# An Effective Analysis of Weblog Files to improve Website Performance

[1] T.Revathi, [2] M.Praveen Kumar,[3]R.Ravindra Babu, [4]Md.Khaleelur Rahaman, [5]B.Aditya Reddy
Department of Information Technology, KL University, Vijayawada, AP, India.
[1]revathi.talari@gmail.com
[2]prav.rockzzz@gmail.com
[3]ravindra.rompicharla@gmail.com
[4]khaleel420@gmail.com
[5]adityareddy.bommareddy@gmail.com

## Abstract

**As there is an enormous growth in the web in terms of web sites, the size of web usage data is also increasing gradually. But this web usage data plays a vital role in the effective management of web sites. This web usage data is stored in a file called weblog by the web server. In order to discover the knowledge, required for improving the performance of websites, we need to apply the best preprocessing methodology on the server weblog file. Data preprocessing is a phase which automatically identifies the meaningful patterns and user behavior. So far analyzing the weblog data has been a challenging task in the area of web usage mining. In this paper we propose an effective and enhanced data preprocessing methodology which produces an efficient usage patterns and reduces the size of weblog down to 75-80% of its initial size. The experimental results are also shown in the following chapters.**

**Keywords: Web usage mining, Preprocessing, weblog.**

## 1. Introduction

Web usage mining (WUM) is one of the applications of data mining techniques which discover the usage patterns from web usage data. The outcome of this web usage mining can be used for web personalization, website modification, system improvement, and marketing etc. Generally web usage mining[4] consists of 4 stages 1.Data collection 2.Data preprocessing 3.Pattern discovery 4. Analysis and knowledge discovery as shown in fig1.

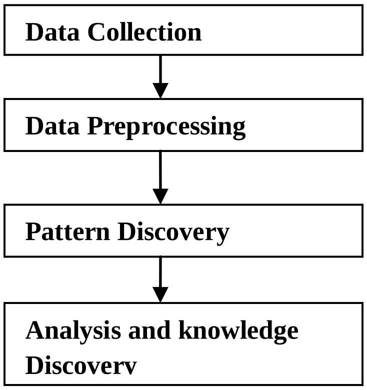


**Figure1: the structure of web usage mining**

Here data preprocessing phase consists of data cleaning, user and session identification and generating the database of log file. Pattern discovery is phase which extracts the usage patterns from the preprocessed data. Pattern analysis and knowledge discovery is final phase of WUM which filters the uninteresting patterns and rules from the interesting patterns. This is a crucial phase of WUM as it identifies the interests and web watching behavior of user's.

Analyzing the user behavior has been a challenging task of web usage mining. If we are able to identify the user interests automatically, the goal of WUM is almost accomplished. This paper deals with the three main stages of WUM i.e preprocessing, pattern discovery and analysis. An efficient preprocessing methodology is discussed with a log file taken from a real time e-commerce website and experimental results are also produced. All frequent patterns of log data, general, access and activity statistics of web log data are shown.

## 2. Data preprocessing

Data preprocessing is a significant step in web usage mining which often consumes more time and effort. This starts with data preparation stage where the original data will be integrated and transformed into a suitable form on which specific data mining operations can be applied. It's shown below[1].

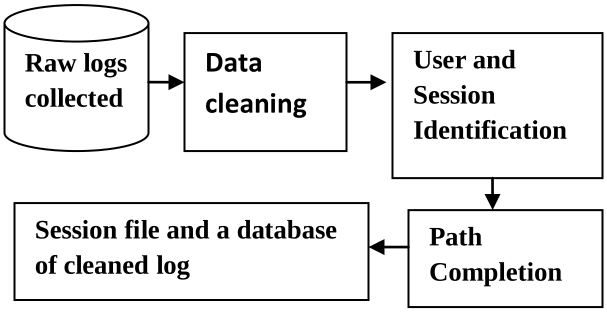

Raw logs collected
Data cleaning
User and Session Identification
Path Completion
Session file and a database of cleaned log

**Figure2: the phases of data preprocessing**

### 2.1 Data collection

The main sources of weblog files in web usage mining are 1.Web Servers 2.Web proxy Servers 3.Client browsers. Here Server logs supply the best usage data than the remaining log files. If at all there are multiple weblog files, they should be converted into a joint weblog file. This joint file will undergo further operations. In this paper a server log file of 7.8 MB is considered for the purpose of analysis.

