Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 9 (2010) 1208–1213

Procedia
Social and Behavioral Sciences

WCLTA 2010

Log analyzer programs for distance education systems

İhsan Güneş[a] *, Muammer Akçay[b], Gökhan Deniz Dinçer[a]

[a] *Anadolu University, Eskisehir, 26470, Turkey*
[b] *Dumlupınar University, Kütahya, 43100, Turkey*

Abstract

Distance education is mostly performed through the internet nowadays. The number of students using the internet services is increasing and parallel to this, the amount of data regarding the usage of services on the internet is increasing. According to the results of data analysis, some information can be acquired. This statistical information can be useful for determining the profiles of students, which is important for decision makers to improve the online learning system. There are several software programs used for data processing and analyzing. In this study these programs are explored and compared in terms of some features.

Keywords: Web log; log analysis; distance education; user profile;

1. Introduction

The Web, which was introduced not more than twenty years ago, is now the environment in which people of different cultures, languages and all ages carry out their daily digital lives in. Web users are encircled by a network, infrastructure of devices, and applications at all hands regardless of time, location or reason (Jansen, Sping & Taksa 2008).

Nowadays, internet is becoming the most important media for collecting, sharing and distributing information. Web-based applications and environments for social networking, search engines, electronic commerce, distance education, news broadcasts, etc., are becoming common practice and widespread (Zaiane&Luo 2001). Mostly distance education is a web based technology and it is used for presenting online course. Online learning environments have been improving rapidly. Web-based learning environments , such as Virtual-U, Web-CT, sakai and moodle generally include course content delivery tools, synchronous and asynchronous conferencing systems, quiz modules, eportfolio, virtual workspaces for sharing resources, white boards, grade reporting systems, logbooks, assignment submission components, etc.

E-Learning portals serve for many users. Therefore, in order to design a e-Learning portal, one should bear in mind that each learner may have different learning style. For this reason, different kinds of learning components are needed to be used. These components not only be the component which the user can study on his own such as e-book, e-audio book, e-video, e-exercise program, e-test etc. but also be the components which gives user

* İhsan GÜNEŞ. Tel.: +90-222-335-0580/2426.
E-mail address: ihsang@anadolu.edu.tr.

1877-0428 © 2010 Published by Elsevier Ltd.
doi:10.1016/j.sbspro.2010.12.308

opportunity to make online team works (Mutlu, Kip & Kayabaş, 2008). Although more than one type of technology are used to improve these learning components, they are not enough to satisfy learners standalone. Using these components all together provides varying styles of the learner (Moore & Kearsley, 2005).

Web servers can be defined as web use-based course delivery systems, as in web-based application or any web site, in order to provide line in to resources and applications. Each request sent to a Web server is kept in an access log which mainly registers the origin of the request, the resource requested and a time stamp. The request may be for a web page containing an article from a course chapter, the answer to an on-line exam question, or a participation in an on-line conference discussion. The web log supply a raw trace of the learners' excursion and activities on the site. The learning system can be enhanced by using them in order to process these long entries and extract valuable patterns.

Many web log analysis tools are available. Most of these tools, such as Nihuo, AWStats, Analog, Webalizer, and Sawmill Analytics, etc., provide statistical analysis of web log data (Zaiane 2001).

In the next sections we describe log file type, log analysis, and we compare some log analyzer programs.

2. Log file type

The data about the learners visiting e-Learning portal e.g the learning components they use and how much time they spend with them, can be saved in log files of HTTP servers. HTTP servers which enable web files to be presented (IIS, Apache ect.) can save action of web site visitors as log files. Figure 1 shows W3c log file field and Figure 2 explain these fields.

