JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 10(1): 73-88, 2022

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2022.10.1.6

Utilization of Log Data Reflecting User Information-Seeking Behavior in the Digital Library

Seonhee Lee
Korea Institute of Science and Technology Information, Seoul, Korea
E-mail: wisdom@kisti.re.kr

Jee Yeon Lee*
Department of Library and Information Science, Yonsei University, Seoul, Korea
E-mail: jlee01@yonsei.ac.kr

ABSTRACT

This exploratory study aims to understand the potential of log data analysis and expand its utilization in user research methods. Transaction log data are records of electronic interactions that have occurred between users and web services, reflecting information-seeking behavior in the context of digital libraries where users interact with the service system during the search for information. Two ways were used to analyze South Korea's National Digital Science Library (NDSL) log data for three days, including 150,000 data: a log pattern analysis, and log context analysis using statistics. First, a pattern-based analysis examined the general paths of usage by logged and unlogged users. The correlation between paths was analyzed through a χ^2 analysis. The subsequent log context analysis assessed 30 identified users' data using basic statistics and visualized the individual user information-seeking behavior while accessing NDSL. The visualization shows included 30 diverse paths for 30 cases. Log analysis provided insight into general and individual user information-seeking behavior. The results of log analysis can enhance the understanding of user actions. Therefore, it can be utilized as the basic data to improve the design of services and systems in the digital library to meet users' needs.

Keywords: log data, information-seeking behavior, log pattern analysis, log context analysis, digital library, South Korea

Received: February 25, 2022 **Revised:** March 6, 2022
Accepted: March 7, 2022 **Published:** March 30, 2022

***Corresponding Author:** Jee Yeon Lee
https://orcid.org/0000-0001-6885-4684
E-mail: jlee01@yonsei.ac.kr

This paper is excerpted from Seonhee Lee's doctoral dissertation, *A study on utilization of log data reflecting user information seeking behavior on science and technology information science*, from Yonsei University Graduate School in 2019 (Advisor: Jee Yeon Lee).

1. INTRODUCTION

Users attempt to search for information stored in digital libraries to fulfill their information needs. Traces of tasks that users perform on digital libraries are left in the server of the digital library system in the form of log data. A transaction log is an electronic record of interactions between a web search system and users searching for information on the digital library during a searching episode. Therefore, log data can provide valuable insights into understanding the information search process users undertake on the web (Jansen, 2006). Computer-based methods of capturing user behavior, known as transaction log analysis, automatically capture the type, content, or time of transactions made by a user from the terminal connected to that system (Rice & Borgman, 1983).

To provide information services suitable for users' needs, librarians should understand users. Therefore, librarians have conducted surveys, in-depth interviews, and usability studies to understand users in the libraries. However, these methods are limited to the number of survey respondents and involve high costs and considerable time and effort. If the log data of the digital libraries are analyzed and the results are revealed of users' information-seeking behaviors, it will reduce time, cost, and efforts for research investigating users in the libraries. Furthermore, services and content that users need can be designed and provided.

This exploratory research was conducted to understand user information-seeking behavior by analyzing the log data left on the National Digital Science Library (NDSL, the previous service of ScienceON). The study was divided into a log pattern analysis that focuses on transaction log analyses alone and a log context analysis that combines transaction log data with user information. The research revealed significant insights into the information-seeking behavior of domestic researchers in the digital library. Furthermore, the data pertaining to identified individuals can be analyzed from various angles through log analysis. This study aims to understand the potential of log data and increase its utilization by using log data as a user research method.

1.1. Problem Statement

This study sought to answer the following research questions to understand log data that reflects the information-seeking behavior of researchers searching a digital library in Korea. Up to 85% of all NDSL users did not log in to use the NDSL services and content. In the log pattern analysis, information-seeking behavior was analyzed based on pageviews by logged and unlogged users. On the other hand, information-seeking behavior was analyzed utilizing log data with logged users in context log analysis.

- RQ 1: Can general information-seeking behavior patterns be assessed through log pattern analysis of the NDSL?
- RQ 2: Can contextual information-seeking behavior be assessed by analyzing and visualizing the log data created by identified users?

1.2. Definition of Terms

The followings are definitions of the core terms for log data analysis.

- Information use path: A user's record of using the services and content from a digital library is shown based on time; the details can be verified from the log data.
- Pageviews: This refers to viewing one page on the Internet and is frequently used in website evaluations. Pageviews contain an assortment of information, such as the services or content used on a specific page on the NDSL service. In this study, pageviews were regarded as a single task unit or the unit of analyzing each user's session.
- Log pattern analysis: Analysis of information-seeking behavior such as information use paths for services and content in the log data from website pageviews, stored in the log data.
- Log context analysis: Analysis to understand the information-seeking behavior of each identified user based on the information use paths, including the details of the user's usage time, service, and content accessed in the digital library.

1.3. Literature Review

1.3.1. Log Analysis and Information-seeking Behavior in the Context of the Digital Library

Libraries have traditionally performed user Transaction Log Analysis (TLA) on Online Public Access Catalogs (OPACs) for user research since the 1980s. The Online Computer Library Center (OCLC) studied log data from 1981 to 1983 by analyzing transactions and in-depth interviews. A TLA or online monitoring analysis of the interaction time and content is generated by the user

 https://doi.org/10.1633/JISTaP.2022.10.1.6

from the system terminal, which is collected automatically (Rice & Borgman, 1983). Since the emergence of the Internet in 1993, TLA has been redefined as the analysis of research interactions that are electronically recorded between an online search system and users who search for information on that system (Peters, 1993). Many attempts have been made to implement log analyses in user studies. Transaction log data provide a method for data collection while logging analysis servers as a method of data research. A TLA is defined as an online system and user study method that analyzes the system's performance and user behavior (Jansen et al., 2009).

On the other hand, online data monitoring, known as a type of TLA, is one of several ways to capture the detailed process of how users search for information at a reasonable price. TLA is a data collection technique rather than a research design. It can be used in empirical or field research either by itself or alongside other data collection methods (Borgman et al., 1996). TLA was reported as an appropriate technique for evaluating a search process because it is unobtrusive and provides a qualitative analysis of a quantitative model (Penniman & Dominick, 1980; Rice & Borgman, 1983). Log data are related to the qualitative evaluation of digital libraries and the qualitative evaluation process of interaction with services from a digital library (Koch et al., 2004). Transaction log data is also used as an evaluation tool for improving the user interface or system design. Further, it even supports qualitative evaluations of information behavior paths in an information retrieval system (Borgman, 1986; Rice & Borgman, 1983). As the information use behavior can be analyzed through log analysis, the log analysis method is argued to be an appropriate method for research in the web search field, which involves diverse, heterogeneous user classes (Park & Lee, 2007). Accordingly, studies that analyze search behaviors based on user log data have been conducted (Choi et al., 2018; Choo et al., 2000; Park, 2011; Park & Lee, 2007, 2013).

