Log Analysis of Academic Digital Library: User Query Patterns

Hyejung Han[1], Wooseob Jeong[1] and Dietmar Wolfram[1]
[1] University of Wisconsin - Milwaukee

Abstract
This study analyzed user queries submitted to an academic digital library for four weeks (July 2012 to August 2012). We examined users' query behaviors and compared external and internal users' query patterns for image-based collections. The results of this study identified the most frequently occurring queries, the mean of query strings, the term frequency, the most frequently used word pairs and the relationship between query terms. Transaction log analysis is useful to examine users' information seeking behavior effectively due to the richness of data. The query analysis of this study shows not only users' information seeking behaviors for image-based collections but also the differences between external and internal users' query patterns clearly.
Keywords: transaction log analysis, academic digital library, user queries, query analysis
Citation: Han, H., Jeong, W., & Wolfram, D. (2014). Log Analysis of Academic Digital Library: User Query Patterns. In *iConference 2014 Proceedings* (p. 1002–1008). doi:10.9776/14346

Contact: hanh@uwm.edu, wjj8612@uwm.edu, dwolfram@uwm.edu

1 Introduction and Literature Review

Users search in various types of IR systems for their information by formulating queries in a search box. Search facilities for different types of Web IR environments may appear similar, but the contents can be different with the search box and execute button (Wolfram, 2008). For the examination of users' information seeking patterns, transaction log data are used in many studies. Log data helps researchers identify and understand hidden and invisible user visit patterns (Zhang, 2008). Jansen (2006) noted 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. Many 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 websites. Wolfram (2008) analyzed query characteristics in a bibliographic databank, OPAC, search Engine, and specialized search system such as HealthLink. Wang, Berry and Yang (2003) investigated an academic website's query trends and patterns with transaction log data during a four-year period.

However, there is little investigation about transaction log analysis of users' query in academic digital libraries, in particular image-based digital collections, although there have been studies to analyze users' queries for image retrieval by surveys and interviews (Choi and Rasmussen, 2003). The purpose of this study is to identify and understand users' query searching behavior and their query formulation patterns in an academic image-based digital library. It also compares the characteristics of external queries and internal queries.

2 Research Questions

This study investigates the following research questions:

a) What characteristics of querying in an academic image-based digital library can be identified?
b) Are there different characteristics of querying between external and internal users for digital image collections?

c) What differences, if any, are there between experts and novices in query formulations in an image-based digital library in terms of advanced search options such as Boolean operations?

3 Method

Transaction log data for four weeks (from July 29, 2012 to August 26, 2012) were made available from the University of Wisconsin-Milwaukee libraries digital collections. The collections consist of over 54,000 photographic images, maps, and special collections. The masked IP address field and the referral field were extracted from the raw transaction log file and saved as a plain text file. There are 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 internal search interface uses "CISOBOX1" as a field name for the primary search function, while the external searches were identified with "q" for the query field. To compare the difference between the internal search and the external search, all the 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 masked 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 or icons. 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. Relationships among query term pairs were compared using the network analysis software Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

4 Results

4.1 Query Analysis

External and internal query data were analyzed to compare the different characteristics between two groups. There were 2,825 external query lines and 11,194 internal query lines. Table 1 shows the 30 most frequent query strings for external searches and those for internal searches. The most frequent query in external data is the two-word query "family notices" with 24 occurrences (Table 1).

External Query Strings	Frequency	Internal Query Strings	Frequency
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___fo rman	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___mercur y	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

Table 1: External and Internal Query Strings

The most frequent internal query is the three-word "tsybikoff g ts" (a Russian explorer named: Gombojab Tsybikov (Romanized as Tsybikoff)) with 1,481 occurrences (Table1). The top query strings are historical topic-related. As Jones and others (2000) found, the queries included users' space error such as "subjectarticles" and spelling differences between UK and American systems such as "pearl___harbour". Figure 1 shows that the mean of external query strings is 2.45 with standard deviation of 1.569. The mean of internal query strings is 1.96 with standard deviation of 1.341 (Figure 2).

