A Design and Implementation of a Web Server Log File Analyzer

∗Yu-Hsin Cheng[1], Chien-Hung Huang[2]
[1]Department of Information Management,
Ling Tung University
No. 1, Ling tung Rd., Taichung, Taiwan
[2]Department of Computer Science and Information Engineering,
National Formosa University,
No.64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan

Abstract: -More and more enterprises use network to communicate with the suppliers or the customers, and also in receiving order forms, data transmission, produces goods, stock warehousing in all enterprise management procedure, carry on to the e-movement and enterprise competitive ability, this act not only may save the production cost, moreover may fast reflect the correlation information. But the enterprise owners or the website administrators often cannot judge precisely whether their commercial websites play the leading role in marketing function. In the settlement of this problem, an intelligent website analysis software can often provide good and effective assistance to business owners or website administrators, and then improve the marketing results of the websites. The website analysis software usually uses the log files in the system to produce statistical-related data, and then obtains the useful information through the analysis. In this paper, we discuss the implementation of an enterprise website analysis software we have developed, and by the modification and the expansion the Analog from free software to analyze the log files in the Apache web server, provides more important information about visitors. We provide a user friendly interface in the system for the enterprise administrators to accumulate the important information of log files, in establishing the system, we add on functions which detect illegal information, make the protective action, and provide the dynamic E-Mail report function to the user.

Key-Words: - log file, web server, website analysis, Configuration Interface Subsystem, Illegal Information Detection Subsystem, Report Delivery Subsystem

1 Introduction

At present most of the commercial websites use Microsoft software, which is so called the Windows operation system, IIS and ASP, and the database of MS SQL, the message system uses Exchange server. Microsoft's software not only is extremely expensive, also requires frequently the version upgrading, each time the upgrading costs a quite considerable expense. On the one hand, because the marketing idea of software design of Microsoft software is covering the range from the enterprise to small and medium business, therefore the entire software product line has become a huge, complex, hard to use monster. At last, the whole system carries out the potency of failure, in order to deploy smoothly, the hardware specification requirement for Microsoft software is usually high, especially requires frequently compromise. Briefly speaking, Microsoft software plan is expensive and complex. This costs a matter regarding small and medium business.

On the contrary, the design idea of the Linux system is to make the source code public and free to use. Because of the opening source code, the computer expert in the world may make comprehensive accuracy and the potency on the system program with thoroughly examines, and unceasingly expands function of the system. By all means, strength of the Linux people can provide the quite outstanding system potency with hardware which no need to be high-end level. More importantly, nearly all software of the Linux system is free, when vendor charges the fee, mostly is because of providing the collection, the reorganization or the other software service expense. For small and medium business, this may save some big money. Therefore, this research takes Linux as the developing platform.

Although quite a lot of enterprises have already implemented website in order to pursue e-commerce and improve the competitiveness, but the enterprise owners or the website administrators often can not judge precisely whether their commercial websites play the leading role in marketing function. In the settlement of this problem, an intelligent website analysis software can often provide good and effective assistance to business owners or website administrators, and then improve the marketing

results of the websites. [1-10]. For example we may tell by looking at overhead architecture of the website by analysis customer's visits of the commerce website each time whether the website designs appropriate or not, having explanation of which homepages with bad design to cause the visitor giving up to continue to browse, which important keywords may guide the visitor to its website, most can attract in which entrances website advertisement visitor and so on. All these information can assist the website administrator to adjust its website's architecture and the content to a better design, and then improve the marketing status of the website.

A website analysis software usually applies the historical log of the system to statistic of all the related data, and derives useful information. Currently, the commercial institute [1-5] and Open Source Community [6-10] both develop this kind of product. In the matter of small and medium business, buying business software will absolutely increase the cost of that company. That concludes the significant of providing free log analysis software. In this research, we discuss the implementation of an enterprise website analysis software we have developed, and by the modification and the expansion the Analog from free software to analyze the log files in the Apache web server, provides more important information about visitors.