The berrypicking model as the process of acquiring information through web browsing was proposed by Bates (1989). The users may change their search query, but their searches may shift toward a new direction depending on the search results, or they may change information in each stage of their search process. Users will move from one search information source to another through various methods, such as footnotes, journal citations, reference literature, or local searches, or they may change the domain, and even when they are searching for a different information source. Agosti et al. (2012) analyzed the ten years of research publications about web service and digital library logs. This research showed that web search engine log analyses addressed matters related to representing, saving, organizing, and approaching important information items composing webpages. In the last ten years, this field of research was divided into three parts: analyzing the details of queries made by users from a search engine, how users interact with the search engine, and how the search engine organizes the results. A new log trend pertains to mobile devices, social networks, personalization, and multimedia service technology related to user information-seeking behaviors or system use (Agosti et al., 2012). A log analysis for the digital library is done by analyzing the access log statistics of the library user community to evaluate the digital library and user community (Bollen & Luce, 2002).

1.3.2. Log Analysis of Digital Library in South Korea

Many public libraries in South Korea have also utilized log data to analyze user information behaviors. For example, a study was conducted on the behavior of archive users based on weblog analyses (Lee & Yim, 2015). Another study proposed solutions for strengthening connections between websites on the National Archives of Korea and further developing the search service through analyzing search queries from users and the paths through which they end up on websites from the archives (Jin & Rieh, 2018). Finally, another study analyzed the status of library culture programs based on user participation log data through a case from the National Library of Korea in Sejong (Choi et al., 2018).

Various studies have been conducted through analysis of the NDSL. First, user log files were used to analyze the information needs and use behavior of users of the NDSL (Yoo, 2002). Search log keywords from the NDSL were analyzed, and the automatic category information of the document selected by the user and the categories between words were analyzed (Lee et al., 2012). Finally, an analysis was performed on the correlation between the categorized results of search queries and the categories of accessed documents. One study proposed a user login service policy after conducting a log analysis to promote use of the NDSL (Kim et al., 2013). Another study analyzed the characteristics of queries in one year's worth of log data from the NDSL. In this study, 70% of the queries consisted of one or two words, and 27.29% of the users spent one minute or less on the site (Park & Lee, 2013). However, the session length of users on the NDSL was longer than the time spent on general commercial search services. Moreover, log data from the NDSL reflected time passing over

three years. The queries and browsing requests of long-term users were analyzed, and it was found that while the simplicity of queries and shortness of usage duration remained the same as time passed, the research subject seemed to change with the passage of time. Search queries with a high frequency of use were analyzed, and the information use behavior was analyzed (Park & Lee, 2016). Since log analysis for libraries or information services in Korea primarily constitutes research regarding statistics on the number of uses or search terms, studies are needed to analyze log data from various aspects to understand users' information behavior. Therefore, different from the previous studies of log data of the NDSL, this study focused on revealing users' information behavior rather than search terms. Therefore, the utilization of log data could be expanded.

2. METHOD

2.1. The Process of Research

Table 1 shows the research procedure and details of this exploratory study. First, literature reviews were conducted. Second, log pattern and context analyses were conducted to answer the research questions of this study. For the log pattern analysis, general information behavior regarding the overall usage of the NDSL was analyzed. For the log context analysis, personal information was added to the log information to reveal individual information-seeking behavior based on the log data for each identified user. Finally, the visualization diagrams of the user information-seeking behavior were extracted by analyzing log data for specified users.

2.2. Data

The Korea Institute of Science and Technology Information (KISTI), a research institute funded by the government, has provided science and technology information content and services to users in South Korea for 60 years, since 1962. KISTI had provided services through the NDSL for ten years from 2010 to 2020, and NDSL moved to ScienceON in 2020. ScienceON is a service over the web that provides the knowledge infrastructure needed by researchers in one place by linking and converging science and technology information, research data, information analysis service, and research infrastructure.

As of 2018, NDSL had provided more than 120 million contents in NDSL, such as journal papers, technical reports, patents, and trends for searching and connecting to full-text. According to the 2017 annual report, the NDSL was used over 1 billion times throughout the year in terms of pageviews, and had full-text downloaded approximately 6 million times by users. Users of NDSL were professors, students, and researchers in private companies and public institutes in science and technology fields in South Korea. Users accessed NDSL through the digital library on the web, using portals such as Google and Naver, and other Open APIs from other institutes. However, the data analyzed in this study were limited to the NDSL data accessed through the web. Up to 85% of all NDSL users do not log in to use services and content on NDSL through the web.

The log data were collected and stored on the server. Table 2 shows information about the field of the log master table schema of the NDSL. About 150,000 pageviews of the randomly selected three-day NDSL logs in 2017 and 2018 were extracted and analyzed. There were 49,121 pageviews (32.7%) on September 13, 2017, 39,879 pageviews (26.5%) on January 16, 2018, and 61,750 pageviews (53.8%) on March 12, 2018. Pageviews refer to one usage of the webpage by one user and have become the basic unit of web data analysis. The term 'pageview' referred to is the most frequently used method for measuring the usage frequency for a certain website. Upon examining the frequency of IP usage, redundancies in the volume of use of the NDSL were eliminated for the log data of the three

Table 1. Research procedure and details

Research procedure	Methodology	Research details
Theoretical review	Literature review	▪ Assess the theoretical background and academic context of the research
Analysis of log data	Log pattern analysis	▪ Pattern log analysis of information use path - General usage frequency analysis for content and services - Establish research hypothesis and verify significance using a χ^2 analysis
	Log context analysis	▪ Demographic analysis of identified users ▪ Analysis of usage frequency and cross tables for content and services ▪ Analysis of usage and information-seeking behavior of identified users through the visualization diagram

