, concurrent sessions. We have assumed that consecutive actions associated with the same IP address represent the same session. However, a period of inactivity between actions for the same IP address may also indicate a different session. How does one determine a cutoff point for periods of inaction that delineate different sessions for the same IP address? The cutoff determination for adjacent sessions with the same IP address should not be arbitrary. We employ a novel method for estimating the cutoff time between adjacent sessions for the same IP address. The cutoff time between actions served as the session boundary for actions associated with the same IP address on the same day. To identify the cutoff point for inactivity, we analyzed the distribution of time intervals in combination with the occurrence of those time intervals. To easily interpret the result, the data were transformed using logarithms. The distribution of intervals between actions exhibited a "V" shaped curvilinear pattern. We assumed that the inflection point implies the time interval that would be long enough to show the termination of a session. Before the inflection point, two adjacent actions were considered to occur in the same session. As the distribution is non-linear, Kernel regression was applied using MATLAB to find the inflection point. Kernel regression is a non-parametric estimation method to predict the conditional expectation of a random variable (Shawe-Taylor and Cristianini 2004). Unlike parametric regression, kernel regression includes a smoothing function as an estimation method, which consists of the kernel and bandwidth:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{x - X_i}{h})}{\sum_{i=1}^{n} K(\frac{x - X_i}{h})}$$

In this study, an Epanechnikov kernel was selected as a weight function (Li and Racine 2007), while the bandwidth was calculated based on Silverman's rule (Silverman 1986). The kernel regression result revealed the time interval cutoff between adjacent sessions was: $10^{2.4}$ = 251.19 seconds. This is reflected in Figure 2.

As a further limiter to sessions, because system content is accessible with direct links from outside sources, only sessions beginning with the home page or one of its subordinate PHP pages were included in the analysis.

The resulting cleaned log file was analyzed using descriptive statistics for the different types of available actions. A network analysis of adjacent actions undertaken within sessions was conducted using UCINET 6 (Borgatti, Everett, and Freeman 2002) to reveal the centrality and relationships among



〈Figure 2〉 Kernel regression of time interval distribution (log transformed)

different actions. Since users' search processes can be represented as transitions in search actions, we decided to focus on the changes of users' actions. A network analysis is a compelling tool to analyze and visualize those transitions. In this study, we viewed each type of action as a node connected by a directed edge to the adjacent action in the same session. The network analysis summarizes the entire transitions of user actions during search processes, and highlights the most frequent transition patterns of search actions in searching the digital library. Even though the network analysis does not show sequential search action patterns in individual sessions, it enables us to understand the overall users' search patterns comprehensively. By employing network analysis, we were able to identify which user actions play a central role in the search process and what types of transitions occur most frequently.

3.2 Query Analysis

An analysis of queries submitted during the data collection period was also conducted. There were two kinds of search query strings: those originating from the digital collection site (internal) and those from outside sources such as search engines (external). The identification of queries in the raw transaction log file was challenging. Internally submitted queries were identified by locating transaction log lines that contained the feature "CISOBOX1" as a field name for the primary search function, while the external searches were identified with "q=" for the query field. In order to compare the difference between internal searches and the external searches, all the transaction lines with the string "CISOBOX1=" in the referral field and all lines with the string of "q=" in the referral field were extracted into two files.

For both files, the entries were sorted in alphabetic order of the encrypted IP addresses. From the sorted data set, all lines with the repeated same referral fields from the same IP addresses were eliminated. The repeated lines were generated because often one displayed page has multiple components such as images, icons, and CSS files. In this way, all the unique queries were identified.

To analyze the records at the query and term level, only the query strings were extracted from the URL encodings at each referral field. The extracted query strings were further cleaned by removing all the non-alphabetic characters (digits and special characters) to make counting simpler. All the remaining alphabetic characters were transformed into lower case for more accurate word and query counting.

