# Analysis on Science & Technology Information Service Utilize based on NDSL usage logs

Hyejin Lee, Mihwan Hyun, Teasuk Yi, Hyesun Kim
Korea Institute of Science& Technology Information, Korea
E-mail : hyejin@kisti.re.kr, mhhyun@kisti.re.kr, tsyi@kisti.re.kr, hskim@kisti.re.kr

## 1. Introduction

With the expansion of the digital economy, a big data environment is emerging in which immeasurably large amounts of information and data are being produced. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization[4]. Accordingly, attention is being drawn again to web analytics, which analyzes user connections in bulk and uses log data obtained from online channels such as web and mobile. In other words, web analytics may be a movement to utilize more actively data which is achieved in real time. Web analytics can be utilized in various ways beginning from its utilization for optimizing operation costs to offering service actively by analyzing individually the online activities of website users.

NDSL (National Digital Science Links, http://www.ndsl.kr) is an STI (Science and Technology Information) integrated service platform provided by KISTI. NDSL links to science and technology information, both domestically and internationally. The contents of NDSL are various such as papers, patents, technical reports, trend analysis reports, standards, human resources, and fact data. Also, regarding usage, there are about 0.5 million members in total, 3 million net visitors a year, and 100 million page views a year.

The purpose of this study is to review the current overall use of NDSL based on usage log data, which is one type of big data, and to conduct a basic analysis of usage log in order to discover the key service in information service.

## 2. Usage log system and analysis method

The usage log table of NDSL is composed of 21 fields including serial number, connection IP, use code, number of search, search formula, CN number, date and time, cookie, connection URL, reference URL, information use approach type, and service code. The CN number means a unique identification number by contents, reference URL means the page URL which reach immediately prior to current page. It is possible to analyze the pattern of use by a user based on reference URL and connection URL.

Regarding the data of a usage log table, the key in analyzing and interpreting this is to analyze service code, use code, and data approach type code by log. The service code has a three-step code system for diverse services offered by NDSL. The first step is the division of page view, search, and the use of a DB, while the second step code is composed of codes by contents. Meanwhile, the third step code consists of codes for various functions. The utilization code shows how a user utilized a service, and it is made up of emails, viewing the PDFs, and download. The data approach type code means a classification of information use through a variety of channels, and is composed of approach, mobile connection, and mailing service from external websites.

This study aimed at approximately 250 million usage log data of NDSL as of 2013 and included all web and mobile services. Also, the study excluded log data used for any service test, which was confirmed through IP and user ID, and data of which the number of connections was abnormally high (over 10,000 times) in the cookie information.

## 3. Usage log analysis results

This study analyzed the utilization code and the service code for considering the actual utilization of service and function regarding the usage log.

**Overall result**

As it is possible to use the available NDSL services without log in, the use of services through log in was found to be very low at about 14%. The number of user IDs for which the number of connections was over 10,000 after logging in for 1 year was 16. Also, user IDs for which the number of connections ranged between 5,000 to 10,000 totaled 42; thus, a total of 58 IDs was discovered to be for users who record very high use.

**Use code**

As a result of the analysis, NDSL users were shown to use the order of search execution, simple view of search result, and viewing original text. This indicates that the service use pattern is concentrated on search and viewing of an original text. In particular, application for copying the original text numbered 79,759, export numbered 61,406, print numbered 10,364, and email numbered 4,347. These accounted for 0.06% of the total logs, which implies that this is a very low function in terms of use.

**Service code**

As a result of the analysis, it was shown that frequency of use was in the order of article DB, report DB, and patent DB. Among article DB, Domestic journal articles were used most frequently. Domestic journal articles were the most frequently used contents among individual contents. Among report DB, Domestic research reports were used most frequently; while among patent DB, Korean registration patents were used most frequently.

In reviewing the use rate compared with the number of creation, it was found to be in the order of Domestic Journal Articles (=2.07), KS (Korean Standards) (=1.98), Domestic Reports(=1.67), Domestic Proceeding Articles (=1.4), and Journal/Proceedings (1.2). Regarding KS, it ranked upper 15th in terms of total use; however, in the use rate compared with the number of creation, it ranked 2nd; thus, usage by contents appeared high.

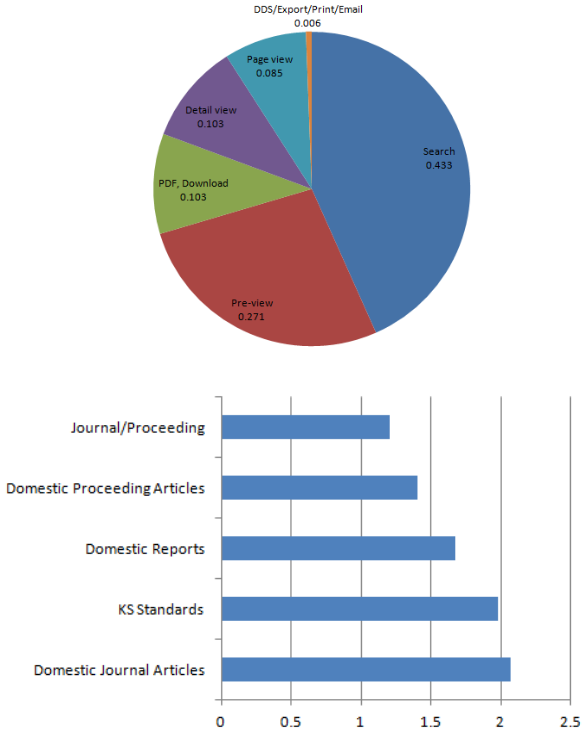

Figure 1. Use rates by use codes(top) and contents(bottom)

## 4. Proposal

This study analyzed arithmetic statistics by information service and by function based on weblog data, which is a type of big data. As a result of the analysis, domestic information was discovered to record a high use rate and high use by contents. Afterward, web measurement indicators for various phenomena occurring at NDSL based on the web log analysis of this study will be developed and the key service analyzed through measurement indicators.

## 5. References

[1] Hwang, Ok-Kyung (2007). The current status of the electronic journal usage statistics at the academic library. Journal of Information Management, 38(4), 68-87.

[2] McDonald, J. D. (2006). Understanding online journal usage: A statistical analysis of citation and use. Journal of the American Society for Information Science and Technology, 57(13), 39-50. doi: 10.1002/asi.20420

[3] McKinsey, “Big Data: The Next Frontier for Innovation, Competition, and Productivity”, Mckinsey & Company. 2011.

[4] Wikipedia http://en.wikipedia.org/wiki/Big data [cited by 2014. 5.10]