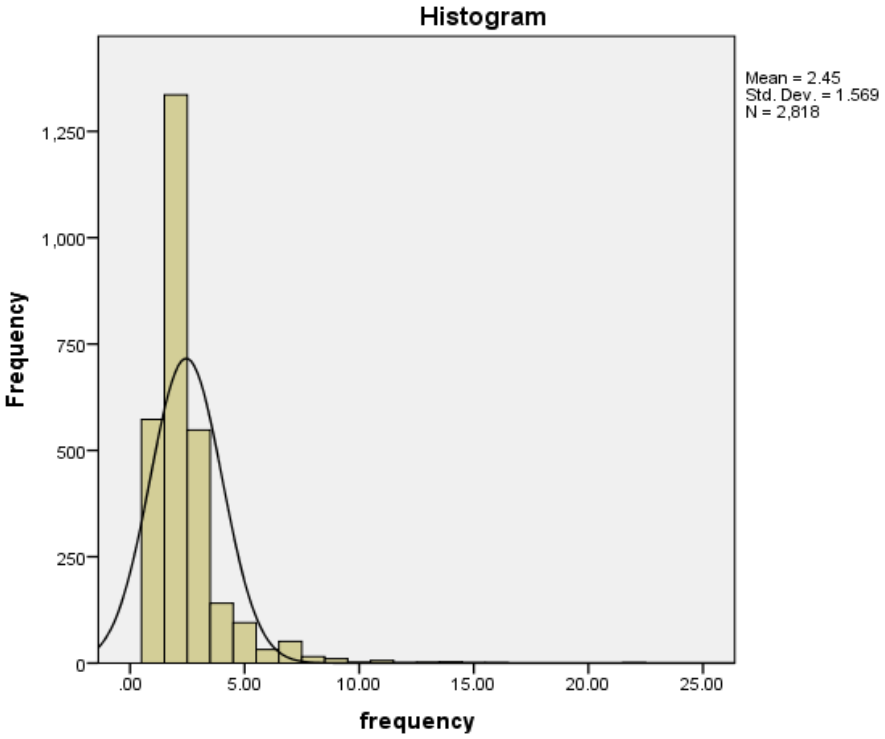

Figure 1: External Query Strings

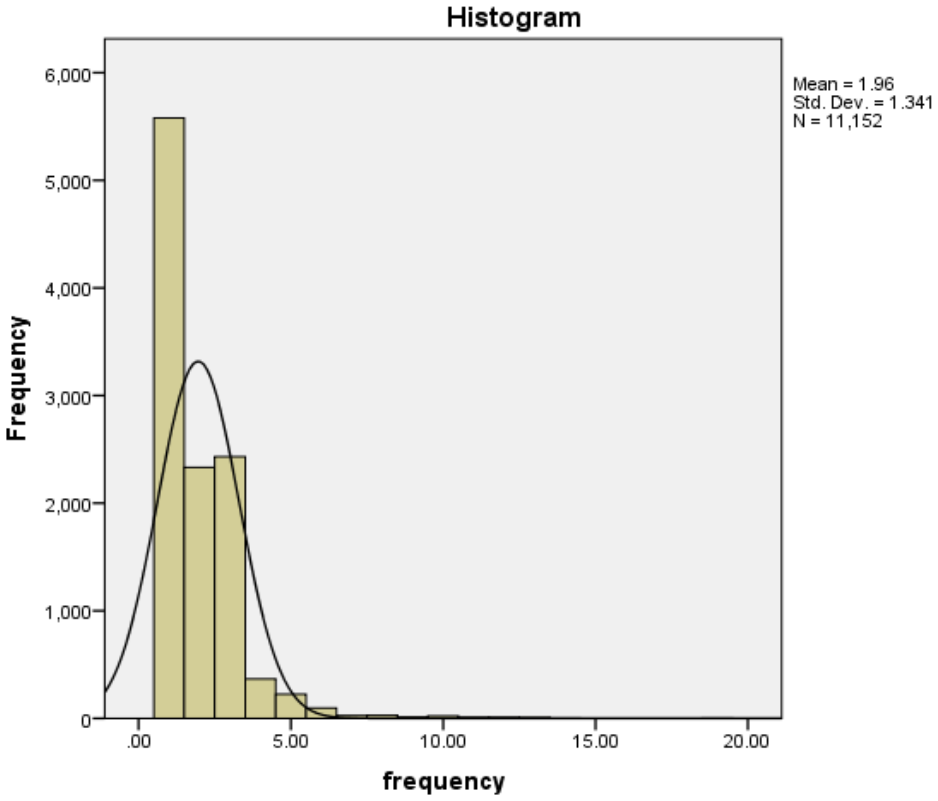

Figure 2: Internal Query Strings

4.2 Term Analysis

For the analysis of individual term used in queries, the most frequent 30 query terms from the external queries and the internal queries were identified respectively. Since the purpose of this study is to explore term frequency patterns, all observed words were contained in the analysis, including Boolean operator "AND" and "OR" as well as stop words (e.g. the, of, and, a, etc.). In the external data, besides the stop words, the most frequently used query terms are "Name related" for conducting people searches (e.g. John, William, James, Thomas, Henry, Peter, etc.). In the internal data, the most frequently occurring terms are "Geographic places related" such as Tibet, Milwaukee, Wisconsin and China rather than stop words. The data also revealed that the top individual terms are much more consistent with the top queries among internal queries than external queries.