4. Results

4.1 Session Level Analysis

The transaction log analysis revealed 9758 sessions with an average of 10.86 page views or actions per session (median 2; min. 1; max. 446; s.d. 33.60) with a frequency distribution that was highly skewed (skewness 6.53). A summary of the total actions and page requests extracted from the transaction log file appears in Table 1.

The tallies of actions undertaken independent of the sessions provide an indication of which actions users are most likely to employ. Users are most likely to engage in requests for specific items (item_viewer.php) or query result (results.php) requests during sessions. The frequency of browsing actions was not

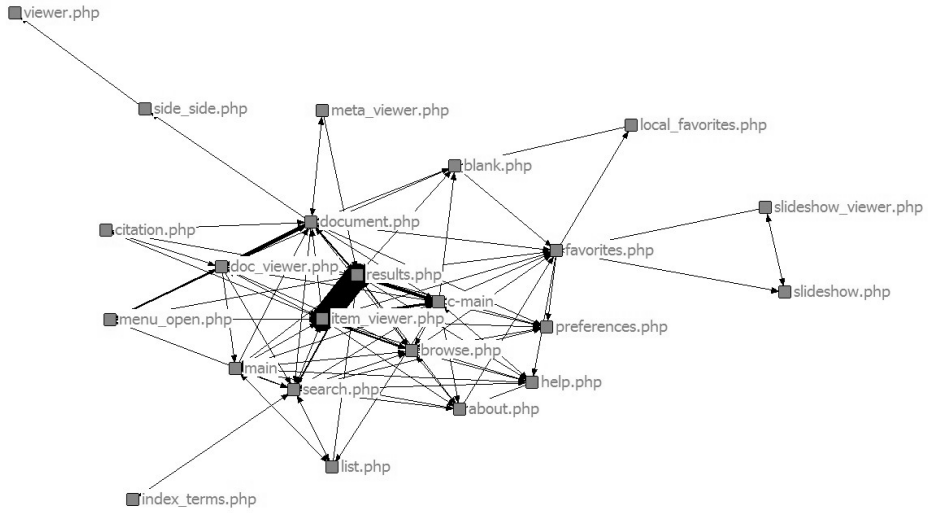
〈Table 1〉 Summary of the frequency of actions for all sessions

Page Type	Action	Total	Percentage
about.php	View the overview introduction of the collections	18	0.02%
blank.php	Encountered an empty page	179	0.16%
browse.php	Browse items by selected category	3546	3.10%
collection main	Enter the main page for a specific collection	3362	2.93%
doc_viewer.php	View a textual document with metadata	3524	3.08%
document.php	View an individual textual document page	3380	2.95%
favorites.php	Add the current item to user's personal list	27	0.02%
help.php	View help pages	12	0.01%
item_viewer.php	View a non-textual item such as photos/maps	53519	46.71%
list.php	Browse a list of items in a specific collection	135	0.12%
main	Enter the main page of the Digital Collections	401	0.35%
menu_open.php	View menu options in a specific collection	474	0.41%
preferences.php	Narrow down the current result	17	0.01%
results.php	View search result pages or browsing results	44796	39.10%
search.php	Enter an advanced search page	1135	0.99%
Other	Other infrequent actions	45	0.04%
Grand Total		114570	100.00%

detectable because in many cases selecting topic categories led to "results.php" instead of "browse.php" actions. When a user selected a certain topic in the predefined browsing categories, it was treated as query input and search results were provided to users. A certain amount of "results.php" entries were the consecutive actions after users' browsing action. Therefore, it was not possible to determine to which extent browsing actions were applied. Advanced search options and help pages, however, were less frequently used at 0.99% and 0.01%, respectively. Searches may have also been initiated outside of the system, with a search result link to the collection serving as the entry point. This is also reflected in the small number of requests for the main page (main, cmain), indicating that users' entry points to the digital collections are frequently a subpage and not the main page. This is also supported by the first page request for each session, where users are more likely to enter the system through a result (39.10%) or item viewer page (46.71%) than one of the two main pages (3.28%).