Figure 1. W3c log file fields

- **Date (date)**: the date on which the request occurred.
- **Time (time)**: the time, in Coordinated Universal Time (UTC), at which the request occurred.
- **Client IP Address (c-ip)**: the IP address of the client that made the request.
- **User Name (cs-username)**: the name of the authenticated user who accessed your server. Anonymous users are indicated by a hyphen.
- **Service Name (s-sitename)**: the site instance number that fulfilled the request.
- **Server Name (s-computername)**: the name of the server on which the log file entry was generated.
- **Server IP Address (s-ip)**: the IP address of the server on which the log file entry was generated.
- **Server Port (s-port)**: the server port number that is configured for the service.
- **Method (cs-method)**: the requested action, for example, a GET method.
- **URI Stem (cs-uri-stem)**: the Universal Resource Identifier, or target, of the action.
- **URI Query (cs-uri-query)**: the query, if any, that the client was trying to perform. A Universal Resource Identifier (URI) query is necessary only for dynamic pages.
- **Protocol Status (sc-status)**: the HTTP or FTP status code.
- **Protocol Sub-status (sc-substatus)**: the HTTP or FTP substatus code.
- **Win32 Status (sc-win32-status)**: the Windows status code.
- **Bytes Sent (sc-bytes)**: the number of bytes that the server sent.
- **Bytes Received (cs-bytes)**: the number of bytes that the server received.
- **Time Taken (time-taken)**: the length of time that the action took in milliseconds.
- **Protocol Version (cs-version)**: the protocol version, HTTP or FTP, that the client used.
- **Host (cs-host)**: the host name, if any.
- **User Agent (cs(UserAgent))**: the browser type that the client used.
- **Cookie (cs(Cookie))**: the content of the cookie sent or received, if any.
- **Referer (cs(Referer))**: the site that the user last visited. This site provided a link to the current site.

Figure 2. Log file fields details

It is not necessary to keep all the fields indicated above in the web server. Which field to choose is depend on the user.

3. Log analysis

It will be helpful to use log analysis tool to aggregate the data and find meaningful patterns when web log data is collected. Several commercial log analysis tools are prevalent that enable you to analyze usage and feedback routinely. The issues that users have with both the product and the help system will be brought out with analysis of web logs. it can be identified the problematic areas for your students by Analyzing data. You can see the figure of your content and realize which part is satisfying for your students and which one is not by examining this data (Raiken 2005).

Web usage mining, is a process by which user browsing and access patterns can be discovered automatically from Web servers. Organizations with distance education sites collect large volume of data. These are the data which were automatically generated by Web servers and collected in server access logs. Referrer logs serve as sources for user information. They are the sources that contain information about the referring pages for each page reference, and user registration or survey data gathered via CGI scripts. With the help of analyzing such data organizations can determine, cross studying patterns across subjects, effectiveness of a web site and structure the thinking styles of learners. It can also provide information about restructuring a Web site to create a more effective Web site presence, and shed light on more effective management of collaborative study group communication and Web server infrastructure (Park et al., 2000).

Usage statistics of the learners on web system with the aim of learning is not enough to reach the information provided by web mining algorithm. However, they can be seen as a starting point to evaluate e-Learning system. As mentioned above, usage statistics can be kept by using web log analysis tools. By means of these tools you can record data. For instance; ‘during “T” period, for “P” page, “N” quantity clicking in existence”

4. Log analyzer programs

The lines in the log files may be seem as purposeless at first glance. However they can be made meaningful by Log Analyzer Programs. These programs can analyze each registered line. By doing so, they can prepare a report with visual richness by means of graphics. Webalizer Program in Figure 3 indicates the frequency of daily access to the e-Learning portal in August, 2010.

