ISSN:0975-9646

Cluster Diagnostics and Verification Tool for Effective and Scalable Web Log Analysis

Shakti Kundu
School of Computer Applications,
Lovely Professional University,
Jalandhar-Delhi GT road, NH-1, Phagwara, Punjab, India.

***Abstract* -- The aim of this paper is to analyze web transaction log data that reveal user behavior in the internet world. To achieve this, clustering tool is used for performing two phases. First, an information gathering is done on web transaction data. Second, an analysis is done on web access log files to analyze the user transaction behavior and user's log-in approach. In this paper, the cluster diagnostics and verification tool is used to discover hidden relationships among the Web server data and access patterns.**

Keywords: Cluster, Log file, Tool, Web.

I. INTRODUCTION

Web Data Mining refers to the extraction of hidden predictive information from large amounts of web database. Data mining involves various web log analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Web data mining not only refer for collecting and managing web data but also followed for analysis and prediction.

In the internet world, Web mining is a very effective data mining approach. Web data mining is applying for traffic analysis of websites based on web log files which basically helps in increasing the efficiency of the websites.

There are various formats of web log files which are having rich source of information about the activities of visitors. The issue here is size of the web log file in the e-commerce is quite large which is in GBs or TBs. This huge volume of information has to be analyzed with clustering tool which will helps us in understanding the overall study of web log analysis.

K-means algorithm was used to cluster the data and the retrieved content (data features) including users' navigation paths, page viewing time, hyperlink structure, and page content. Then they were used for building up a vector space model in order to find the Web users' usage patterns represented as online user profiles [1-2].

To discover the access patterns and sequential navigational patterns, association-rule mining and clustering had been used in many research projects [3-4]. For general access pattern tracking, some of the popular projects are 'Weblog Mining' [5], 'WUM' [6] and 'WebSIFT' [7]. Customized usage tracking research includes 'Adaptive Web sites' [8] and WebWatcher [9] and so on.

The clustered diagnostics and verification tool [10] were further used for learning the trend patterns by using statistical analysis methods. In order to make the analysis more intelligent we also used the clustered data to predict the daily and hourly traffic including request volume and page volume.

In the subsequent sections, the remainder of the paper is organized as follows: Section II. describes an overview of Web server Log File, Section III. refers to the discussion about Cluster Diagnostics and Verification Tool and Section IV. concludes the paper.

II. WEB SERVER LOG FILE

A web server log is a log file which is automatically created and maintained by a server of activity performed by it. For example, a web server log which maintains a history of page requests. The WWW maintains a standard format (the Common Log Format) for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information.

These files are usually not accessible to general Internet users, only to the webmaster or other administrative person. A statistical analysis of the server log may be used to examine traffic patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, adequate hosting resources and the fine tuning of sales efforts can be aided by analysis of the web server logs. Marketing departments of any organization that owns a website should be trained to understand these powerful tools.

Table 1. An Example of Server Log Fields

Originating IP	123.237.23.247
Timestamp	[22/Jan/2012:04:22:02 +051800]
HTTP Command & Protocol Version	"GET /distance/general/scentre.htm HTTP/1.1"
Status Code	200
Browser	"Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.75 Safari/535.7"
Referring URL	http://www.google.com

Web browsers and servers communicate using the stateless hypertext transfer protocol (HTTP). The header of an HTTP request message contains the attribute value pairs that a web server can record in its log file. Therefore, the web log file contains fields that describe each request that a browser makes from a server. By combining what can be inferred from this data, coupled with our understanding of how this information was derived, we can accurately analyze user activity [11]. For example,

123.237.23.247 - - [22/Jan/2012:04:22:02 +051800] "GET /distance/general/scentre.htm HTTP/1.1" 200 26394 "http://www.gjust.ac.in/distance/ddatesheet.htm" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.75 Safari/535.7"

