International Journal of Advanced Smart Convergence Vol.7 No.4 101-107 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.4.101

IJASC 18-4-12

Web Server Log Visualization

Jungkee Kim[†]

[*,†] *Department of Glocal IT, Sungkonghoe University, Seoul, Korea*
jake@skhu.ac.kr

Abstract

Visitors to a Web site leave access logs documenting their activity in the site. These access logs provide a valuable source of information about the visitors' access patterns in the Web site. In addition to the pages that the user visited, it is generally possible to discover the geographical locations of the visitors. Web servers also records other information such as the entry into the site, the URL, the used operating system and the browser, etc. There are several Web mining techniques to extract useful information from such information and visualization of a Web log is one of those techniques. This paper presents a technique as well as a case a study of visualizing a Web log.

Key words: *Web Log, Data Analysis and Visualization, Big Data, Web mining.*

1. Introduction

Web log files have a lot of useful information including Web documents, hyperlinks between documents, the client's access patterns on the Web server, etc. The Web leaves these invisible footprints. These resources enable an individuals or a company to promote business, understanding marketing dynamics, new promotions floating on the World Wide Web, etc.

Before the Web was introduced to the world, traditional log analysis can be also found starting from 1980's. The office of the Online Computer Library Center (OCLC) carried out a study of online public access catalogs (OPACs) to figure out what extent current system features were used [1]. The study made the most effort to address integrating different formats of each log file. Jansen and Pooch [2] reviewed the studies of Web transaction logs of Web search engines. Agosi and di Nunzio [3] suggested general methodology for gathering and mining information from Web log files and managed to retrieve, save and analyze the data extracted from log files.

Visualization of Web logs provides intuitive knowledge in temporal and spatial context. There are many Web visualization tools such as Graylog, NXLog, Splunk, Syslog-ng, and Rsyslog. They typically provides useful visualization in temporal as well as spatial aspects. However, many of them use a backend relational databases resulting in the drawback of the serial delay of the process and the limit of scalability. Although the existing visualization tools work well with limited capacity, there are Web servers and sites whose data

Manuscript Received: Oct. 8, 2018 / Revised: Oct. 14, 2018 / Accepted: Oct. 21, 2018
Corresponding Author: jake@skhu.ac.kr
Tel: +82-2-2610-4715, Fax: +82-2-2610-4390
Department of Glocal IT, Sungkonghoe University, Seoul, Korea

amounting to terabyte or even more. It is not possible for such visualization tools to process such data in a timely manner. In this paper, we propose to utilize some sampling methods and suggest a future study on how to choose an appropriate sampling method for each analysis.

Park, Lee and Bae [4] focused on the analysis of transaction logs for Web search engines and proposed a log cleaning, session definition, and query classification. Suneetha and Krishnamoorthi [5] analyzed Web log data of NASA site to retrieve information of a web site, errors and visitors of the site. They asserted that the achieved results of the study can help system administrators and Web designer to increase the effectiveness for their Web maintenance. Goel and Jha [6] developed a log analyzer tool called Web log expert to predict the behaviors of users who access an astrology website.

The rest of this paper is organized as follows. Section 2 describes the characteristics of Web log. Section 3 presents the performance study with Web log visualization. In Section 4, we summarize and suggest future work.

2. Web logs

Web servers run NCSA's Hypertext Transfer Protocoal daemon. In turn, each copy of this daemon maintains a W3C's standard Common Log Format [7] for Web server log files, but other variables of formats exist. The Common Log Format is a well-defied set of facts about each hit that a Web server processed. The format contains lines of eight fields – host, ident, authuser, date-time, zone, request, status, and bytes. Each field is separated by a single space. If a field is not known or available, a dash is used as the field value instead. The ident field is rarely used, and authuser is used only at sites that require registration. While the referrer log records the point of entry into a site, the agent log identifies the operating system and browser used by the client.