 https://doi.org/10.1633/JISTaP.2022.10.1.6

Table 2. Log master table schema

Code of field	Name of field	Type
LOGSEQ	Log serial number	Produce (20)
USERID	User ID	Varchar (30)
USERIP	User IP	Varchar (15)
REGDATE	Input date and time	Time
SITECODE	Site code	Varchar (20)
SVCCODE1	Service-main-class	Varchar (10)
SVCCODE2	Service-sub-class	Varchar (10)
USECODE	Application code	Varchar (3)
CCTTCODE1	Content-main-class	Varchar (10)
CTTCODE2	Content-sub-class	Varchar (10)
RETCOUNT	Number of cases	Varchar (10)
COMMENT	Remark	Varchar (50)
YY	Input year	Varchar (4)
MM	Input month	Varchar (2)
DD	Input day	Varchar (2)
HH	Input time	Varchar (2)
LOGSID	Access identifier	Varchar (30)
URL	URL	Varchar (255)
R_URL	Reference URL	Varchar (255)
IP_CNAME	Organization name	Varchar (50)
IP_CITY	City name	Varchar (20)
IP_DISTRICT	District (gu) county (gun) name	Varchar(20)
IP_VILLAGE	Dong name	Varchar (20)
IP_NATION	Country	Varchar (10)
REGIONCODE	Area code	Varchar(2)
PAGECODE	Page code	Varchar (10)
CONTENT_ID	Content serial No.	Varchar (100)
JOURNAL_ID	Content journal serial No.	Varchar (100)
SESSION_KEYWORD	Search formula the given session	Varchar (4000)
SELECT_SEQ	Search result Top-N of selected content	Number
CONTENT_COUNT	Related content count	Varchar (2)
ACCESS_TYPE	Access type	Varchar (2)
LOGCID	Cookie ID	Varchar (255)
LOGSID	Session ID	Varchar (255)

Source: Kim et al. (2013).

days, which resulted in about 22,000 IPs. An IP from public places cannot be regarded as a singular user. Therefore, personal identification is the most accurate means of identifying a user when a person logs data in using an user ID to access the NDSL and then leaves log data.

3. RESULTS

3.1. Usage Statistics

As exhibited across the NDSL, the information-seeking behavior of users can be observed in the following log fields: 'Service-main-class,' 'Service-sub-class,' 'Content-main-class,' and 'Content-sub-class,' expressed by statistics using SPSS Statistics 22 (IBM Co., Armonk, NY, USA). From the log data containing the user log history, the frequent pageviews of September 13, 2017 (32,355 pageviews), January 16, 2018 (9,073 pageviews), and March 12, 2018 (41,242 pageviews), i.e., a total of 82,670 pageviews, were used for information seeking path analysis. More detailed information obtained by frequency analysis of the log data usage is available in Tables 3 and 4. In Table 3, frequently used items selected on a nominal scale for 'Service-main-class' included 'Search,' 'Detailed view,' and 'Full-text view.' Similarly, three frequently used items selected on a nominal scale for 'Service-sub-class' were 'Quick search,' 'Detailed view,' and 'Full-text view.' From 'Service-main-class' to 'Service-sub-class' in Table 3, the path from 'Full-text view' of 'Service-main-class' toward 'PDF full-text view' of 'Service-sub-class' was frequently taken. The main purpose of the NDSL users is to obtain the desired 'Full-text' that satisfies their information needs. Therefore, downloading or viewing a full-text version of a paper on the screen through 'PDF full-text view' is often considered to be the final step in satisfying user information needs. In Table 4, the most frequently used 'Content-main-class' in the NDSL was 'Paper,' while 'Domestic paper' was most frequently used in 'Content-sub-class.'

3.2. Log Analysis of Information-Seeking Behavior: Log Pattern Analysis

This study attempted to determine whether the information behavior link pattern representing the user's in-

Table 3. Frequency of log data usage: service-main-class and service-sub-class (unit: case, %)

Service-main-class	Frequency	Percent %	Cumulative %	Service-sub-class	Frequency	Percent %	Cumulative %
Detailed view	35,621	43.09	43.09	Detailed view	35,621	43.09	43.09
Full-text view	24,482	29.61	72.70	PDF full-text view	19,554	23.65	66.74
				Link to other institutions for full-text	4,928	5.96	72.70
Search	22,567	27.30	100.00	Quick search	22,567	27.30	100.00
Total	82,670	100.00		Total	82,670	100.00	

Table 4. Frequency of log data usage: content-main-class and content-sub-class (unit: case, %)

Content-main-class	Frequency	Percent %	Cumulative %	Content-sub-class	Frequency	Percent %	Cumulative %
Journal	2,488	3.01	3.01	Journal	2,488	3.01	3.01
Paper	59,130	71.53	74.53	All papers	1,8149	21.95	24.96
				Domestic paper	27,367	33.10	58.07
				International paper	6,595	7.98	66.04
				Theses and dissertations	7,019	8.49	74.53
Patent	5,924	7.17	81.70	Domestic publicized patent	2,418	2.92	77.46
				Domestic registered patent	3,506	4.24	81.70
Report	11,772	14.24	95.94	All reports	2,876	3.48	85.18
				Domestic research report	8,896	10.76	95.94
Researcher	3,356	4.06	100.00	Researcher	3,356	4.06	100.00
Total	82,670	100.00		Total	82,670	100.00	

 https://doi.org/10.1633/JISTaP.2022.10.1.6

formation use behavior appears in the log data to answer research question 1 (RQ 1).

- RQ 1: Can general information-seeking behavior patterns be assessed through log pattern analysis of the NDSL?

This study attempted to identify the characteristics of the paths with the link pattern of the moving route. First, users entered the NDSL and used services and content. Then for purposes of this study, the second-stage usage path for the search link log pattern analysis was analyzed, rather the 'Service-main-class' and 'Content-main-class,' followed by the 'Service-sub-class' and 'Content-sub-class.'

3.2.1. Information Seeking Path Analysis: Service-main-class and Content-main-class

Usage frequencies were studied to determine the relationship between variables in the information-seeking path 'Service-main-class' and 'Content-main-class.' Accordingly, a research hypothesis was formulated, and cross-tabulation analysis was performed using the χ^2 Test to determine a group difference or correlation between the attributes of 'Service-main-class' and 'Content-main-class.'

Research Hypothesis 1: There is a difference in the usage frequency of 'Content-main-class' types within the same 'Service-main-class.'

In order to determine the statistical significance of 'Service-main-class' and 'Content-main-class,' the log data for 'Service-main-class' of January 12 was tested. The items 'search,' 'detailed view,' and 'full-text view' were observed to be the most frequently used nominal variables. On the other hand, the more frequently used items in 'Content-main-class' were 'Journal,' 'Paper,' 'Patent,' 'Report' and Researcher.' Based on this sample data, a cross-tabulation analysis was conducted to test the independence of 'Service-main-class' and 'Content-main-class.' The results of the analysis are presented in Table 5. Furthermore, for the test of significance between 'Service-main-class,' and 'Content-main-class,' a cross-tabulation analysis was performed, and the χ^2 value was observed to be statistically significant at the 0.05 level.