First we have an IP Address (123.237.23.247) of the client which has queried this web server at a given standard date and time [22/Jan/2012:04:22:02 +051800] for the index page with request method (GET/). The proxy server fulfilled this request and returned with status code 200 which means request resulted in successful response. There is an associated search string as (http://www.gjust.ac.in /distance/ddatesheet. html) with User Agent (Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.7)

III. THE CLUSTER DIAGNOSTICS AND VERIFICATION TOOL

The Cluster Diagnostics and Verification Tool (ClusDiag.exe) is a tool that performs basic verification and configuration analysis checks and creates log files to help system administrators identify configuration issues prior to deployment in a production environment. ClusDiag.exe can capture all relevant log files and event logs from each node of a server cluster and merge them into a single file for easy analysis and troubleshooting. Administrators can analyze these log files with built-in filtering, merging, and bookmarking functionality and generate various diagnostics reports. ClusDiag.exe can also create graphical and text-based reports of cluster disk and network configurations, as well as generate a graphical view of the cluster resource dependency tree.

A. Where to Download the Tool

You can obtain ClusDiag.exe from the Microsoft Windows Server 2003 Resource Kit, or you can download it from the Cluster Diagnostics and Verification Tool (ClusDiag.exe) Web site [10].

B. Opening the Cluster Log

The following figure shows the dialog box you see when you first open the Cluster Diagnostic and Verification Tool