Agent logs record the browser type, the browser version and the operating system from which the client accessed the Web site. These help us determine prevalent operating systems and browsers that are used to visit the Web site. Generally, Web servers keep typical four types of log files as follows:

Table 1. Types of Web Server Logs [8]

Types of log files	*Actions*	*Extracted Knowledge*
Access log	Records all users' request processed by server and record information about users	Users' profiles Frequent patterns Bandwidth usage
Agent log	User browsers and browser's version	Agent version and Operating system used
Error log	List of errors for users request made by server	Types of errors Generated errors IP address Date and time of error occurred
Referrer log	information about link and redirects visitor to Site	Browser used Keywords. Redirect link content.

In our experiment, the host, the date-time, and the request fields are used. The status item is also used for noise eliminating because the status indicates whether an access was successful or not. If a non-exist page is accessed, it would produce a failure code. Since such accesses are noises, we can remove the noise by checking the status item. The possible status codes are classified in table 2.

Table 2. HTTP/1.0 server status codes

Success Code				*Failure Code*			
200	OK	300	Multiple Choices	400	Bad Request	500	Internal Server Error
201	Created	301	Moved Permanently	401	Unauthorized	501	Not Implemented
202	Accepted	302	Moved Temporarily	402	Payment Required	502	Bad Gateway
203	Non-Authoritative Information	304	Not Modified	403	Forbidden	503	Service Unavailable
204	No Content			404	Not Found		

The host field shows the remote host which visited the Web page. The request field has a method, the visiting page and HTTP version. We extract only the visiting page information. The date and time information is used too. The remote host, access page, and access time are used to represent each axis in the visualization.

3. Web log analysis and visualization

Web Analysis is widely used in many applications to estimate the usefulness of a Web site. Jansen [9] addressed that there are three stages for the log analysis – collection, preparation, and analysis. The Web log data usually collected in the form of a file by Web server demon. To analyze the log data, we need to transform the log data to a proper format. Debian Unix system provides Wwwstat package for processing the log files to prepare the analysis. Gwstat Perl script in the Unix system that allow us to collect the daily, hourly, monthly, weekly, subject, and other temporal and special statistics. Note that recent AWstats, Google Analytics and other tools are freely available as open sources to show Web usage statistics.

This paper provides an example of Web log analysis and visualization from our previous experiences. Cyclic visualization of visits to the Web site of the NPAC [10] is generated from logs of the NPAC site. To gain some meaningful page request data, some logs of accessing all day long from a specific machine were filtered. The shell script collected the log files and Perl programs generated daily, hourly, monthly, weekly, domain, and subject graphic files. However, these tools only shows simple counts ordered by the number of visits or total transferred size. Using an advanced visualization tool, we could gain some different patterns that are hard to find by utilizing the basic Unix packages. For pre-processing the log data, a database management system was used. Only three data fields are extracted from the log files – remote host, access date, and access pages. A Java program generated a distinct remote host name file and a distinct access page file. Those files are coverted numbers with an order from 1 to the number of each items. When producing an ASCII input file with a Perl program, the access time converted to minutes. For example, “01:03” converted to “63”. The ASCII input file is necessary for the visualization tool – Scivis[11]. An overview of the visualization process is presented in Figure 1.

From a day data set, 17,368 access points were extracted to generate several interesting two or three dimensional graphs. Figure 2 shows three dimensional graphs. One has some labels due to the tool’s adjusting the ratio of each axis automatically and the other provide a three dimensional graph without the scale labels for access time, remote host, and access pages. All those three dimensions are linearized by the alphabetic ordering and converted to numbers. A small number of remote hosts accessed some particular web pages within a short time, and many remote hosts usually accessed the Web pages in the working hours.