As presented in Table 5, 'Detailed view' was most frequently accessed in 'Service-main-class,' followed by 'Full-text view' and 'Search.' In 'Content-main-class,' 'paper' was the most intensely used item, which is also observed to be the most widely used item in 'Search' and 'Full-text view.' This demonstrates that most NDSL services are related to theses and dissertations. This finding can also be confirmed by the log data analysis of the information-seeking paths. The item 'Paper,' which pertains to the content available in the NDSL services, is utilized with various services. Therefore, it is necessary to develop strategies to create more diversified services that use such contents (theses and dissertations), or to identify any negative user experience in existing services to improve the contents or services from the users' perspective.

Table 5. χ^2 test: frequency of content-main-class in the service-main-class

Category			Content-main-class					Total	χ^2
			Journal	Paper	Patent	Report	Researcher		
Service-main-class	Detailed view	Frequency	250	1,667	453	834	337	3,541	1,483.5*
		Expected frequency	98	2,257	246.7	786	153.4	3,541	
	Full-text view	Frequency	-	1,987	179	765	-	2,931	
		Expected frequency	81.1	1,868.2	204.2	650.6	127	2,931	
	Search	Frequency	1	2,129	-	415	56	2,601	
		Expected frequency	72	1,657.8	181.2	577.4	112.7	2,601	
Total		Frequency	251	5,783	632	2,014	393	9,073	
		Expected frequency	251	5,783	632	2,014	393	9,073	

*p<0.05.

3.2.2. Information Seeking Path Analysis: Service-sub-class and Content-sub-class

This analysis investigates and identifies users' information-seeking paths in the log data's 'Service-sub-class' and 'Content-sub-class' shown in the log data. Lower in the hierarchy, located in the lower layer of the corresponding main classification, the sub-classifications of services or contents can be reached by taking a path through the 'Service-main-class' or 'Content-main-class.' To determine if there is a group difference or correlation between the attributes of 'Service-sub-class' and 'Content-sub-class,' a research hypothesis was formulated and then a χ^2 test, i.e., a cross-tabulation analysis was performed.

Research Hypothesis 2: There is a difference in the usage frequency of 'Content-sub-class' types within the same 'Service-sub-class.'

Using the log data of January 12, 2018 as the sample data, a cross-tabulation analysis was performed to test for the group independence of 'Service-sub-class' and 'Content-sub-class.' The results are presented in Table 6. The results of the cross-tabulation analysis, which was conducted to test the significance of 'Service-sub-class' and 'Content-sub-class,' showed that the χ^2 value was observed to be statistically significant at the 0.05 level.

For 'Content-sub-class,' 'Detailed view' was the most frequently used in 3,541 out of 9,073 cases. Users frequently used 'Detailed view' contents, such as 'Domestic paper,' 'International paper,' and 'Theses and dissertations.' 'Detailed view' is a high-frequency service because it allows users to view abstracts and bibliographies or information about authors and references, as well as the full-text version of texts. A close look at the frequency of 'Service-sub-class' and 'Content-sub-class' for each information-seeking characteristic of a path reveals that 'all papers' of 'Content-sub-class' made through 'Quick search' in 'Service-sub-class' was the most frequently used item.

In this path, a brief list of all papers is displayed as the results for 'Quick search,' making it suitable for users who prefer performing a quick search. Another 'Service-sub-class' with a high usage frequency of 'Domestic paper' was 'PDF full-text view.' This path indicates the users' tendency to download a full-text paper after quickly searching and retrieving information with search terms that suit their information needs. In the information-seeking path for 'Service-sub-class' and 'Content-sub-class,' 'Domestic paper' through 'Detailed view' was also popular among users. 'Theses and dissertations' through 'Link to other institutions' for full-text were further observed to be popular among users. User information-seeking paths to access the contents available in other institutions can be observed from 'link to other institutions for full-text,' which is a popular item. Therefore, the process of satisfying user information needs could be identified along the path by analyzing the log data.

3.3. Log Analysis of Information-Seeking Behavior of Identified Users: Log Context Analysis

For the context log analysis, personal information was used in addition to log data to track user information-seeking behavior against the individual's information environment and demographic environment. Users who did not log in to use the NDSL make up 85% of all service users. The log data pattern analysis reviewed information-seeking behaviors based on pageviews without distinguishing individual users. Therefore, use behaviors demonstrated in log data were analyzed based on pageviews to understand the widespread use of the NDSL service. However, as users searched the NDSL individually, it was necessary to conduct a log analysis that focused on the users. Therefore, this paper tried to answer research question 2 (RQ 2) regarding whether contextual information-seeking behaviors can be analyzed from the log data of identified users.

- RQ 2: Can contextual information-seeking behaviors be assessed by analyzing and visualizing the log data created by identified users?

Of the service users studied over three days, the identities of 30 users who performed searches after logging in to NDSL were extracted randomly. Clearer identification of individual users was possible by identifying individuals who used the system by accessing it with their user IDs rather than by distinguishing them through session-by-session log analysis.

3.3.1. Demographic Statistics of Identified Users

The demographic characteristics of users' subjects and preferred information types, identified through the contextual analysis of log data are demonstrated in Table 7. The number of male users was 2.75 times larger than the number of female users among the 30 users selected for the analysis. As for the users' academic fields, seven users (23%) pursued electrical engineering, four users (13%) were in bioengineering, three users (10%) were in computer engineering, and three users (10%) were in medi-

 https://doi.org/10.1633/JISTaP.2022.10.1.6

Table 6. χ^2 test: frequency of content-sub-class in the service-sub-class

Category			Content-sub-class										Total	χ^2
			Journal	All paper	Domestic paper	International paper	Thesis and dissertation	Domestic publicized patent	Domestic registered patent	All reports	Domestic research report	Researcher		
Service-sub-class	Detailed view	Frequency	250	-	795	521	351	168	285	-	834	337	3,541	14,229.4*
		Expected frequency	98	792.7	973	208.4	283	89	157.7	156.1	629.9	153.4	3,541	
	Full-text view	Frequency	-	-	1,632	1	-	-	-	-	758	-	2,391	
		Expected frequency	66.1	535.2	657	140.7	191.1	60.1	106.5	105.4	425.3	103.6	2,391	
	Linking another library	Frequency	-	-	5	-	349	60	119	-	7	-	540	
		Expected frequency	14.9	120.9	148.4	31.8	43.2	13.6	24	23.8	96.1	23.4	540	
	Quick search	Frequency	1	2,031	61	12	25	-	-	400	15	56	2,601	
		Expected frequency	72	582.2	714.7	153.1	207.8	65.4	115.8	114.7	462.7	112.7	2,601	
Total		Frequency	251	2,031	2,493	534	725	228	404	400	1,614	393	9,073	
		Expected frequency	251	2,031	2,493	534	725	228	404	400	1,614	393	9,073	

*p<0.05.

