

Log Analysis of Academic Digital Library: User Query Patterns

Hyejung Han¹, Wooseob Jeong¹ and Dietmar Wolfram¹

¹University 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

Copyright: Copyright is held by the authors.

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	89
Baumgarten	9	rman	81
china__military__police	7	manila	81
Gwalior	7	california	79
vegetation__map__of__asia	7	hong__kong	72
		turkey	70

shanghai__evening__post__and__mercur	7	east__side	69
y			
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	dwelling	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).

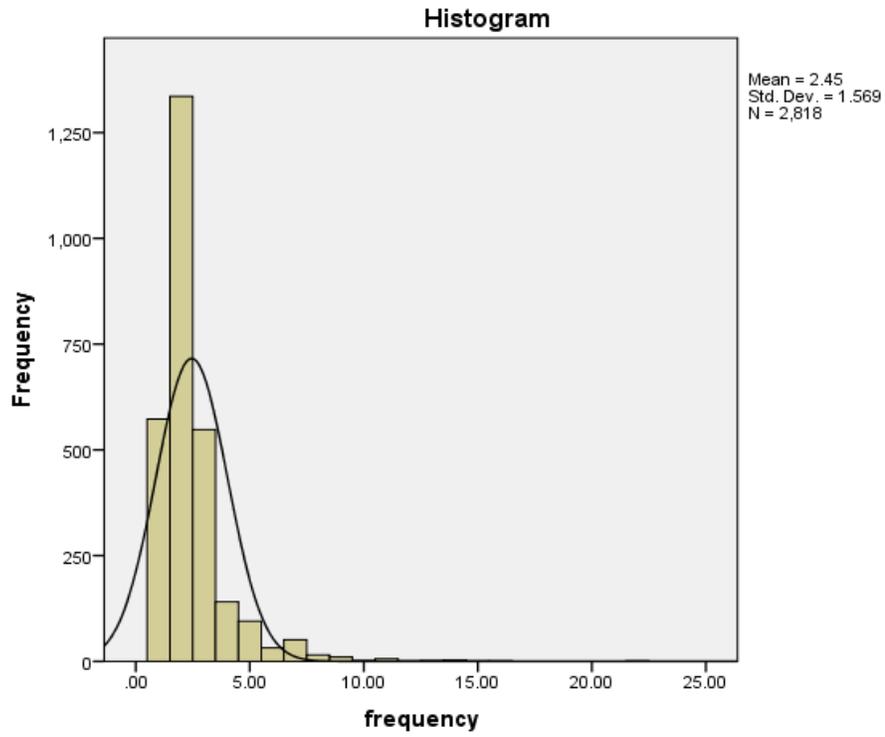


Figure 1: External Query Strings

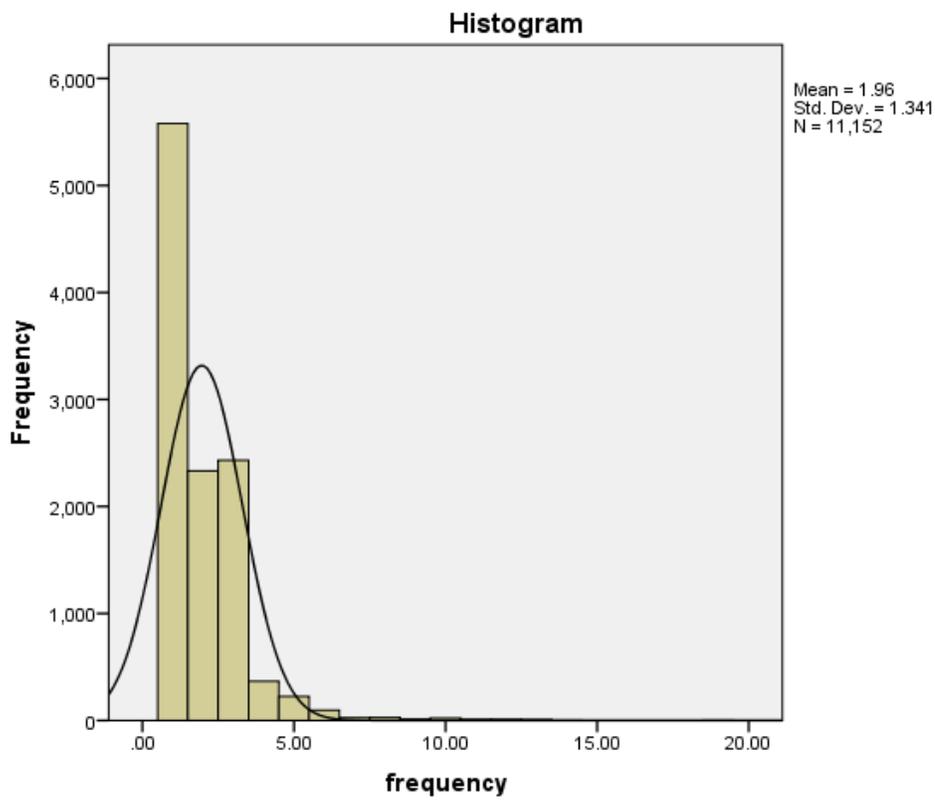


Figure 2: Internal Query Strings

- 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

Figure 1: External Query Strings	1005
Figure 2: Internal Query Strings	1005
Figure 3: The relationship among external search terms	1006
Figure 4: The relationship among internal search terms.....	1007

8 Table of Tables

Table 1: External and Internal Query Strings	1004
--	------