4.3 Word Pair Analysis

To analyze the relationship among query terms, all the possible pairs of terms from each query string were created by a PHP script. The most frequent word 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 internal top query strings. 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 by Pajek, one of the most popular network analysis tools. The data were converted to Pajek format using "txt2pajek.exe" with a weighted option for the pair frequency. To avoid too complex displays and to obtain meaningful outputs, only the top 100 word pairs from each data set were included (the ties were counted). Figure 3 illustrates the relationship among external search terms and Figure 4 shows that among internal ones.

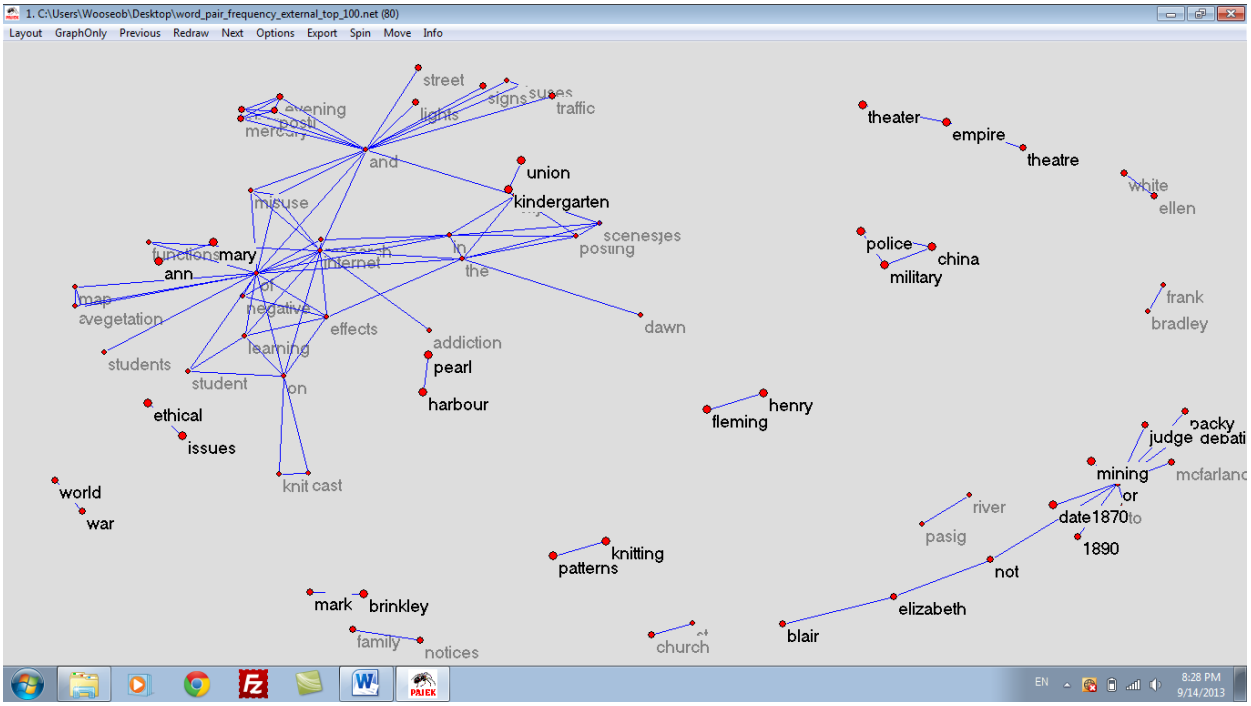

Figure 3: The relationship among external search terms

4.4 Differences between Experts and Novices

Experts tend to use various search tactics more frequently in their searches than novice users. We could observe the differences between experts and novice users by analyzing the search queries using Boolean

operators and wildcard characters, although their frequencies are low. Only 26 Boolean search queries were observed and only 4 queries using a wildcard character were identified.