Table 7. Frequency analysis on subjects and preferred information types for 30 users

Subjects					
	Subject area	Frequency	Percent %	Valid percentage %	Cumulated percentage %
Valid	Engineering	18	60.0	60.0	60.0
	Science	6	20.0	20.0	80.0
	Medicine/health	6	20.0	20.0	100.0
	Total	30	100.0	100.0	
Preferred information types					
	Information type	Frequency	Percent %	Valid percentage %	Cumulated percentage %
Valid	Paper	246	40.53	40.53	40.53
	Trend	1	0.16	0.16	40.69
	Report	305	50.25	50.25	90.94
	Journal/proceeding	22	3.62	3.62	94.56
	Patent	31	5.11	5.11	99.67
	Favorite information	2	0.33	0.33	100.00
	Total	607	100.0	100.0	

cine. The 30 logged-in users had 607 pageviews altogether. The 30 identified users and their log data were analyzed to assess whether they demonstrated individually unique information-seeking behavior in their circumstances and/or contextual information needs. The preferred types of information were 246 cases (40.54%) for 'Theses,' 305 cases (50.25%) for 'Report,' and 31 cases (5.11%) for 'Patent.' The type of information NDSL users preferred was analyzed by reflecting the cases where an individual requested more than one type of information. The frequency of 'Patent' or 'Journal' was also not high. On the other hand, 305 cases (50.25%) of 'Report' indicate that 'Report' was the most frequently used. The result showed that the most frequent use of content by identified users was 'Report.' The identified users had a higher preference for 'Report' than for 'Paper.' This shows a different aspect from the overall usage frequency statistics of NDSL in terms of log pattern analysis.

3.3.2. Visualization of Log Context Analysis

This study visualized user information-seeking behavior using log data, since log data was an electronic code and showed the user information-seeking behavior that the user interacted with while accessing NDSL. From the perspective of log analysis, log context analysis included user information and log pattern analysis. It would be desirable to analyze the user's information-seeking behavior based on the model, improve the systems, and provide the preferred information to all users. Since most users accessed the NDSL without logging in, it will be possible to provide various services only after analyzing personal information and researching the use of log information of users who have accessed the system after logging in. Only a few steps are currently presented, but adding various elements will make it possible to create subdivided service conditions that satisfy users.

3.3.3. Creating Visualization Diagram of Log Data

The visualization diagram was applied to User 27's cases, as shown in Fig. 1. This figure visualized the flow of User 27's behavior as time passed when they accessed the NDSL and used the services and content that leave log data. The x-axis shows pageviews, and the flow of time is seen from left to right. The y-axis shows services and content as one set and is displayed as arbitrary numbers. For example, User 27's first pageview in Fig. 1 is labeled 33. Label 33 here means that the user checked the 'Detailed view' and 'Report,' as shown in the center-right of Fig. 1. The user's second pageview was 40, which shows that the content report was viewed through 'Full-text view.' This figure visualizes how User 27 accessed services and content through 21 pageviews.

Microsoft Excel (Microsoft, Redmond, WA, USA) was used to visualize the flow of time, content, and services

 https://doi.org/10.1633/JISTaP.2022.10.1.6

for the visualization diagram. Given the nominal scales of log data, services and content were bundled as shown in Table 8, and an arbitrary number was attached to each pair to create an Excel cross table that visualizes this data. Combinations of seven Service-main-classes and seven Content-main-classes (7×7) resulted in 49 types of service and content pairs, each of which was given an arbitrary number from 1 to 49. Fig. 1 was created based on these categories.

The columns in Table 8 are the commonly used service types, and the rows at the top of the cross table are the list of user content types. The number in each cell shows that the service on the left and the content on the right were used. The service users' patterns of using NDSL can be seen by looking at the category to the left of the number to understand each user's combination of services and content. The content can be seen in the content category from the rows above. For example, the service and content that the user accesses can be understood by checking the number on the line graph in Fig. 1, and then finding the number it corresponds to. For example, number 49 means that the user searched to look up a 'paper'; the number 42 means that the user looked up a 'paper' through 'full-text view.' This information is indicated as numbers on the diagram, showing which service or content was used from the relevant pageview.

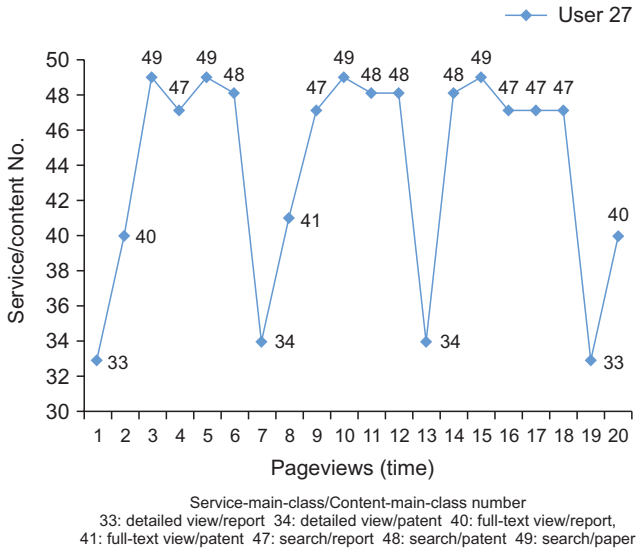

Fig. 1. Visualization diagram of log context analysis: example of User 27.

3.3.4. Application of Visualization Diagram Targeting Identified User on NDSL

The visualization diagram of log context analysis was used for 30 cases as shown in Fig. 2. The lines in the line graph represent the information-seeking behavior of 30 users; these 30 users had different use patterns, and each of their unique information-seeking behavior was visualized through a line. This is consistent with the fact that the information-seeking behaviors of users, which are regarded as the most critical factor in the user study model, are complex according to the situation, context, and social environment and are dynamic according to the context and process. Therefore, the users' information-seeking behaviors may be understood better through a user model rather than a technical description. The x-axis that represents time in the search formula indicates users' information-seeking behaviors over time. The y-axis represents a pair of services and content, and uses an arbitrary number based on the user's type of service and content the user availed. For example, User 3 viewed two 'Papers' through Search' before ending their search. User 2 only looked at 'Paper' using services and content. This shows constant changes in information-seeking behaviors in the line regarding individuality, circumstances, and context, as pointed out in various user studies.