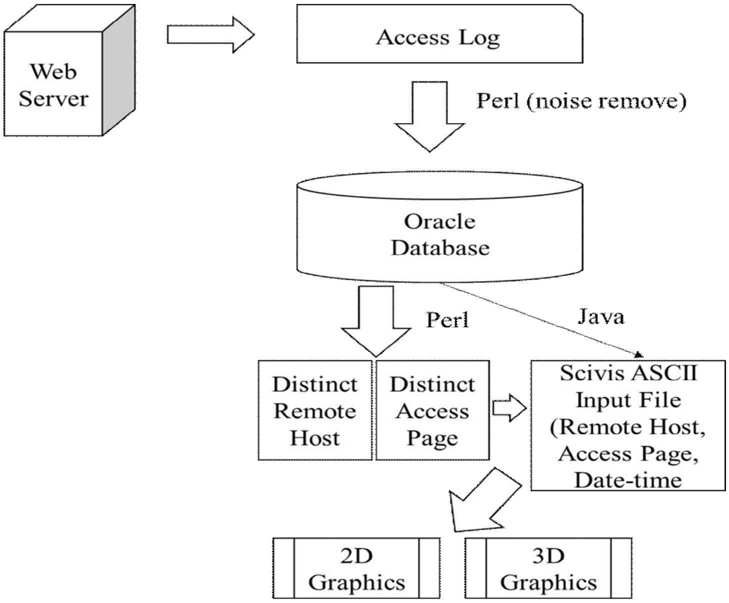


Figure 1. Preparation chart on the Web log analysis