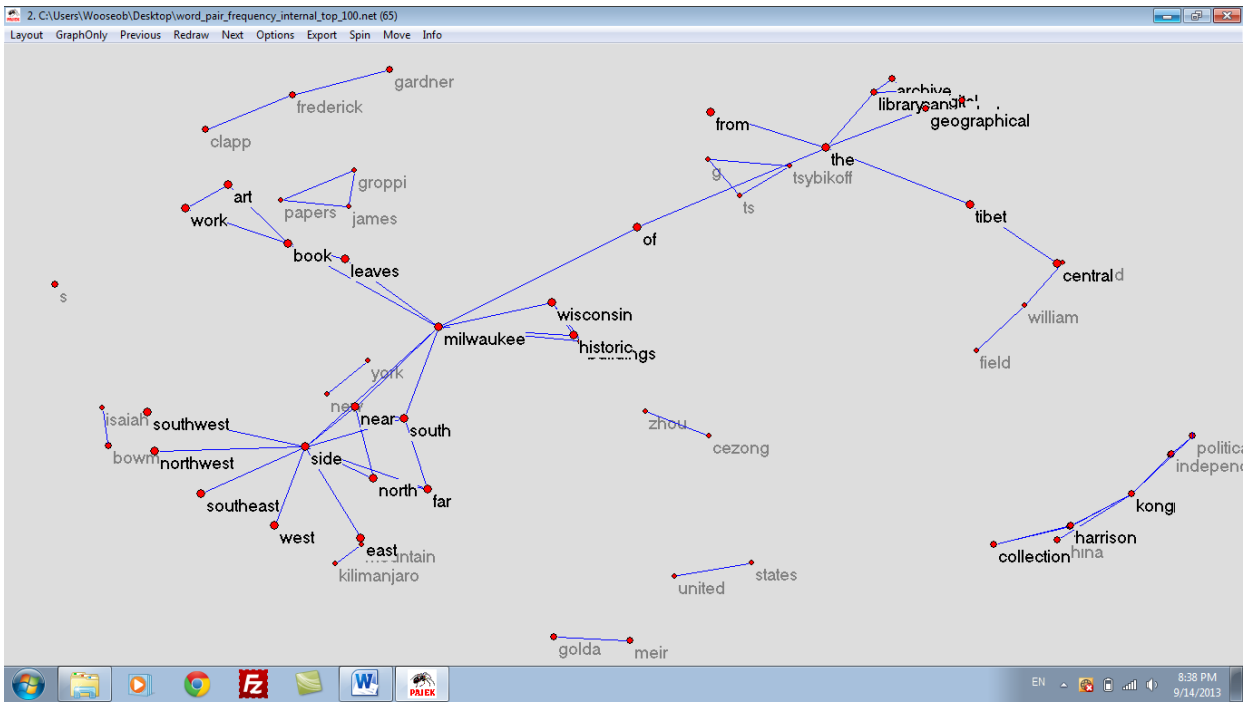

Figure 4: The relationship among internal search terms

5 Discussion and Conclusion

The query analysis in this study shows the differences between external and internal query patterns clearly. Internal query strings, query terms and word-pairs show the consistent relationship among each other while external ones do not show this consistency.

User queries between this academic digital library and the search engine Excite (Jansen) show some similarities. The average length of queries in both IR systems is short, with a mean of around 2 terms per a query. However, there exist differences in the search topic. Queries of this academic digital library include historical and geographical related topics while the search topics from search engines range from sexual topic to entertainment.

The limited size of the data collection does not allow broad generalizations. The limitation of this study is to analyze query term frequency literally. For example, Hong Kong is two-term place name but in the single term analysis, it was divided into two single terms and they were ranked 9th and 10th as the most frequently used query terms. Due to a small amount of data, spelling errors were not examined. In a larger study in near future with additional data, more sophisticated data examination will be conducted.

The results of this study are useful to understand users' searching behavior, especially with their query patterns in an image-based academic digital library. There is little research about transaction log analysis of digital libraries. Hence, this study will contribute to the better understanding of users' interaction with digital libraries, especially those with image-based collections.

6 References

Choi, Y., & Rasmussen, E. M. (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.

Jansen, B. J., Spink, A., & Saracevic, T. (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., and Spink, A. (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. (2006). Search log analysis: what it is, what's been done, how to do it. Library & information science research, 28(3), 407-432.

Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. International Journal on Digital Libraries, 3(2), 152-169.

Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2001). Searching the web: The public and their queries. Journal of the American society for information science and technology, 52(3), 226-234.

Wang, P., Berry, M. W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54(8), 743-758.

Wolfram, D. (2008). Search characteristics in different types of Web-based IR environments: Are they the same?. Information processing & management, 44(3), 1279-1292.

Zhang, J., Wolfram, D., & Wang, P. (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.

Zhang, J. (2008). Visualization for information retrieval, New York, NY: Springer.

7 Table of Figures

8 Table of Tables