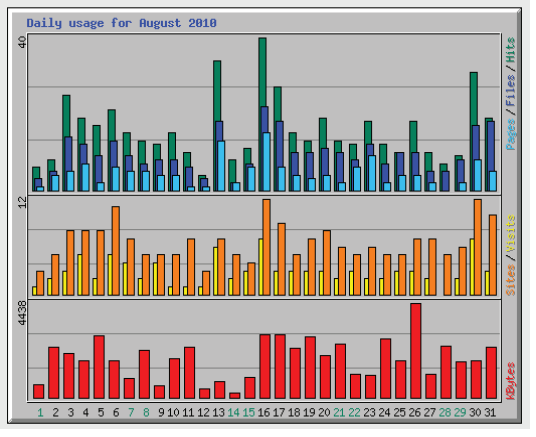

Figure 3. *2010 statistics of the visits for august — Webalizer*

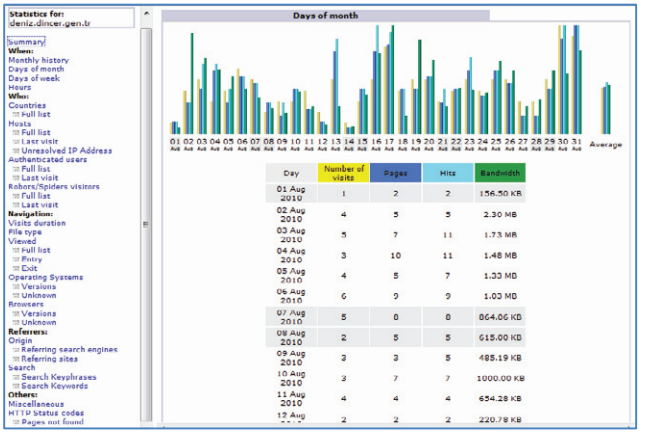

Figure 4. 2010 - statistics of the visits for august – Awstats

Features/Softwares	AWStats	Analog	Webalizer	Sawmill Analytics
Language	Perl	C	C	C/Salang
Price/Licence	Free/GPL	Free/GPL	Free/GPL	From $99 Per Profile Lite/Pro/Ent
Works with Apache common (CLF) log format	All features available with log format (b)	All features available with log format (b)	All features available with log format (b)	All features available with log format (b)
Works with IIS (W3C) log format	Yes	Yes	Need a patch	Yes
Works with personalized log format	Yes	Yes	No	Yes
Analyze Web/Ftp/Mail log files	Yes/Yes/Yes	Yes/No/No	Yes/No/No	Yes/Yes/Yes (+844)
Report number of "human" visits	Yes	No	Yes	Yes (Sessions)
Report session duration	Yes	No	No	Yes
Report countries	From IP location or domain name	Domain name	Domain name	From IP location or domain name
Report cities and major countries regions	Need Maxmind Cities database	No	No	Yes GeoLite City included
Report/Filter robots (nb detected)	Yes/Yes (642**)	Yes / Yes (8**)	No/No	Yes/Yes (250**)
Report most often viewed pages	Yes	Yes	Yes	Yes
Report entry pages	Yes	No	Yes	Yes
Report exit pages	Yes	No	Yes	Yes
Report OS (nb detected)	Yes (71)	Yes (29)	No (0)	Yes
Report browsers (nb detected)	Yes (208*)	Yes (9*)	Yes (4*)	Yes (~20*)
Report screen sizes	Yes	No	No	Yes & Depths
Report HTTP Errors	Yes	Yes	Yes	Yes
Report 404 Errors	Nb + List referer	Nb only	Nb only	Nb + List last date/referer
Analyzed data save format (to use with third tools)	Structured text file or XML	Text files with OUTPUT option	Flat text file	Flat text file/MySQL/MS SQL/Oracle
Graphical statistics in one page / several / or frames	Yes/Yes/Yes	Yes/No/No	Yes/Yes/No	Yes/Yes/Yes

Figure 5. Comparison Table of Log Analyzer Tools

In the same way, daily counts for the access to the portal is indicated by AWStats Program in the Figure 4. Comparison between AWStats and other famous statistics tools can be seen in the Figure 5 (analyzers Comparisons, 2010).

Due to GPL (General Public License) (GNU General Public License, 2010), programs such as Webalizer, Awstats and Analog Log Analyzer can be supplied free. Also Nihuo is a program that can be purchased commercially like Sawmill. It includes properties similar to other Log Analyzers (Nihuo Web Log Analyzer Features, 2010). It is used mainly Nihuo to analyze the e-Learning portal in our institution.

Nihuo can generate information report the data mentioned below as graphics. Figure 7 shows hourly hits in a day.