2 Literature review

The log application has been use widely in the market [11-17], there are all kinds of logs in web servers, recorded all the events of website. Every record in the event is the web page been browsed, the size of web page, what browser has been used, the duration of browsing, and so on. But the log is hard to read and understand; also the number of events is huge. Therefore the system administrator only can consult some dispersible materials, and still unable to make the massive materials the statistics and the analysis. Therefore, the website analyzes the software ability to transform these complex materials information to a easy reading, simultaneously also provided the extremely user friendly interface, then the system administrator will be very easy to obtain the statistics or the numeral. For example the majority of analyses result can demonstrate the graphics, some even can provide the PDF files output and so on. In the free software community, has provided some website analysis software, including PhpOpenTracker [6], Lire [7], AWStats [8], Webalizer [9] and Analog [10] etc. These software have the different supporter, also have the different function, we will introduce as following.

PhpOpenTracker [6] is written in PHP, it put all the requests into mysql. On the other hand, it also provides many API for the software developer, who can call a simple function to execute complicated job. But phpOpenTracker is written in PHP, in order to have a better performance, we recommend not putting the analysis software and target server together.

Lire[7] is a pluggable log analysis software. It is written in Perl and Shell script, integrate with HTTP, email, DNS, FTP and Firewall. Lire converts original log to specific DLF format, and then proceed with its own analysis engine to execute jobs.

AWStats [8] is written in Perl, it analysis website, ftp, and email server log. It also provides the total amount of users and robots that have visited the web site. The so-called robot actually is some automatic searching software. It helps administrator to understand the real commercial behavior between these two kinds of visitors.

Webalizer [9] is written in C, and is also a log file analyzer software. , it can create the statistic result through configuration file, and send the result in HTML format to WebPages for browsing. But Webalizer is not able to process administrator's self-defined log file format, and has not upgraded or patched the software since April 2002.

Analog [10] is the log files analysis software written in C. Analog mainly analyzes the log in the website, currently may process the administrator self-defined pattern from the log, and the load balanced system interaction log, which analysis including in the browsing peak time, the daily capacity in a week, filters which page the robots browse over the most, what browser and operation system the visitor uses and so on. Because in general commercial website, the customer commonly accesses then the website server provides the service, therefore this research first focus on Linux system log files in Apache Web server. Moreover, because Analog is written in C language, therefore it has better system performance compare to Perl and the PHP programming language. This characteristic has been regarding as a great advantage on some website establishment on the non high-end server website of the high browsing rate. Moreover, Analog always has the new version to come, and is stable and better in functions compares to Webalizer. Therefore, in this research we adopt Analog as the analysis software of basic framework.

The Apache server provides the following important logs：

(1) Error log：records all the error information, including CGI script error messages.

(2) Access log：records all the requests for the server, and by the analysis of the log, we will obtain many precious information. For example ：how many customers have recently browse homepages of the website in past week.

(3) Script Log：Records CGI script of input and output data.

(4) Rewrite Log：Records the detailed analysis to explain how the Rewrite engine does transform the request.

Analog allows users to configure different report format of statistical content, anyhow the configuration is in text format, would easily have several hundred kinds of different configuration. In addition grammar of the configuration is quite complicated, therefore it is hard to maintain the configuration. Following is small fragment in a Analog configuration files：

DAILYSUM "OFF" # we don't want a Daily Summary
DAILYREP "ON" # we want a full Daily Report instead
HOSTNAME (Spam Widgets Inc.) # Spaces, so quotes or brackets needed
LOGFILE logfile1.log, \
Logfile2.log # this line and the previous one are one command

Since the configuration is quite complicated to the users, therefore this study shows how we will provide users an interface with easy configuration. First, we will study the Analog configuration files to provide the significant option and the hypothesis grammar. Then the plan designs a user-friendly interface, also provides a homepage picture, and provides the user not only having canceled chooses and combines some different option, comes automatically to complete setting work of configuration.