The transaction log also records transitions from one type of action to another. This allows the development of transition networks to identify the relationships among actions during sessions. A directed matrix of action transitions was created. We observed the two most frequently occurring transitions of search actions were "results.php to item_viewer.php" (16715) and "item_viewer.php to results.php" (13191). This reveals that users iteratively evaluate search results and individual items to achieve their search goals.

The resulting network created in UCINET appears in Figure 3. The line thickness indicates the strength of the relationship. It's not surprising that the strongest connections would be between those actions that occur most frequently. The observed transitions in user actions imply that a hub and spoke model (Catledge and Pitkow 1995) can also be applied to this DL environment. Catledge and Pitkow observed that users frequently return to a primary node such as an entry page and then proceed to other links during Web browsing. In this study, the most frequent search strategy was iterative search result evaluation, with frequent transitions between results.php and item_viewer.php pages; however, users did not frequently return to an entry page. What this reveals is that consideration of iterative loops between search results and items is essential. The efficiency of the hub and spoke structure should be enhanced in the design of image DLs. We also observed that advanced search options and help pages were infrequently used. This suggests the user tendency to engage in least effort activities during search processes. Implicit help features could better benefit users as they are not likely to use explicit help features such as help instructions. Similarly, providing more information to users about individual images (e.g., metadata) as part of the retrieval list and not just a thumbnail may reduce the need for users to alternate between the results list and individual item pages as frequently.



〈Figure 3〉 Session action transition network

4.2 Query Level Analysis

Query data were analyzed separately from external and internal data to compare the different characteristics between two groups. Queries were attributed to internal and external users based on the query origin, explained above. Internal queries were those originating from within the UWM digital collections. External queries originated from outside the UWM domain, for example, public search engines.

There were 2825 external query lines and 11194 internal query lines. Table 2 shows the 30 most frequent query strings for external searches and those for internal searches. The most frequently occurring query in the external data was the two-word query "family notices" constituting 24 occurrences. The most frequent query for the internal submissions was the three-word query "tsybikoff g ts" (a Russian explorer name: Gombojab Tsybikov (Romanized as Tsybikoff)) representing 1481 occurrences (Table 1). The top query results in this study were history-related. Internal queries reveal users are more interested in local information such as "wisconsin", and "milwaukee". Also, many terms are related to geographical information, which reflect the content of the digital collections. The UWM Digital Collections contain many collections that address local history and geographical information. In particular, the UWM Digital Collections have special collections from the American Geographical Society Library, which is part of the UWM Libraries. This shows that users' queries are associated with the topics and subjects of the collections. In particular, internal users would come to the UWM Digital Collections site with a specific intention to look for the collections of the American Geographical Society Library. This

result included users' spacing errors such as "subjectarticles" (Table 2: External query strings) during the query formulation. Also, spelling differences between British and American English such as "pearl harbour" were observed, as Jones et al. (2000) also found. The mean number of query terms per external query string is 2.45 with a standard deviation of 1.569. The mean of query terms per internal query strings is 1.96 with a standard deviation is 1.341. The lower number of terms per query may be due in part to internal users selecting categories for browsing, which are treated as queries by the system.

〈Table 2〉 The most frequent external and internal query strings

External Query Strings	Freq.	Internal Query Strings	Freq.
family__notices	24	tsybikoff__g__ts	1481
scenes__in__the__city__posting__the__messages	23	central__tibet	543
subjectarticles	15	wisconsin	182
kindergarten__union	13	near__south__side	145
knit__cast__on	11	china	138
frank__bradley	11	milwaukee	120
the__dawn	9	james__groppi__papers	98
knitting__patterns	9	s__s	90
mark__brinkley	9	hong__kong__harrison__forman	89
baumgarten	9	manila	81
china__military__police	7	california	79
gwalior	7	hong__kong	72
vegetation__map__of__asia	7	turkey	70
shanghai__evening__post__and__mercury	7	east__side	69
vietnam	7	downtown	69
pearl__harbour	7	new__york	66
hiroshima	6	united__states	64
refugees	6	people	59
empire__theater	6	am	59
deegan	6	forman__harrison	57
empire__theatre	6	nanniwan	56
schomberg	6	southeast__side	55
ellen__white	6	dwellings	53
thailand	5	near__north__side	53
asylum__hill__hotel	5	documents	51
ellen__brown	5	henan	48
functions__of__internet	5	west__side	47
morey	5	afghanistan	46
use__and__misuse__of__internet	5	hong	45