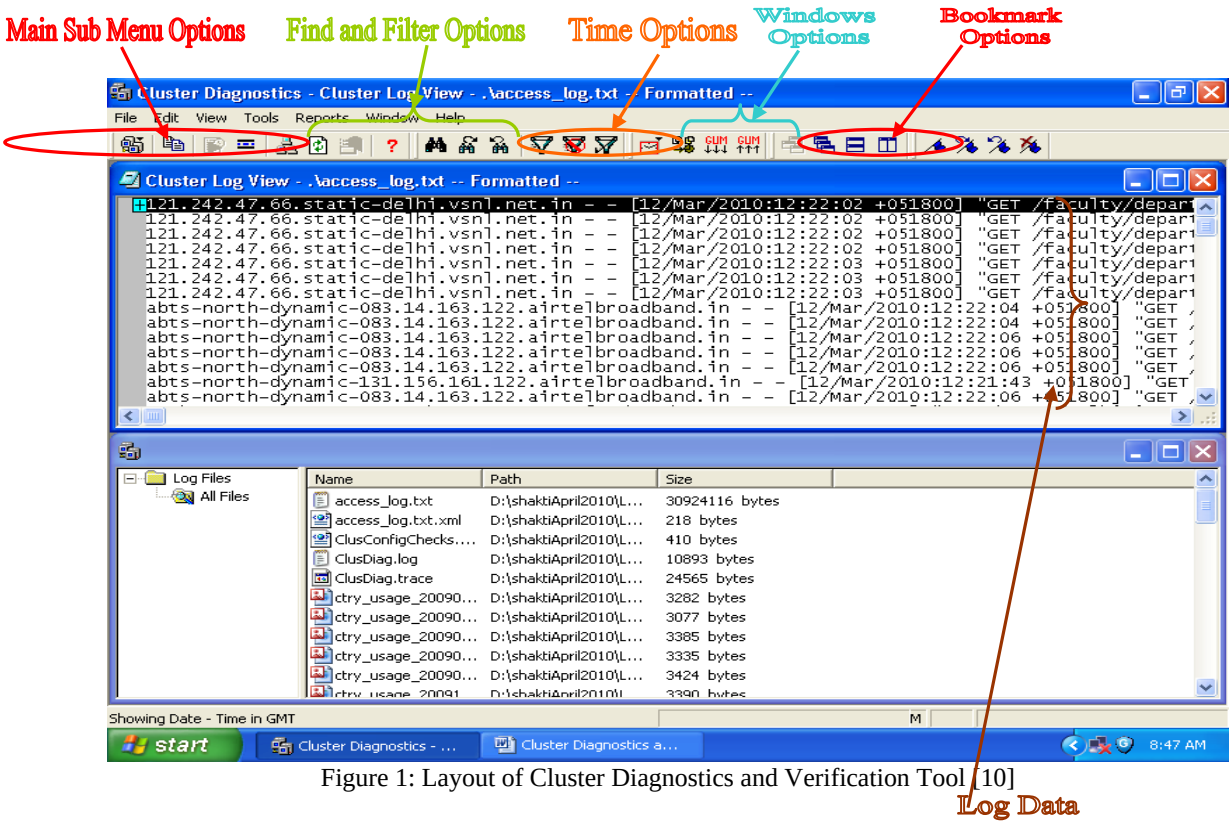

Figure 1: Layout of Cluster Diagnostics and Verification Tool [10]

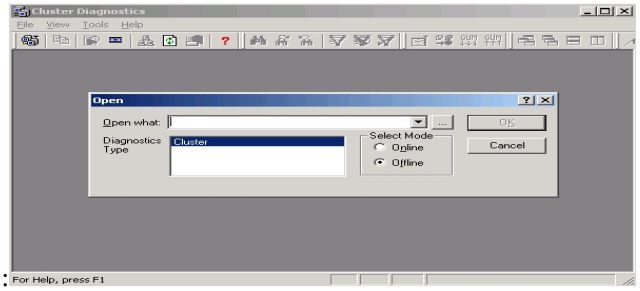

Figure 2: Opening the ClusDiag Tool

You can use this dialog box to go to the folder that contains the cluster log file.

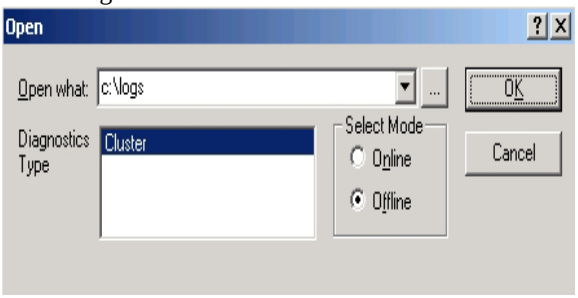

Figure 3: Open Dialog Box

C. Filtered View of the Cluster Log

The following figure shows a filtered view of the cluster log that you see when you first open the cluster log file:

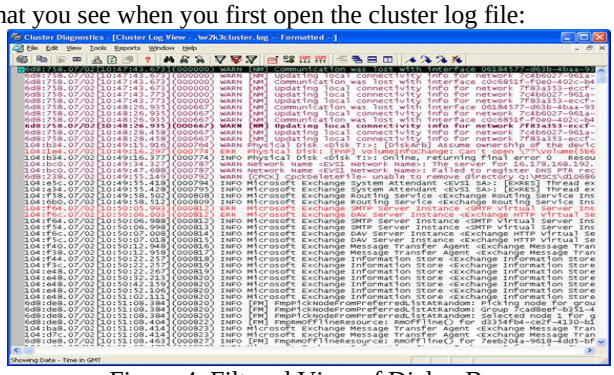

Figure 4: Filtered View of Dialog Box

You can see the default filter by clicking View, clicking Filter, and then clicking Show Filter, or by pressing F4. The following figure shows the default filter:

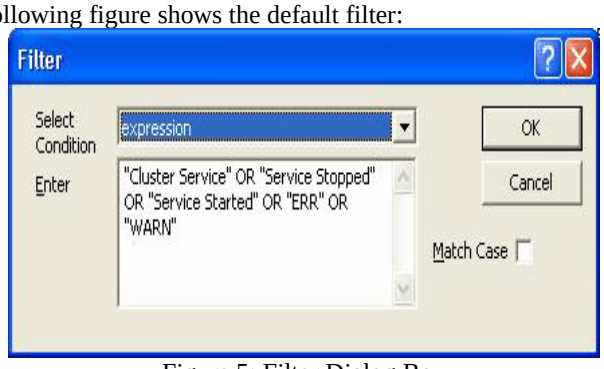

Figure 5: Filter Dialog Box

The default filter is set up to indicate a warning or something of interest for troubleshooting purposes is believed to have happened.