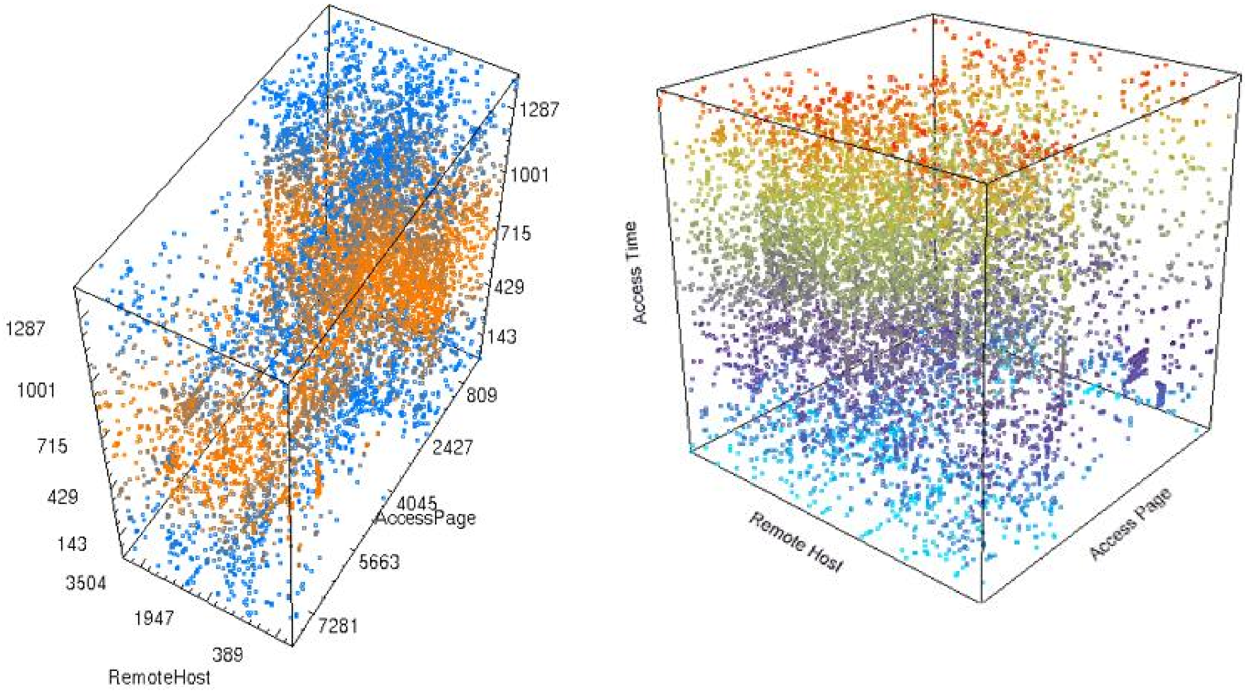
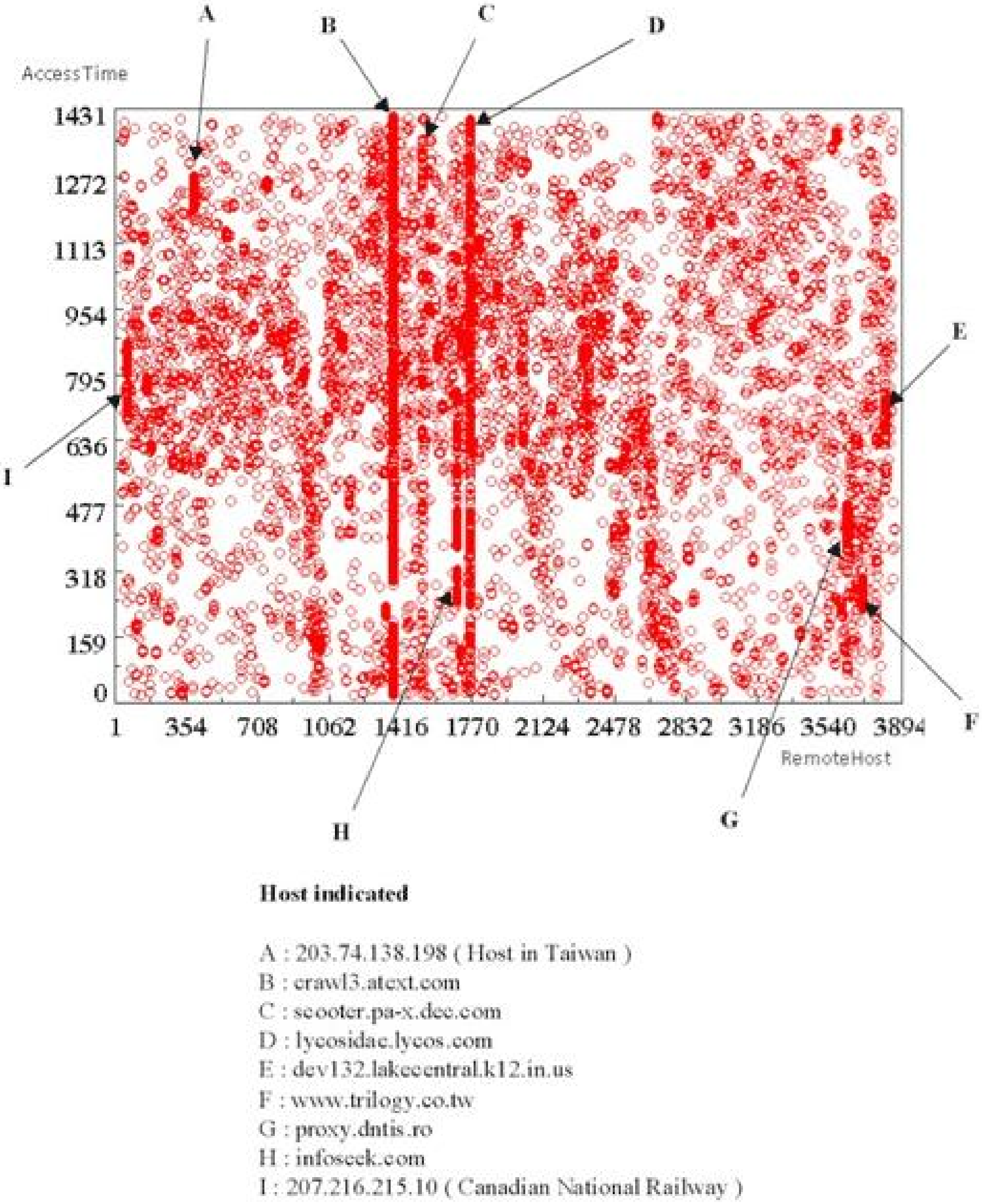