3 System design

3.1 System architecture

This research establish the log files analysis system (Log File Analyzer, LFA), the system will provide the users an easier, more perfect Web log files analysis software. The system is divided into 3 subsystems as shown in Figure 1 according to its requirements： Configuration Interface Subsystem (CIS), Illegal Information Detection Subsystem (IIDS), and Report Delivery Subsystem (RDS). We introduces their responsible separately as follows:

(1) CIS Subsystem

- Chooses Log file which Web Server produced to create the statistical data analysis.
- Provides self-defined configuration function to establish the log information which the log files analysis software can produce.
- According to different user's demand in order to capture the specific report information.

(2) IIDS Subsystem

- Detects the intrusion and the attacking according to the illegal request of Log File.

(3) RDS Subsystem

- Provides dynamic E-Mail function to the user.

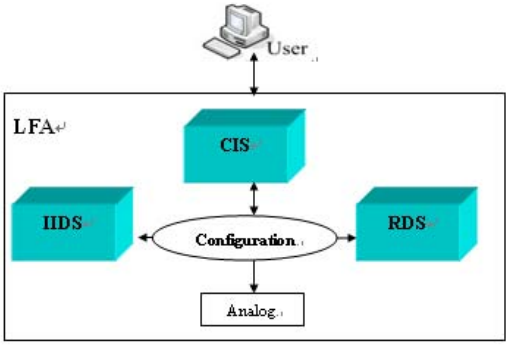

Fig. 1: LFA Architecture

3.2 Feasible solution assessment

This study analyzes the log files produced by the Apache Web server installed on a Linux-based system, and uses the free software log analyzer with open source as the basic architecture to develop additional functions.

- Performance: If the system runs very slow and the processing time is too long in the execution.
- Stability: If the system works stably and hangs.
- Extensibility: If the system functions are extendible in the future.
- Ease of development: If there is room for future development of the system.

Common software log file analyzers include the PhpOpenTracker, Lire, AWStats, Webalizer and Analog as described above. This study uses the Analog analyzer as the basic architecture to develop fuller additional functions by means of extension and with reference to other log analyzers in design.

The Analog is selected as the log analyzer in this study because it is a C-based log analyzer and has better performance and stability when running on Linux-based systems. Second, it uses open source and is a freeware. It has been widely used and is ready for shared debugging. Therefore, version updates are faster, functions are more powerful, and bugs are fewer.

3.3 Requirements and design of system interface

3.3.1 System interface requirements

The system must be able to send data via HTTP protocols and users (browsers); and subsystems must be able to complete function parameter setup and run these functions by accessing a common configuration file.

3.3.2 System interface design

(1) CIS subsystem needs to penetrate the PHP program in order to complete the access and the modification of configuration files parameter.
(2) IIDS subsystem needs to grasp the parameter of the configuration file user defined through C shell Script, then generate the result after compare to signature database.
(3) RDS subsystem requires the user defined parameter of the configuration file, and then configure the scheduling program to send the report periodically.
(4) Finally, LFA system WebPage allow users to access through HTTP protocol.

3.4 Three Subsystems

3.4.1 CIS subsystem

The CIS subsystem provides users a simple operational GUI (scroll bar, the graphic, Option Box, Checked Box... HTML table component), and users can configure the scheduling analysis software by using Internet Browser equipped computer, this allow users to produce all kinds of different combination analysis report. After analyzed the above subsystem function, there are 2 components in this subsystem: "HTML GUI" and "Interactive Webpage Routine", like Figure 2 shows, following describes these two processing components separately.

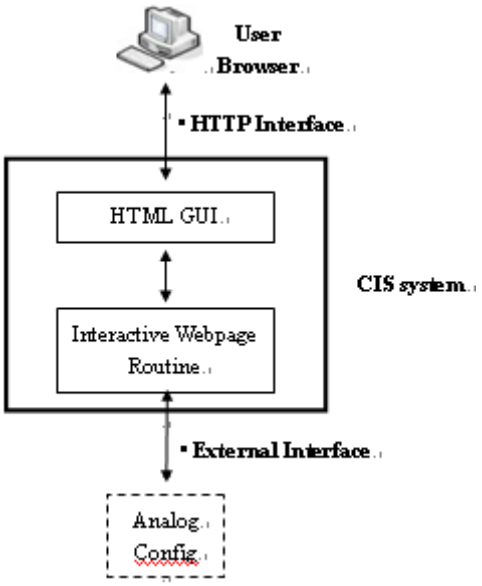

Fig. 2: CIS subsystem architecture

- HTML GUI

When the user uses browser and connects to the CIS subsystem, penetrates this part to show the configuration files he to examine, and is modified by user thought the same HTML WebPage for making the modification, or remove the function.