The final thing to notice is that some items are color coded. You can see and change the color coding by clicking Tools, clicking Options, and then clicking the Color Codes tab.

D. Adding a Bookmark

Cluster logs can be quite large, and one of the features that are very useful in ClusDiag.exe is the ability to place a bookmark in to the cluster log file. You create a bookmark by clicking the far left side of a line in the log file as shown in the following figure:

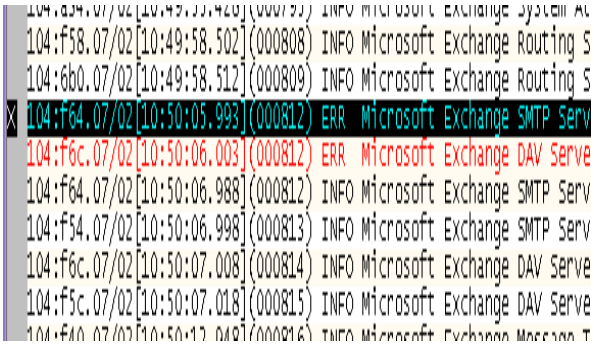

Figure 6: Adding a Bookmark

Bookmarks are not saved when you exit the Cluster Diagnostic and Verification Tool.

E. Adding a Comment

You can add a comment to any line in the cluster log. You add a comment by right clicking a line, clicking Comment, and then clicking on Edit. A line with a comment attached appears by default with yellow highlighting; if you place your mouse cursor over the line, you see the comment displayed as a tool tip as shown in the figure 7.

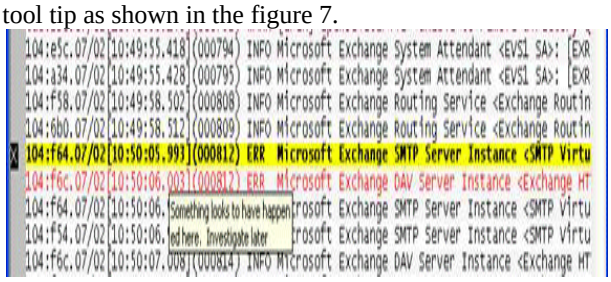

Figure 7: Cluster Log View

Comments are saved in an XML file that is reread when you open the cluster log. The file containing the xml is in the same folder as the original log file. For example:

```
<XML>
<LINECOUNT>8575</LINECOUNT>
<COMMENT>
<LINE>1933</LINE>
<STRING> …….. </STRING>
</COMMENT>
</XML>
```

F. Moving Around In ClusDiag

Beyond the regular methods of navigating in ClusDiag.exe, a quick way to jump around the cluster log is to jump between bookmarks. You can use the flag icons on the toolbar to create, move between, and clear bookmarks. The first flag adds a new bookmark, the second flag moves the position in the log file to the next bookmark, the third flag icon moves you to the previous bookmark, and the last flag clears the current bookmark.

Table 2: Different Types of Files needed to run a test by ClusDiag

Files Needed to Run a Test and Full Capture in Online Mode		
Exe files	DLLs	Configuration files
Wddump.exe Spsrv.exe Spsrvcl.exe Spfail.exe	Clusdiagdll.dll Wttsdk.dll Ntlog.dll Wttlog.dll Msxml3.dll (Windows 2000 only) Msxml3r.dll (Windows 2000 only) Msvcp60.dll (Windows 2000 only)	Windiag.ini

Running a test in online mode requires that the following files be installed on the target nodes. These files are automatically installed when the Run Test or Capture Logs (Full capture) operations are performed in online mode. The files are automatically uninstalled when the operations are complete