Table 8. Visualization cross table for numbers of service and content pairs

Category	Paper	Patent	Report	Trend	Journal	Researcher	Saved information
Search	49	48	47	46	45	44	43
Full-text view	42	41	40	39	38	37	36
Detailed view	35	34	33	32	31	30	29
Send	28	27	26	25	24	23	22
Document delivery service	21	20	19	8	17	16	15
Translate	14	13	12	11	10	9	8
Save	7	6	5	4	3	2	1

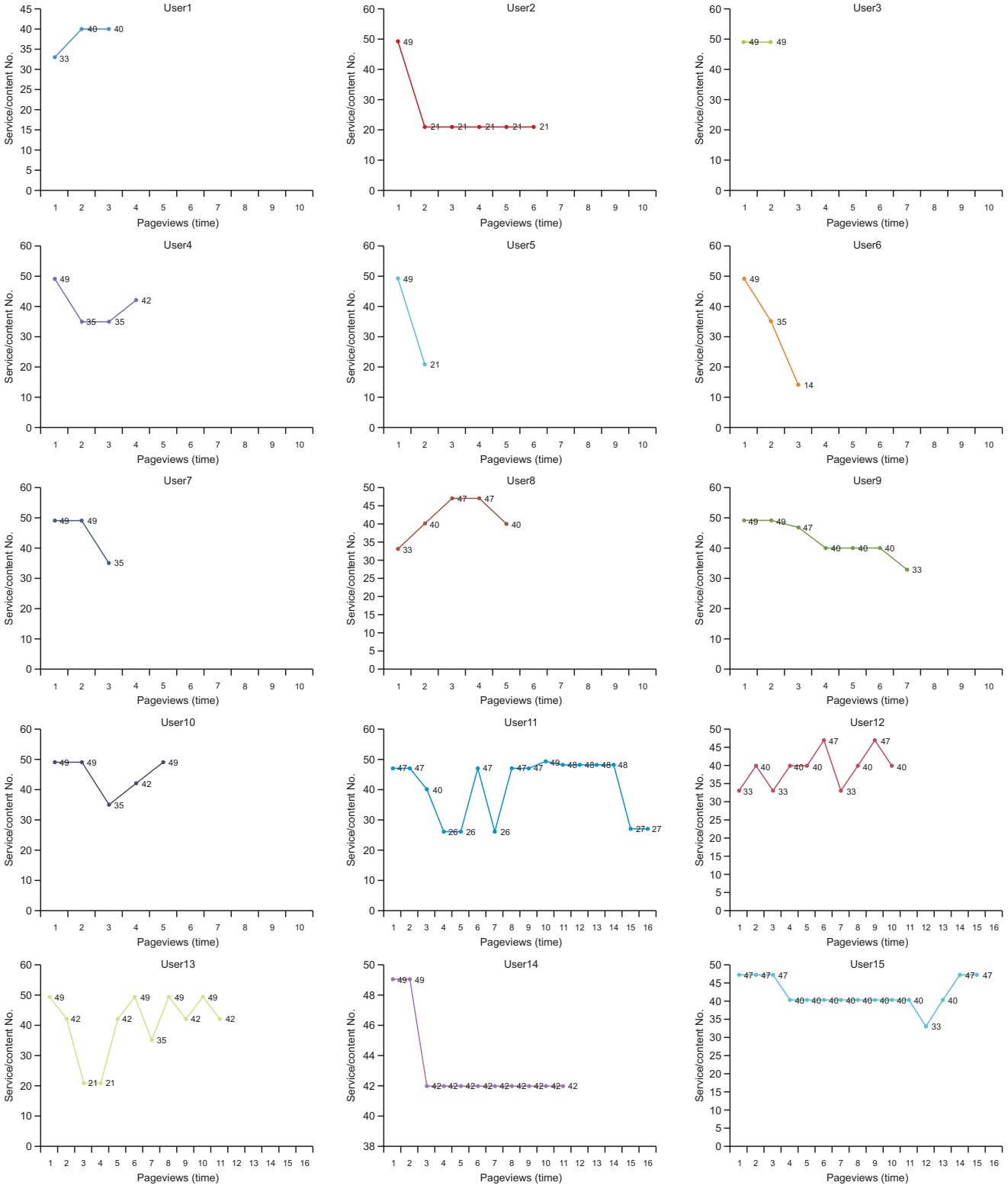

Fig. 2. Visualization diagram of context analysis: 30 users.