- Interactive WebPage Routine

There are two main processing functions: 1. Reads from the Analog configuration files and takes to demonstrate the modification of the parameters; 2. Receives parameters setting from GUI HTML transmission, and then penetrates this part to the modification of the Analog configuration files.

3.4.2 IIDS subsystem

The IIDS subsystem analyzes the request URL in Log file that dwell in Apache Web Server, and detects whether there is any illegal intrusion attack, after reorganizes the useful information to generate the report for the user's quick understanding, and lead the user to make the following protective action. After analyzed the above requirement of subsystem, there are three processing parts in this subsystem: "Intrusion Detection Routine", "Signature Database" and "Script Filtering", like Figure 3 shows, following describes three processing parts separately.

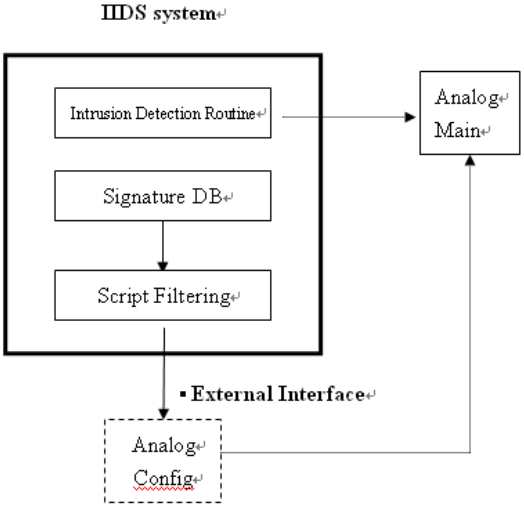

Fig. 3: IIDS subsystem architecture

- Signature Database

Because the URL attack may divide into many different types, possibly the virus attack, the network Hacker attack, abnormal access...and so on, therefore we need to use a signature database in this subsystem to verify many abnormal URL requested in the network, this part is the special character string or the random code contains in abnormal URL request.

- Script Filtering

Because the signature database was not setup for this system to design, therefore many useless information in the system may be useless, therefore needs to use this part to do filtering.

- Intrusion Detection Routine

This component is a must in the Analog main program, mainly is the logs of the normal URL request in the configuration files of Analog, when all URL in Log file does not related to the comparing action, then will detect possible abnormal URL request and generate another statement analysis.

3.4.3 RDS subsystem

After users have configured the RDS subsystem, the system would send out the report automatically to the users, it uses the e-Mail for transmission, and help the rapid adjustment managing the website.

When the analysis of the subsystem has finished, there are 3 process component been executed, "Scheduling Program Execution", "E-Mail Shell Script" and "Mail Server", like figure 4 shows, following we describes three components separately.

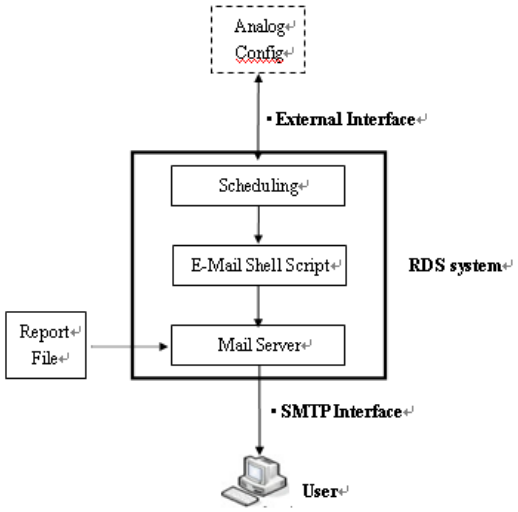

Fig. 4: RDS subsystem architecture

- Scheduling Program Execution

The execution time establishes which CIS subsystem goes to the automatic execution Shell Script to mail the report information.