4.3 Term Level Analysis

Table 3 summarizes the 30 most frequent query terms from the external queries and the internal queries, respectively. Since one of the purposes of this study is to explore term frequency patterns, all observed words were included in the analysis, including grammatical words (e.g., articles, conjunctions, prepositions) such as *the*, *of*, *and* and *a*. Although these types of terms could be viewed as stop words by an IR system for indexing and retrieval purposes, they more accurately reflect user input and search intentions. Yi et al. (2006) as well as Spink, Wolfram, Jansen and Saracevic (2001), for instance, also include these terms in their analyses to more accurately report user query terms. For the purpose of identifying potential differences in groups of users, including these words can also reflect differences in how groups of users approach the search process, for example, through the use of natural language input by users versus subject-based searches or the use of controlled vocabularies or metadata. This is evident in the differences between internal and external query submissions to the system. For the external data, besides grammatical words, the most frequently used query terms are "Name related" for conducting people search (e.g. John, William, James, Thomas, Henry, Peter, etc.). Furthermore, the most frequently occurring terms in the internal data are "Geographic places related" such as Tibet, Milwaukee, Wisconsin and China. The data also revealed that the top individual terms are much more consistent with the top queries among internal queries than external queries. Again, because internal queries may represent, at least in part, browsing categories selected by the user and not user-entered query terms, fewer grammatical words may be present in the internal queries.

〈Table 3〉 The most frequent external and internal query terms

External Query Terms	Frequency	Internal Query Terms	Frequency
the	156	Tsybikoff	1505
of	137	G	1497
and	106	Ts	1481
internet	105	Tibet	619
in	87	Side	603
john	82	Milwaukee	594
william	65	Central	555
james	52	Wisconsin	413
on	42	Hong	298
thomas	42	Kong	256
henry	39	Forman	236
mary	37	South	231
city	34	China	228
research	34	Harrison	228
ellen	33	Near	227

External Query Terms	Frequency	Internal Query Terms	Frequency
online	32	S	187
notices	30	Groppi	135
scenes	28	James	123
family	28	And	122
george	26	Collection	117
posting	26	North	117
peter	25	Papers	114
messages	25	Of	110
war	24	East	107
church	23	Far	105
a	23	the	98
use	23	new	89
theatre	22	manila	84
asia	22	photographs	84
china	21	west	82