Figure 2. Three dimensional graphs of a day access

The relation between access time and remote host is shown as a graph in Figure 3. The long lines show some typical robots sites. A several short thick lines represent that particular remote hosts accessed many times within a short time period. They are also kinds of robot sites which were turned on in a short time.

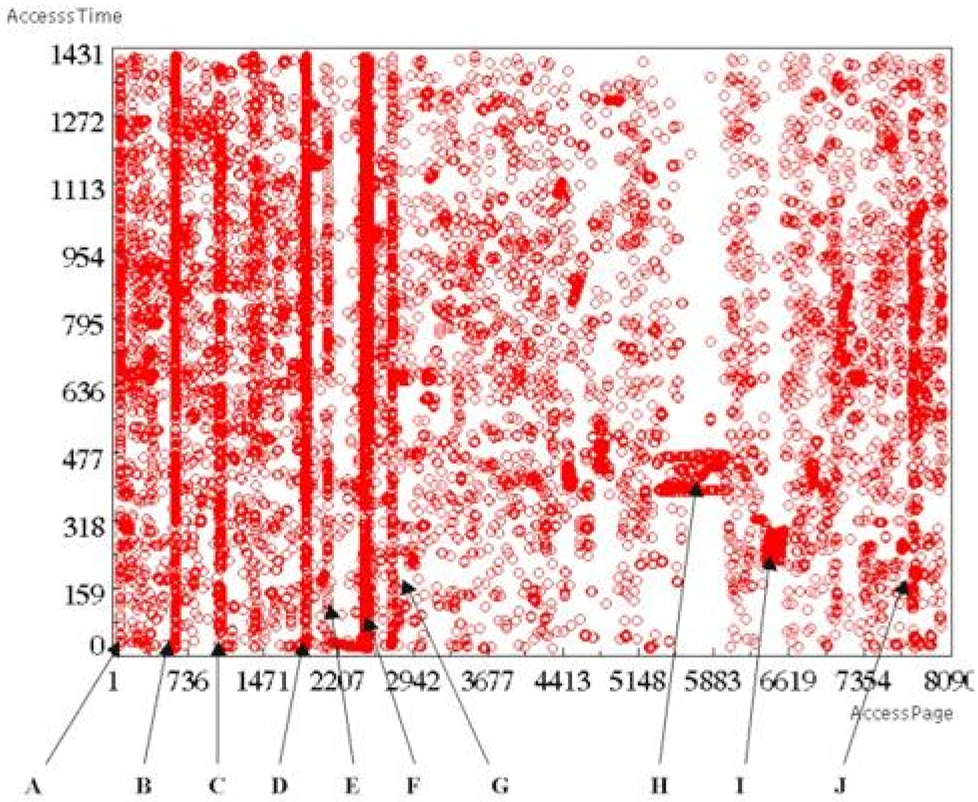


Host indicated

A : 203.74.138.198 (Host in Taiwan)
B : crawl3.atext.com
C : scooter.pa-x.dec.com
D : lycosidae.lycos.com
E : dev132.lakecentral.k12.in.us
F : www.trilogy.co.tw
G : proxy.dntis.ro
H : infoseek.com
I : 207.216.215.10 (Canadian National Railway)

Figure 3. Two dimensional graph between access time and remote hosts

The relation between access time and access page is also shown as a graph in Figure 4. It has the vertical lines which represent that some web pages are accessed continually. The most frequently referred web sites such as Visual human and kids Web can be indicated as popular pages. The reason why the “robots.txt” is accessed open is that it is usually accessed when a robot visits. That file addresses to the robot which pages are allowed to be visited. A small number of interesting patterns are also found not a line form but geometrical figures. The “Z” shape indicates that some consecutive pages are access one by one within a short time. The rectangular shape marked with “I” illustrates the Java tutorial materials are accessed one by one.