 https://doi.org/10.1633/JISTaP.2022.10.1.6

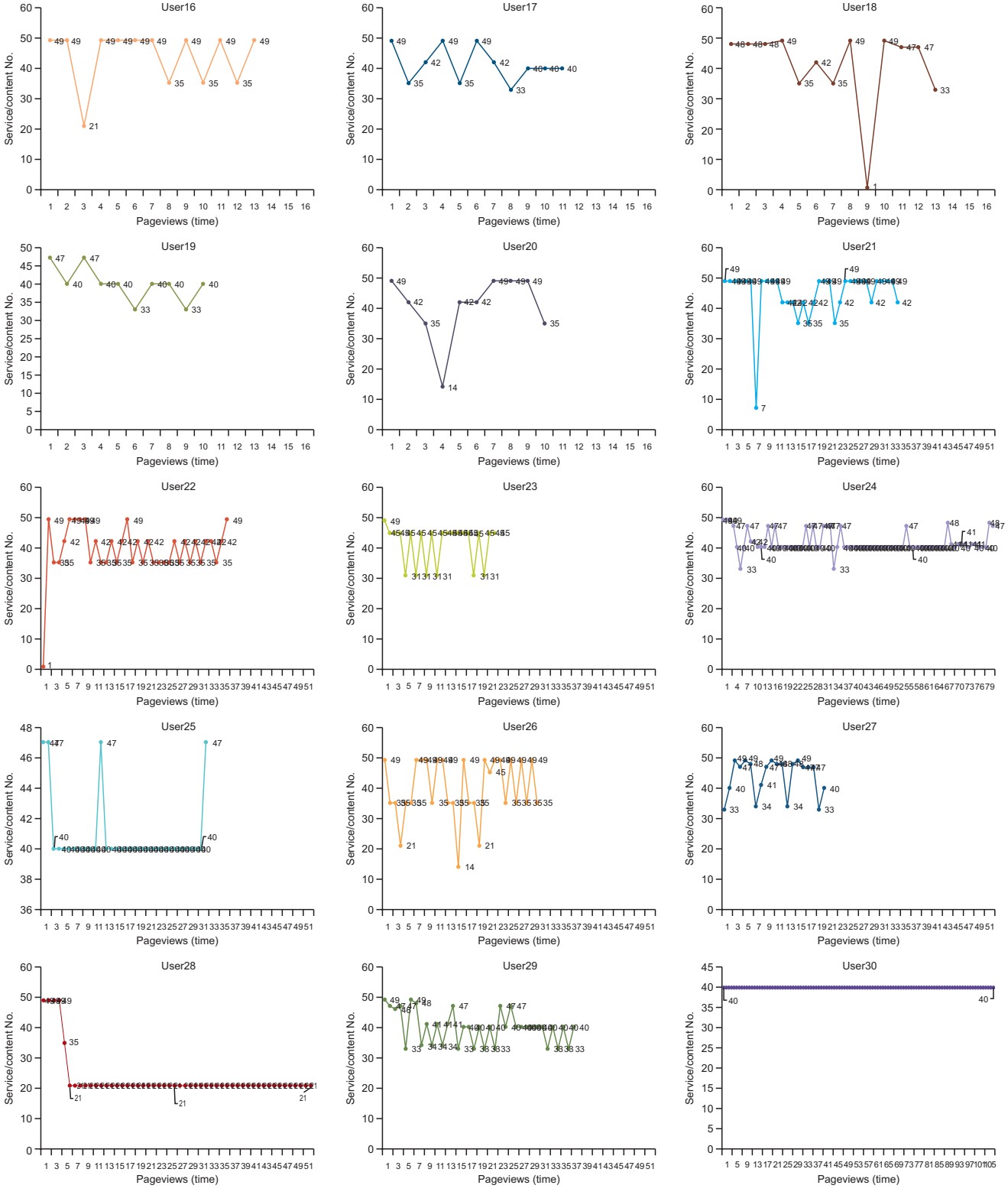

Fig. 2. Continued.

In the future, a method that can group these use behaviors while emphasizing each user's individuality will be sought. With respect to content, users could be grouped into users who prefer 'Papers,' users who prefer 'Reports,' and so on. In terms of services, although there are various types of users, such as those who only use 'Search' and 'Full-text view,' users who typically use 'search,' 'Detailed view,' or 'Full-text view,' and users who use customized services like 'Save' after logging in, if users can be grouped, it would be best to establish a way to provide services based on these groupings. If services can be provided based on understanding the information behavior of users, it will become easier to help users fulfill their information needs.

4. DISCUSSION

Log data from the NDSL were collected randomly and analyzed statistically to answer the two research questions. To conduct an exploratory study on log data that contain the usage records of NDSL users, a log pattern analysis and log context analysis were conducted on the log data. Analyses of the information use path were conducted using log data in pageviews for the log pattern analysis. Pageviews are seen as task units of information use, and they were therefore used as the base unit for data analysis. Log pattern analysis was performed to analyze general usage patterns in the NDSL rather than the patterns for particular individuals. Services and content were divided into main-class and sub-class, and the multi-frequency path was analyzed under these categories: 'Service-main-class,' 'Content-main-class,' 'Service-sub-class,' and 'Content-sub-class.' The correlation between paths was also analyzed through a χ^2 analysis. Logs of users who utilized the NDSL after logging in with their ID were analyzed for the log context analysis. The results of log pattern and context analysis results showed differences in preferred services or content. However, since log pattern analysis is generalized using an average through quantitative analysis, it is necessary to consider the associated limitations of not reflecting individual characteristics.

In the context analysis, users frequently looked up content or services that could only be accessed by logging in. However, the information-seeking behavior of log-in users shown in the log context analysis also revealed the characteristics of a small number of information services that could be overlooked in the overall log data analysis. Therefore, it is desirable to perform a complimentary analysis of the entire log data and the individual user data can be considered in a complementary role as it is possible to analyze the use of NDSL from different angles. Therefore, the results of log context analysis results can be used to assess information behaviors for individuals to provide personalized services that cater to the individual needs of users.

The log pattern analysis and log context analysis were conducted through a frequency analysis to objectify user information-seeking behavior. Log pattern analysis is similar to seeing a forest, and log context analysis is akin to seeing a tree. The log context analysis revealed information-seeking behavior that is not reflected in the pattern analysis, particularly the diversity of information-seeking behaviors and changes in context and circumstances. If the log pattern analysis and log context analysis are not used in conjunction, only a skewed understanding of information-seeking behavior will be obtained. Since the log data tracking analysis applied in this study was useful to assess the patterns and context of user information-seeking behavior, it validates the potential for log data as a method in user studies. In the future, a way may be derived to develop library services and contents through log analyses based on attempts to expand the utilization of log data, and may be developed as a user study method.

This paper provided detailed records and descriptions of the information behavior of individuals who use the NDSL services by analyzing and visualizing user information behaviors that appear in the log data. Three components of duration (time), service, and content were visualized two-dimensionally. The visualization tool is based on Excel; it can be used in similar cases that analyze log data. If user demands and usage patterns are researched through a highly usable log analysis method, it will be possible to make significant contributions to establishing a way to improve and design systems and services. Furthermore, the model should be structured to play a role in showing the dynamics of the information-seeking process according to the users' situations and contexts. The model used in this study was derived inductively through actual log analysis. Visualization by model schematic diagram can be used to understand individuals' distinctive use patterns by making it easy to recognize the information-seeking behavior of NDSL users. It can be said that the usability of log data can be expanded.

Understanding information-seeking behavior through log analysis can increase the usability of log data by expanding the domain in log analysis, focusing on a statistical analysis of the frequency of use or simple frequency analysis of existing query words (Lee et al., 2012; Park & Lee, 2013, 2016). In addition, it is possible to supple-

 https://doi.org/10.1633/JISTaP.2022.10.1.6

ment the usability evaluation conducted in the electronic library to analyze the information-seeking behavior of users. Since the usability evaluation targets some users, the results may be slightly distorted, but log data analysis is more likely to reflect the characteristics of the entire population of users since it is analyzed for all users. Log usage tracking analysis can be performed with existing log data, which is inexpensive. User research through log analysis has the advantage of analyzing realistic user information pursuit behaviors of users without falsehood. In addition, the log data provides clues to the users' contextual understanding, as this essential element of user research can track their situation and context.

There is an increasing number of libraries introducing log analysis systems or attempting log analysis using Google Analytics. However, there are cases in which log analysis is of limited use. It does not provide much help in grasping the user information-seeking behavior. Furthermore, there are many cases in which log data is not fully utilized. This study attempted to examine the extent to which log data reveals information-seeking behavior and how much we can use the log. Since log data is an electronic sign of the information-seeking behavior of the users of the corresponding digital library, if this sign is well interpreted and analyzed properly, the user information-seeking behavior can be analyzed in various ways (Jansen et al., 2009; Park & Lee, 2013; Peters, 1993). Based on this, the contents or services desired by users can be better provided, so the usability of the log can be improved.