Table 3: Log File types supported by ClusDiag

Log File types supported by ClusDiag	
.log	A file that lists actions that have occurred. For example, Web servers maintain log files listing every request made to the server.
.txt	A text file is a kind of computer file that is structured as a sequence of lines. A text file exists within a computer file system. The end of a text file is often denoted by placing one or more special characters, known as an end-of-file marker, after the last line in a text file.
.xml	XML (Extensible Markup Language) is a set of rules for encoding documents electronically. It is defined in the XML 1.0 Specification produced by the W3C, and several other related specifications.
.html or .htm	HTML, which stands for Hypertext Markup Language, is the predominant markup language for web pages. It provides a means to create structured documents by denoting structural semantics for text such as headings, paragraphs, lists etc as well as for links, quotes, and other items. It allows images and objects to be embedded and can be used to create interactive forms. It is written in the form of HTML elements consisting of "tags" surrounded by angle brackets within the web page content.
.evt	Event log files

IV. CONCLUSIONS

The relationships between the behaviors of user were studied by analysis of web access log data from Guru Jambheshwar University of Science and Technology (G.J.U.S&T), Hisar website [12]. In doing this, a Cluster Diagnostics and Verification Tool was employed. In the analysis of the full data set, we can gain a rough idea of non member's reach ratio for this site.

The Cluster Diagnostics and Verification tool capture log files as well as a analyze log files which found to be quite useful for the study of web log analysis.

REFERENCES

[1] Martı´n-Bautista MJ, Kraft DH, Vila MA, Chen J, Cruz J. User profiles and fuzzy logic in web retrieval. Springer JSoft Comput 2002;65(5):365–72.

[2] Pazzani MJ, Billsus D. Learning and revising user profiles: the identification of interesting web sites. J Machine Learning 1997;27(3):313–31.

[3] Berkan RC, Trubatch SL. Fuzzy logic and hybrid approaches to web intelligence gathering and information management. In Proceedings of 2002 World Congress on Computational Intelligence, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE(02) Special Session on Computational Web Intelligence (CWI), Honolulu, Hawaii, USA 2002;2002:1033–8.

[4] Chen PM, Kuo FC. An information retrieval system based on an user profile. J Syst Software 2000; 54:3–8. Zaı¨ane OR. Building virtual web views. J Data Knowledge Engng 2001;39(2):143–63.

[5] Zaı¨ane OR, Xin M, Han J. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Proceedings of Advances in Digital Libraries Conference, Santa, Barbara, California, USA 1998;1998:19–29.

[6] Spiliopoulou M, Faulstich LC. WUM: a web utilization miner. In Proceedings of Workshop on the Web and DataBases (WebDB(98), Valencia, Spain 1998;1998:109–15.

[7] Cooley R, Tan PN, Srivastava J. WebSIFT: the web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling (WebKDD'99) Workshop on Web Mining, an Diego, CA, USA; 1999. pp. 163–82.

[8] Perkowitz M, Etzioni O. Adaptive web sites: automatically synthesizing web pages. In Proceedings of the 15th National Conference on Artificial Intelligence and 20th Innovative Applications of Artificial Intelligence Conference (AAAI(98, IAAI(98), Madison, Wisconsin, USA 1998;1998:727–32.

[9] Joachims T, Freitag D, Mitchell T. WebWatcher: a tour guide for the world wide web. In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI(97), Nagoya, Japan 1997;1997:770–5.

[10] Microsoft Windows Server 2003 Resource Kit, Avail: http://www.microsoft.com/downloads/

[11] M. Rosenstein, "What is Actually Taking Place on Websites: E-commerce lessons from Web Server Logs", *ACM Conference on Electronic Commerce (EC-00)*, October 17-20, 2000, Minneapolis, Minnesota, USA.

[12] Guru Jambheshwar University of Science and Technology Web site, Available: <http://www.gjust.ac.in>

Shakti Kundu received his MCA from Kurukshetra University Kurukshetra, India in 2006, MPhil in Computer Science from Chaudhary Devi Lal University, Sirsa, Haryana, India in 2008, MTech in Computer Science & Engineering from Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India in 2010. Currently he is pursuing his Ph.D in Computer Science from Manav Bharti University, Solan, Himachal Pradesh, India. The author current research interests are web mining and web testing. He is individual member of Computer Society of India and life member of Indian Society for Technical Education (I.S.T.E), New Delhi.