4.4 Term Co-occurrence Analysis

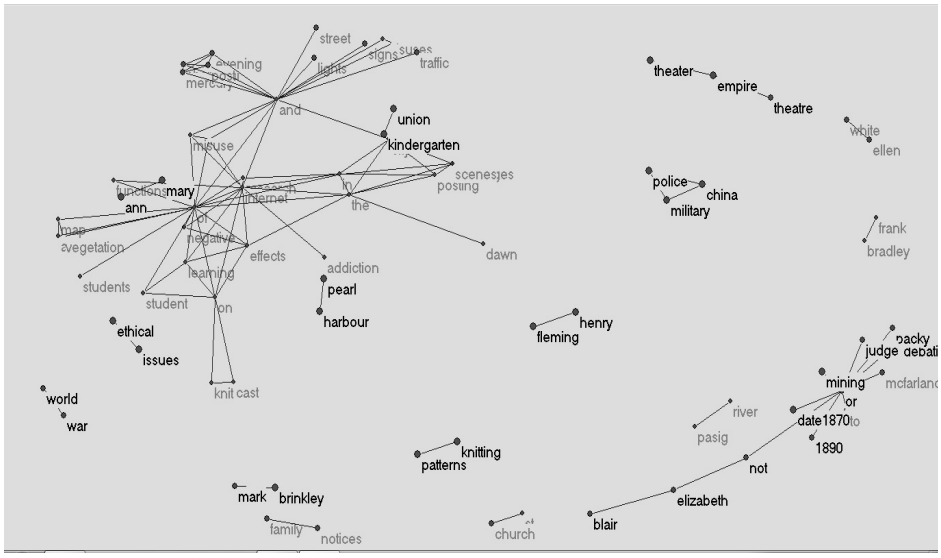
To analyze the relationship among query terms, all the possible pairs of terms from each query string were created by a PHP script. Table 4 provides the list of the most frequent term pairs for the external and the internal data. The most frequently occurring term pairs in the external data are “use and of”, “or and or”, and “map and of”. In the internal data, the most frequent word pairs are “g and ts”, “tsybikoff and g” and “tsybikoff and ts”. Thus, the internal top query pairs are consistent with the internal top query strings.

In order to visually examine the relationship among query terms, visualization of the relationship among all the terms in each data set (external vs. internal) was attempted in Pajek, one of the most popular network analysis tools (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). The data were converted to Pajek format using “txt2pajek.exe” with a weighted option for the pair frequency. Only the top 100 term pairs from each data set were included to avoid complex displays and to get meaningful outputs. Figure 4 displays the relationship among external search terms and Figure 5 illustrates those among internal search terms.

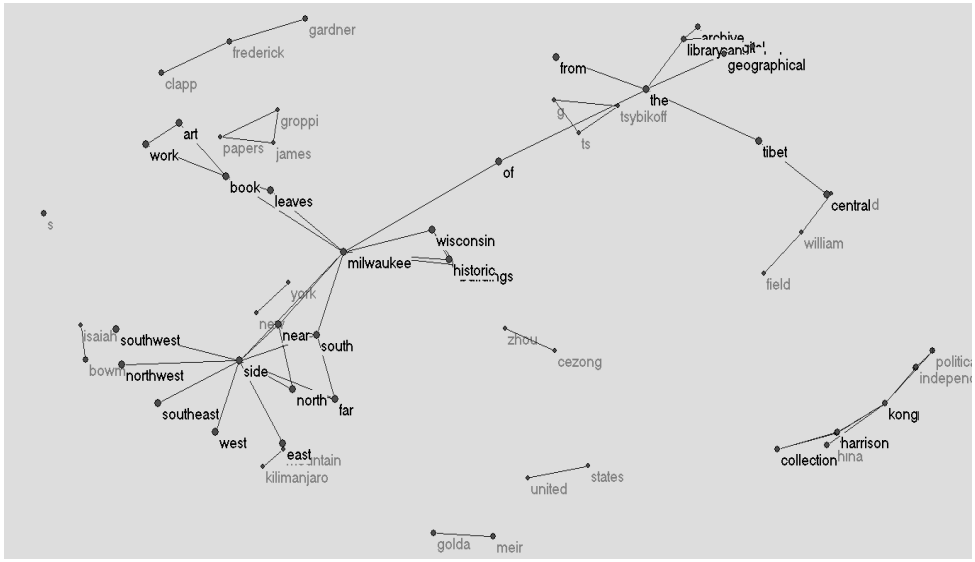
〈Table 4〉 The most frequent external and internal term pairs

External Word-Pair	Frequency	Internal Word-Pair	Frequency
use__of__	12	g__ts__	1481
or__or__	12	tsybikoff__g__	1481
map__of__	12	tsybikoff__ts__	1481
world__war__	11	central__tibet__	550
use__misuse__	11	hong__kong__	253
negative__of__	11	near__side__	227