- How many visitors came to the web site.
- How visitors browse your web site.
- Where visitors come from e.g. from search engine or from other sites, from United States or Europe, accurate to cities.
- Which pages they viewed and which ones they ignored
- Whether a marketing campaign is successful, etc.
- How long visitors spent on your web site and each page.
- Advanced visitor filters and hits filters
- Automated scheduling
- Tracked files statistics
- Adjustment of the time zone
- Auto detecting log file format
- Stolen object statistics
- Downloading log files via FTP or HTTP
- Analyze log file data stored in an ODBC database
- Easy use of GUI
- Support Win 9x/Me/NT/2000/XP/2003/Vista
- Support Apache and IIS W3C web log format
- Support IPV6 Apache web log
- Support URL parameters analysis
- Support GZ, BZ, BZ2 and ZIP compressed log files
- Support multiple log files
- Support command line operation
- Support reverse DNS lookup

Figure 6. Nihuo Tools Properties

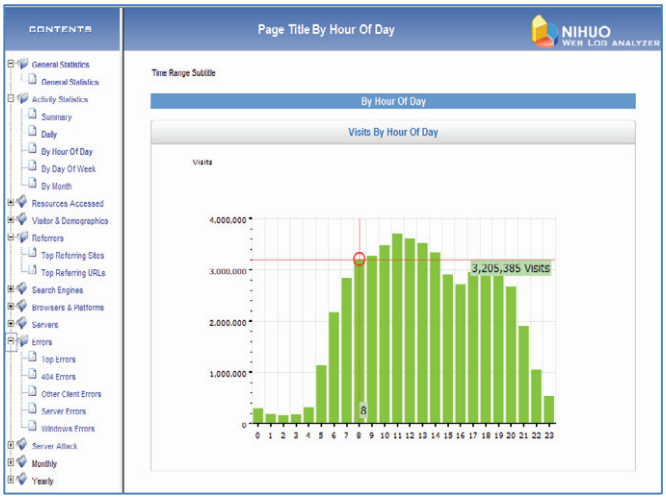

Figure 7. Range of visits considering the hours of the day – Nihuo

Web applications are developed very fast recently so the amount of logs or user access data saved in servers has increased. For instance, our e-Learning servers produce logs reaches to terabyte by a month. The reason is that we receive many hits in a month. Standard log analyze programs are slow to meet analysis requirements anymore. MapReduce (Dean & Gehawamat, 2004), working upon Hadoop (White, 2009) distributed parallel system, is another application we use to accelerate log analyzing procedure. Hadoop which was developed by apache is a java-based and open source software. MapReduce application on Hadoop frameworks mainly used to process a great deal of data.

Many companies such as Amazon, Google, IBM, Yahoo and Facebook use hadoop to do so. First Google used MapReduce programming model for web log analyzing and page ordering. Then Yahoo has started to Hadoop platform for open source as MapReduce programming model for cloud. Amazon uses the Hadoop platform with same purpose. Also Facebook analyze log data for social web service by using Hadoop platform. Via these analyzes facebook construct the millions of users' profiles (Lee, Kang & Son, 2010).

In this work it is analyzed user access files of our e-Learning portals used in distance education the files include information about student ID numbers, the lessons they log in and access time. At first, it is needed to build up a cluster structure, because Hadoop is a frame work which work with cluster structure. We set a basic framework consist of one master and 3 client in our application. In our studies we use computers with xeon core 2 duo CPU and 2GB ram. We installed Ubuntu 10.4 for operating system and Hadoop 0.20.2 version for Hadoop working environment. You need to install Java 6.0 for Hadoop to work. Computers with cluster structure connected to network environment with 100mbit/bps ethernet interface.

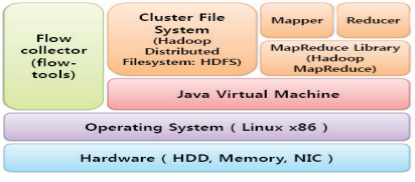

Figure 8. Functional components of a cluster

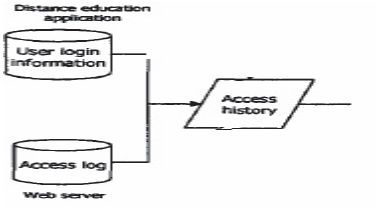

Figure 9. Log files

Course	Hit
"000103"	88634
"000104"	7780
"000105"	6610
"010203"	81538
"010204"	7693
"010205"	6188
"020303"	121162
"020304"	11591
"020305"	9084
"030403"	161490
"030404"	14742
"030405"	12094
"040503"	204231
"040504"	19951
"040505"	15650
"050603"	223446
"050604"	23076
"050605"	18022
"060703"	281921
"060704"	31301
"060705"	24680
"070801"	8884
"070802"	4450
"070803"	362713
"070804"	34852
"070805"	28129
"080901"	26652
"080902"	13997
"080903"	704854
"080904"	97649
"080905"	129907
"091001"	82578

Figure 10. Output file

The process in the MapReduce is as follows: data in the log files of master computers were allocated to the cluster computers and so data are manipulated in parallel with each other. Each cluster computer rotates results to master computer. Master computer gathers these results and create an analyze output file.