Host indicated

A : / (root)
B : /cgi_tools/count.cgi?*
C : /projects/3Dvisiblehuman/*
D : /projects/visihuman*
E : robots.txt
F : /textbook/kidweb
G : /users/dpk/ATM*
H : /users/gcf/familyphotos/*
I : /users/gcf/javatutorial98.1/*
J : /users/mahesh/homepage/atm_tutorial/*

Figure 4. Two dimensional graph between access time and access pages

4. Conclusion and future work

Web log analytics with visualization techniques are very useful for discovering patterns in data sets. Due to many points in the 3D graphs, it is difficult to find some interesting clusters in the graph. However, some explicit patterns are visualized in the graph and we decided to divide it as 2D graphs which enable us to get a particular point. Some vertical and horizontal lines have been showed up and can be interpreted as a particular meaning. The same processes can be done on the commercial sites and it will produce very useful information to decide the future policy to promote their business.

In the previous experiment, only 17,368 points were used to process the analysis on a day data only though some unimportant accesses such as graphic file accesses were removed. Additional graphs can be obtained from visualization covering a longer periodic range like weekly or more. Even a big portal site produces millions accesses and this approach may not be possible to adequately handle the large data points. As a result, one of the most significant challenges of implementing a big data analysis is ensuring effective evaluation. As per our data mining needs, in the instance of a user trying to plot Web Accesses over the past few months, the query would need to read at least 250,000 rows of access data in order to obtain the relevant information. Not only is this process time consuming but also places enormous resource overheads on the database.

An appropriate use of sampling aggregation and corresponding visualization can be useful for management of larger data sets. In the future work, various sampling skills such as simple, stratified, cluster, and systematic random sample can be utilized and compared each other.

References

[1] J. Tolle, “Transactional log analysis: Online catalogs,” in Proc. 6th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR’83. ACM, New York: Association for Computing Machinery, pp 147–160, 1983.

[2] B. J. Jansen and U. Pooch, “Web user studies: A review and framework for future work,” Journal of the American Society of Information Science and Technology, 52(3), pp. 235-246, 2001.

[3] M. Agosti and GM Di Nunzio, “Gathering and mining information from web log files,” in Lecture notes in computer science, vol. 4877. Springer, pp 104–113, 2007.

[4] S. Park, J. H. Lee, and H. J. Bae, “End user searching: A Web log analysis of NAVER, a Korean Web search engine,” The Journal of Library & information science research, Elsevier, Vol. 27, pp. 203-221, 2005.

[5] K. R. Suneetha and R. Krishnamoorthi, “Identifying user behavior by analyzing web server access log file,” International Journal of Computer Science and Network Security, Vol. 9(4), pp. 327-332, 2009.

[6] N. Goel and C. K. Jha, “Analyzing users behavior from web access logs using automated log analyzer tool,” International Journal of Computer Applications, vol. 62(2), pp. 29-33, 2013.

[7] W3C, “Logging control,” http://www.w3.org/Daemon/User/Config/Logging.html.

[8] T. A. Al-Asadi and A. J. Obaid, “Discovering similar user navigation behavior in web log data,” International Journal of Applied Engineering Research, 11(16), pp. 8797–8805, 2016.

[9] B. J. Jansen, “Search log analysis: What it is, what’s been done, how to do it,” Library & Information Science Research, Vol. 28, pp. 407–432, 2006.

[10] Syracuse University, Northeast Parallel Architecture Center, http://surface.syr.edu/npac/.

[11] B. Ki and S. Klasky, “Scivis,” Concurrency: Practice and Experience, Vol. 10, pp. 1107-1115, 1998.