External Word-Pair	Frequency	Internal Word-Pair	Frequency
frank__bradley__	11	wisconsin__milwaukee__	219
map__asia__	11	south__side__	211
knit__on__	11	near__south__	174
knit__cast__	11	harrison__forman__	159
of__in__	10	far__side__	105
military__police__	10	james__groppi__	101
knitting__patterns__	10	james__papers__	98
negative__effects__	10	groppi__papers__	98
of__on__	10	s__s__	90
city__and__	10	hong__forman__	89
and__misuse__	10	kong__forman__	89
internet__use__	9	hong__harrison__	89
vegetation__asia__	9	kong__harrison__	89
the__dawn__	9	east__side__	82
mark__brinkley__	9	north__side__	82
in__internet__	9	united__states__	69
and__and__	9	new__york__	68
vegetation__map__	9	milwaukee__book__	68
functions__internet__	8	west__side__	66
internet__addiction__	8	southeast__side__	60
functions__of__	8	forman__harrison__	58
and__signs__	8	northwest__side__	57
in__research__	8	american__society__	55
vegetation__of__	8	geographical__library__	55



〈Figure 4〉 The relationship among external search term pairs



〈Figure 5〉 The relationship among internal search term pairs

5. Discussion

The query analysis, although only covering a relatively short period of time, provides some insights into the digital collection usage. Internal queries included a certain amount of predefined topic categories, so those queries tended to be more directly related to the collection topics, indicating greater familiarity with the content of the collection, or reliance on system browsing features using predefined topic categories. The query analysis also shows the differences between external and internal query patterns clearly. Internal query strings, query terms and word-pairs show the consistent relationships with each other while the external query data do not show consistency with each other. User image query analysis of this academic digital library indicates that the average length of queries is short, with a mean of around 2 terms per a query. Additionally, users' image queries in this study focused on personal name and geographical place searches. Overall, the queries reflect the content of the digital collections. Because the UWM Digital Collections have many items related to history and geography, it's not surprising that frequent queries also reveal users' needs in those areas. External queries were more diverse and included more grammatical words compared to internal queries. The term pair analysis of co-occurring terms confirms that external queries represent a combination of grammatical words and keywords, while internal queries contain more subject terms chosen from predefined topic terms.

The findings of this study have implications for understanding search behavior and DL design for image collections. The findings reveal that users request search results and individual item view pages

most frequently while relatively fewer numbers of actions were related to query creation and reformulation. Users' most frequent actions were the evaluation of search result pages and individual item evaluation. Compared to these two actions, query related actions occurred less frequently. This finding reaffirms Xie and Joo's (2010) observation that users spend most of their time during the search session in evaluating search results and individual items. Previous transaction log research on image searching has focused more on query creation, reformulation and resulting page views (Goodrum and Spink 2001; Tjondronegoro, Spink, and Jansen 2009). Our study findings reveal that users engage more in evaluation actions rather than query term manipulation. This implies that user studies in digital libraries need to focus on users' evaluation behavior as well as query behavior. In addition, this study found that explicit help was infrequently used. According to a previous study (Xie and Cool 2009), users are less likely to use help pages when encountering problems in search processes. Thus, implicit help would be more useful by providing implicit system features (e.g., query suggestion, query expansion, etc.) for helping users resolving the problems or incorporating help instructions in the search process, instead of separate help pages.

Limitations of transaction log analysis must be acknowledged. One common limitation when relying on a public search system is that users are not required to log in, which then makes it necessary for session boundaries to be estimated in cases where there are no clear session delimiters. Second, the current study relied on a relatively small data set collected over a limited period of time. Session behaviors may vary at different times of the year. Third, although transaction logs provide a wealth of objective data that precisely record the actions taken by users, they cannot reveal why users engage in the actions they do or their level of satisfaction with their search outcomes. Log data provide only the resulting records of user actions. Despite the benefits of objectivity and unobtrusiveness of transaction log data, it does not provide any underlying intentions or contextual information behind users' actions. This, however, was not the purpose of the present study. Future research could combine direct observation and interaction with users of the digital collections to better understand their intentions.