- E-Mail Shell Script

Capture the report information through this component, and execute the E-Mail process according to the schedule.

- Mail Server

This component is the actual transmission report doer to a designated receiver Mail Address.

4 Conclusions

This study establishes the log files analysis software including the configuration interface subsystem, the illegal information detection subsystem and the report delivery subsystem. There is no graphical interface for configuration in Analog needs the manual editing, and then the side can demonstrate the report form. The configuration interface subsystem provides easy-to-use graphical interface (scroll bar, graphics, Option Box, Checked Box...HTML table component). The users need to have Internet Browser equipped computer to setup the configuration files of analysis software, this enable users to produce all kinds of different combination analysis report form. At present, when Analog derives fail, can only distinguish the intrusion manually. The illegal information detection subsystem is able to analyze inside the Log file which Apache Web Server produces from all URL request, and detects whether there is intrusion attack, after reorganizing, the useful information to generate the report for the users to quick understand, can lead the user to make the later protective action.

In the report delivery subsystem, the user will need to input name, E-mail, time through web interface, the system will mail the reports to users according to what users have configured, which will allow to monitor server as necessary. The system will record the requirement to Database. The principle of work is lying on Crontab, which will carefully examine the database on each period of time, and compared the demand of time with various users, and send the report to those whose time is consistent.

References:

[1] ClickTrack:http://www.clicktracks.com
[2] HitBox:http://www.hitbox.com/
[3] Sawmill:http://www.sawmill.net
[4] Summar:http://summary.net/index.html
[5] WebTrends:http://www.netiq.com
[6] phpOpenTracker: http://www.phpopentracker.de/
[7] Lire：http://logreport.org
[8] AWStats：http://awstats.sourceforge.net/
[9] Webalizer ： http://www.webalizer.org/newindex
[10] Analog：http://www.analog.cx
[11] J. H. Andrews, Testing using log file analysis: tools, methods, and issues, *Proceedings of 13th IEEE International Conference on Automated Software Engineering*, 1998, pp. 13-16.
[12] J. H. Andrews, and Y. Zhang, Broad-spectrum studies of log file analysis, *Proceedings of the 2000 International Conference on Software Engineering*, 2000, pp. 105-114.
[13] A. Arona, D. Bruschi, and E. Rosti, Adding availability to log services of untrusted machines, *Proceedings of 15th Annual Computer Security Applications Conference*, 1999, pp. 199-206.
[14] T. Feng, and K. Murtagh, Towards knowledge discovery from WWW log data, *Proceedings of International Conference on Information Technology: Coding and Computing*, 2000, pp. 302-307.
[15] H. Lai, and T. C. Yang, A system architecture of intelligent-guided browsing on the Web, *in the Proceedings of the Thirty-First Hawaii International Conference on System Sciences*, 1998, pp. 6-9.
[16] A.I. Reuther, and D.G. Meyer, Analysis of daily student usage of an educational multimedia system, *in the Proceedings of 27th Annual Conference. Teaching and Learning in an Era of Change, Frontiers in Education Conference*, vol.3, 1997, pp. 412-1417.
[17] Y. K. Woon, W.K. Ng and E. P. Lim, Online and incremental mining of separately-grouped Web access logs, *in the Proceedings of the Third International Conference on Web Information Systems Engineering*, 2002, pp. 12-14.
[18] G. Castellano, A. M. Fanelli, and M. A. Torsello, LODAP: A LOg DAta Preprocessor for mining Web browsing patterns, *Proceedings of the 6th WSEAS International Conference on ARTIFICIAL INTELLIGENCE, KNOWLEDGE ENGINEERING and DATA BASES (AIKED '07)*, 2007, pp. 12-17.
[19] G. Castellano, A. M. Fanelli, and M. A. Torsello, Mining usage profiles from access data using fuzzy clustering, *Proceedings of the 6th WSEAS International Conference on SIMULATION, MODELLING AND OPTIMIZATION (SMO '06)*, 2006, pp. 157-160.