As shown in Figure 9 access history information is comprised access log file and user login information file. As shown in below sample application on MapReduce, it is used nearly 800 MB user login information file. In this application hits were calculated for every course. Thus, students' interest in lessons can be seen. Output file is shown in Figure 10. MapReduce with the application of various analysis done on a large log files. Hadoop cluster analysis of the structure of the processing time can be shortened by increasing the number of machine.

5. Conclusion

Log analysis programs give statistical information about the users. For instance, information about the pages in which visitors log in and out, and information about the duration of use of those pages were obtained. This has implications for education. When the visit statistics of the e-learning pages that include various components are obtained, the components preferred by the learners and what learning style is preferred by the learners can be identified.

Log analyzer programs thanks to IP numbers stored in the log file, it is possible to learn in which city the visitor connect to portal. Investigating the information leads the relation between geographical position and learning style to be assigned. Log Analyzers can analyze the spiders' data and search engines, which leave a trace like a visitor. The key word which connected search engine to the page can be found. Then most frequently used key words can be seen in the reports of the Log analyzers.

Technical properties of the computer used to access the page (such as operating system, Internet browser, screen resolution etc.) can be obtained. This information is needed to be considered by instructional designers and software developer while preparing an e-Learning portal.

Errors turn back from HTTP servers can be reported by log analyzer. By this means, the pages to which visitors fail to connect can be detected and the problem will be solved.

Log analyzer programs generate some statistical information but if we need more detailed statistical information we should use data mining or web mining algorithms and programs.

References

Analyzers Comparisons. (2010, 05 10). *General format.* Retrieved from sourceforge: http://awstats.sourceforge.net/docs/awstats_compare.html.

Dean, J., Ghemawat, S. (2004). MapReduce: Simplified Data Processing on Large Clusters. *In Proc. of the 6th Symposium on Operating Systems Design and Implementation,* San Francisco CA.

GNU General Public License. (2010, 09 10). *General format.* Retrieved from GNU: http://www.gnu.org/licenses/gpl.htm

Jansen, B.., Spink, A. and Taksa, I. (Eds.). (2008). *Handbook of Research on Web Log Analysis.* pp. 506-522., USA, Pensylvania, Hershey: Idea Group Inc (IGI Global).

Lee, Y., Kang W. and Son H. (2010). An Internet Traffic Analysis Method with MapReduce. 1st IFIP/IEEE Workshop on Cloud Management , Osaka.

Mutlu, M.E., Kip, B. ve Kayabaş, İ., "e-Sertifika Programlarında Katılımcıların Öğrenme Ortamı Tercihleri", *Future-Learning 2. Uluslararası Gelecek İçin Öğrenme Alanında Yenilikler Konferansı 2008: e-Öğrenme,* İstanbul, 27-29 Mart 2008.

Nihuo Web Log Analyzer Features. (2010, 01 10). *General format.* Retrieved from Nihuo: http://www.nihuo.com/web-log-analyzer-features.html

Park, C. S., Bae, S. M., & Ha, H. S. (2000). Web Mining for distance education. Korea Advanced Institute of Science and Technology, 716.

Raiken, N. (2005). Analyzing web-based help usage data to improve products, *Professional Development/STC-Related Sessions.*

White, T. (2009) . *Hadoop: The Definitive Guide.* O'Reilly Media, Yahoo! Press.

Zaiane, O.R., and Luo, J. (2001). Towards Evaluating Learners' Behaviour in a Web-Based Distance Learning Environment. *Proc. International Conference on Advanced Learning Technologies ICALT'01.*

Zaiane, O.R. (2001). Web Usage Mining for a Better Web-Based Learning Environment. *Proceedings of Conference on Advanced Technology for Education (CATE'01).* Banff, Alberta.