6. Conclusion

In this study we have attempted to uncover unique search patterns in image-based collections by analyzing click-through transaction log data in an unobtrusive manner, including queries submitted by internal and external users of an image-based digital library, an IR environment that has not been widely studied to date. The use of transaction log data has provided a larger data set to study than direct observation would permit. Unlike most earlier studies of transaction logs, whether for search engines or specialized digital library environments that have focused on queries and their modification, we also examined complete click-through data that included result viewing and other system feature usage. One

key finding of the study is that transitions between results pages and item views were by far the most common actions taken by searchers. We found that the current organization of the DL site encourages a hub and spoke model for browsing search results. System interaction could be made more efficient by providing users with more metadata information for each image without having to click on each image first. The analysis of queries embedded within the transaction log as part of the session behavior revealed patterns in topics searched as well as differences in the content of internally and externally submitted queries, where external queries were more likely to integrate non-content-bearing terms, perhaps indicating that internally generated queries were relying on subject browsing or the use of controlled vocabulary terms.

We continue to explore unique search behavior in the context of digital libraries by analyzing different aspects of the log data. A larger log data set is being analyzed for more detailed session characteristics. As a further analysis, sessions will be classified into different themes based on their page view patterns. In addition, the authors plan to extract users' queries to investigate how the query characteristics affect users' subsequent search activities.

Acknowledgment

The authors would like to thank the staff of the University of Wisconsin-Milwaukee Libraries for providing access to the transaction log data and, in particular, Ann Hanlon for her assistance and helpful insights into the UWM Libraries Digital Collections. Portions of this study were presented at the 35th Annual Conference of the Canadian Association for Information Science meeting (Han, Joo, and Wolfram 2013).

References

Bollen, J. and R. Luce. 2002. Evaluation of Digital Library Impact and User Communities by Analysis of Usage Patterns. *D-Lib Magazine*, 8. [cited 2014.3.14]. <http://www.dlib.org/dlib/june02/bollen/06bollen.html>.

Borgatti, S.P., M.G. Everett, and L.C. Freeman. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies.

Catledge, L.D. and J.E. Pitkow. 1995. "Characterizing Browsing Strategies in the World-Wide Web." *Computer Networks and ISDN Systems*, 27(6): 1065-1073.

Choi, Y. and E. M. Rasmussen. 2003. "Searching for Images: The Analysis of Users' Queries for Image Retrieval in American History." *Journal of the American Society for Information Science and Technology*,

54(6): 498-511.

Choi, Y. and I. Hsieh-Yee. 2010. “Finding Images in an Online Public Access Catalogue: Analysis of User Queries, Subject Headings, and Description Notes.” *Canadian Journal of Information and Library Science*, 34(3): 271-295.

Goodrum, A. and A. Spink. 2001. “Image Searching on the Excite Web Search Engine.” *Information Processing and Management*, 37: 295-312.

Han, H., S. Joo, and D. Wolfram. 2013. Tales from Transaction Logs: User Search Session Patterns in an Image-based Digital Library. In K. Hodges, E. Meyers, & H. O’Brien (Eds.), *Tales from the Edge: Narrative Voices in Information Research and Practice*. [cited 2014.3.14]. <http://cais-acsi.ca/proceedings/2013/HanJooWolfram_submission_23.pdf>.

Huurnink, B., L. Hollink, W. van den Heuvel, and M. de Rijke. 2010. “Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis.” *Journal of the American Society for Information Science and Technology*, 61(6): 1180-1197.

Jamali, H.R., D. Nicholas, and P. Huntington. 2005. “The Use and Users of Scholarly E-journals: A Review of Log Analysis Studies.” *Aslib Proceedings*, 57: 554-571.

Jansen, B. J. 2006. “Search Log Analysis: What It Is, What’s Been Done, How to Do It.” *Library & Information Science Research*, 28(3): 407-432.