A limitation of this study was that the dataset was limited to the data from three days of using a digital library providing information resources for science and technology. Therefore, the results of this study may not apply to general cases. Consequently, it is necessary to have a contextual understanding of the model's content analyzed by the log system, the characteristics of users, and the provided content. Moreover, the proportion of users who log in may be smaller than that of university libraries or research institutions. Since information behaviors may differ depending on the structure of the digital information service, any attempts to apply the model to other digital library services must be preceded by a thorough review.

5. CONCLUSION

Libraries are making efforts to improve their services and systems using log data, which are traces of interactions between users and systems. This study expanded the scope for utilization of log data by exploring its potential as a user study method. For this purpose, the study developed a new model that can extract the contextual and detailed characteristics of information-seeking behavior. Analysis of the transaction log of the digital library's services and content indicated the users' information search behaviors and provides insight into users' information needs, information usage patterns, and search and access methods. The majority of users' overall access pattern and information-seeking behavior was analyzed through log pattern analysis. The specific users' information-seeking behavior was complemented by log context analysis. Based on the configuration of the information-seeking behavior derived through exploratory research, a visualization tool showing the users' research situation and situation characteristics was built. Surveys or in-depth interviews, which are conducted by digital library librarians to provide information services tailored to the user's information needs, require time and considerable cost. There is also a limited sample size. On the other hand, log data can provide accurate information on users' information search behavior, which is the basic data necessary to improve and design library information services and systems, so it can be an efficient tool to grasp users' information pursuit behaviors. By tracking, analyzing, and linking the components of various types of log data provided by the log system, it will be possible to grasp the contents of the richer user information-seeking behavior in more detail. The results of log analysis for information-seeking behavior can enhance the understanding of users. Therefore, it can be utilized as the basic data to improve the design of digital library services and systems for users.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Agosti, M., Crivellari, F., & Di Nunzio, G. M. (2012). Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, 24(3), 663-696. https://doi.org/10.1007/s10618-011-0228-8.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5), 407-424. https://doi.org/10.1108/eb024320.

Bollen, J., & Luce, R. (2002). *Evaluation of digital library impact and user communities by analysis of usage patterns*. http://

www.dlib.org/dlib/june02/bollen/06bollen.html.

Borgman, C. L. (1986). The user's mental model of an information retrieval system: An experiment on a prototype online catalog. *International Journal of Man-Machine Studies*, 24(1), 47-64. https://doi.org/10.1016/S0020-7373(86)80039-6.

Borgman, C. L., Hirsh, S. G., & Hiller, J. (1996). Rethinking online monitoring methods for information retrieval systems: From search product to search process. *Journal of the American Society for Information Science*, 47(7), 568-583. https://www.proquest.com/openview/20c67ac9d22a200ad3f6af80a29769c6/1?pq-origsite=gscholar&cbl=41136.

Choi, D. W., Gang, J. Y., Yang, D., Lee, H., & Oh, H. J. (2018). An analysis of library culture program management based on users' participation logs: A case study of National Library of Korea, Sejong. *Journal of Korean Library and Information Science Society*, 49(1), 293-320. https://doi.org/10.16981/kliss.49.201803.293.

Choo, C. W., Detlor, B., & Turnbull, D. (2000). Information seeking on the web: An integrated model of browsing and searching. *First Monday*, 5(2). https://doi.org/10.5210/fm.v5i2.729.

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. https://doi.org/10.1016/j.lisr.2006.06.005.

Jansen, B. J., Spink, A., & Taksa, I. (2009). *Handbook of research on web log analysis*. IGI Global.

Jin, J. Y., & Rieh, H. (2018). Analysis of users' inflow route and search terms of the Korea National Archives' web site. *Journal of the Korean Society for information Management*, 35(1), 183-203. https://doi.org/10.3743/KOSIM.2018.35.1.183.

Kim, G. Y., Kwon, N. H., Yu, S. Y., & Choi, Y. (2013). *An analysis of NDSL use statistics and its application for developing virtuous circle of NDSL information services*. https://scienceon.kisti.re.kr/commons/util/originalView.do?cn=TRKO201500002288&dbt=TRKO&rn=.

Koch, T., Ardö, A., & Golub, K. (2004, June 7-11). *Browsing and searching behavior in the Renardus web service: A study based on log analysis*. Paper presented at the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tuscon, AZ, USA.

Lee, H., & Yim, J. H. (2015). A case study analysing the users of archives through web analytics. *Korean Journal of Archival Studies*, 45, 83-120. https://doi.org/10.20923/kjas.2015.45.083.

Lee, T. S., Jeong, D. H., Moon Y. S., Park, M. S., & Hyun, M. H. (2012). An analytic study on the categorization of query through automatic term classification. *The KIPS Transactions: Part D*, 19D(2), 133-138. https://doi.org/10.3745/KIPSTD.2012.19D.2.133.

Park, M., & Lee, T.-S. (2013). Understanding science and technology information users through transaction log analysis. *Library Hi Tech*, 31(1), 123-140. https://doi.org/10.1108/07378831311303976.

Park, M., & Lee, T.-S. (2016). A longitudinal study of information needs and search behaviors in science and technology: A query analysis. *Electronic Library*, 34(1), 83-98. https://doi.org/10.1108/EL-04-2014-0058.

Park, S. Y. (2011). Trends and changes of web searching behavior. *Journal of the Korean Society for Library and Information Science*, 45(1), 377-393. https://doi.org/10.4275/KSLIS.2011.45.1.377.

Park, S. Y., & Lee, J. H. (2007). Applications of transaction log analysis for the web searching field. *Journal of the Korean Society for Library and Information Science*, 41(1), 231-242. https://doi.org/10.4275/KSLIS.2007.41.1.231.

Penniman, W. D., & Dominick, W. D. (1980). Monitoring and evaluation of on-line information system usage. *Information Processing & Management*, 16(1), 17-35. https://doi.org/10.1016/0306-4573(80)90003-5.

Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech*, 11(2), 41-66. https://doi.org/10.1108/eb047884.

Rice, R. E., & Borgman, C. L. (1983). The use of computer-monitored data in information science and communication research. *Journal of the American Society for Information Science*, 34(4), 247-256. https://www.dhi.ac.uk/san/waysofbeing/data/health-jones-rice-1983c.pdf.

Yoo, S.-R. (2002). User-oriented evaluation of NDSL information service. *Journal of the Korean Society for Library and Information Science*, 36(1), 25-40. https://doi.org/10.4275/KSLIS.2002.36.1.025.