Jansen, B. J. 2008. “Searching for Digital Images on the Web.” *Journal of Documentation*, 64(1): 81-101.

Jansen, B. J. and A. Spink. 2006. “How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction logs.” *Information Processing and Management*, 42(1): 248-263.

Jansen, B. J., A. Spink, and T. Saracevic. 2000. “Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web.” *Information Processing & Management*, 36(2): 207-227.

Jansen, B.J., D.L. Booth, and A. Spink. 2008. “Determining the Informational, Navigational, and Transactional Intent of Web Queries.” *Information Processing & Management*, 44: 1251-1266.

Jansen, B.J., A. Spink, and S. Koshman. 2007. “Web Searcher Interaction with the Dogpile.com Metasearch Engine.” *Journal of the American Society for Information Science and Technology*, 58(5): 744-755.

Jones, S., S. J. Cunningham, R. McNab, and S. Boddie. 2000. “A Transaction Log Analysis of a Digital Library.” *International Journal on Digital Libraries*, 3(2): 152-169.

Joo, S. and I. Xie. 2012. “Exploring Search Tactic Patterns in Searching Digital Libraries.” *In The Outreach of Digital Libraries: A Globalized Resource Network. Lecture Notes in Computer Science*, 7634: 349-350.

Jörgensen, C. and P. Jörgensen. 2005. “Image Querying by Image Professionals.” *Journal of the American Society for Information Science and Technology*, 56(12): 1346-1359.

Khoo, M., J. Pagano, A. L. Washington, M. Recker, B. Palmer, and R. A. Donahue. 2008. Using Web

Metrics to Analyze Digital Libraries. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 375-384). ACM.

Li, Q. and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Markey, K. 2007. "Twenty-five Years of End-user Searching, Part 1: Research Findings." *Journal of the American Society for Information Science and Technology*, 58: 1071-1081.

Pu, H. T. 2008. "An Analysis of Failed Queries for Web Image Retrieval." *Journal of Information Science*, 34(3): 275-289.

Shawe-Taylor, J. and N. Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Silverman, B. W. 1986. *Density Estimation*. London: Chapman and Hall.

Smith, G., C. Brien, and H. Ashman. 2012. "Evaluating Implicit Judgments from Image Search Clickthrough Data." *Journal of the American Society for Information Science and Technology*, 63: 2451-2462.

Spink, A., D. Wolfram, J. B. Jansen, and T. Saracevic. 2001. "Searching the Web: The Public and Their Queries." *Journal of the American Society for Information Science and Technology*, 52: 226-234.

Tjondronegoro, D.W., A. H. Spink, and B. J. Jansen. 2009. "A Study and Comparison of Multimedia Web Searching: 1997-2006." *Journal of the American Society for Information Science and Technology*, 60: 1756-1768.

Wolfram, D. and H. Xie. 2002. "Traditional IR for Web Users: A Context for General Audience Digital Libraries." *Information Processing & Management*, 38: 627-648.

Xie, I. and C. Cool. 2009. "Understanding Help Seeking within the Context of Searching Digital Libraries." *Journal of the American Society for Information Science and Technology*, 60: 477-494.

Xie, I. and S. Joo. 2010. "Transitions in Search Tactics during the Web-based Search Process." *Journal of American Society for Information Science and Technology*, 61: 2188-2205.

Yi, K., J. Beheshti, C. Cole, J. E. Leide, and A. Large. 2006. "User Search Behavior of Domain-specific Information Retrieval Systems: An Analysis of the Query Logs from PsycINFO and ABC-Clio's Historical Abstracts/America: History and Life." *Journal of the American Society for Information Science and Technology*, 57(9): 1208-1220.

Zhang, J., D. Wolfram, and P. Wang. 2009. "Analysis of Query Keywords of Sports Related Queries Using Visualization and Clustering." *Journal of the American Society for Information Science and Technology*, 60(8): 1550-1571.