### 2.2 Data cleaning

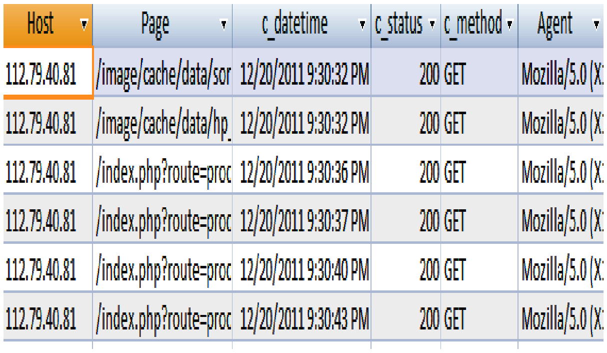
Data cleaning is a process which eliminates the irrelevant and redundant log entries from the original log file and converts the log file into a data base in order to identify individual user's and sessions. That irrelevant information[9] includes 1.Log entries with filename suffixes such as .gif, .jpeg, .jpg, .css, .cgi, can be deleted because these images will auto download with our requested pages. 2. The entries with filename suffixes such as robot.txt can be eliminated as they exceed the range of WUM. 3. Error requests with status codes such as 494(unknown) 404(file not found), 501(inner server error) can be removed. 5. Delete all the entries of request methods except 'GET" method. An algorithm for data cleaning is depicted as shown below.

```
Input: log files (LF)

Output: Summarized log file (SLF)

Begin

1. Read records in LF

2. For each record in LF

3. Read fields (Status code, method)

4. If status code='200' and method='GET' Then

5. Get IP_address and URL

6. If URL_suffix is={.gif,.js,.jpeg,.jpg,.css} Then

7. Remove URL_suffix
8. Save IP_address and URL
   End if
   Else
9. Next record
 End if
 End
```

**Algorithm for Data cleaning**

### 2.3 User identification

User identification means identifying each user accessing website, whose goal is to mine every user's access characteristic, and the make user clustering and provide



| Host | Page | c_datetime | c_status | c_method | Agent |
|---|---|---|---|---|---|
| 112.79.40.81 | /image/cache/data/sor | 12/20/2011 9:30:32 PM | 200 | GET | Mozilla/5.0 (X |
| 112.79.40.81 | /image/cache/data/hp_ | 12/20/2011 9:30:32 PM | 200 | GET | Mozilla/5.0 (X |
| 112.79.40.81 | /index.php?route=proc | 12/20/2011 9:30:36 PM | 200 | GET | Mozilla/5.0 (X |
| 112.79.40.81 | /index.php?route=proc | 12/20/2011 9:30:37 PM | 200 | GET | Mozilla/5.0 (X |
| 112.79.40.81 | /index.php?route=proc | 12/20/2011 9:30:40 PM | 200 | GET | Mozilla/5.0 (X |
| 112.79.40.81 | /index.php?route=proc | 12/20/2011 9:30:43 PM | 200 | GET | Mozilla/5.0 (X |

**Figure 3: A database of cleaned log**

personal service for the users. But user identification is complicated by the presence of local caches, proxy servers. We assume that each user has unique IP address and each IP address represents one user. But in fact there are three conditions: (1)Some user has unique IP address.(2).Due to proxy server, some user may share one IP address .As of now, we propose following rules[4] for user identification: If there is a new IP address, then there is a new user. If the IP address is same, but the operating system or browser are different, then assumption is that each different agent type for an IP address represents a different user. Moreover we give some notations for user identification.

Here Usersi= {User_ID, User_IP, User_Url, User_Time, User_RefferPage, User_Agent), 0<i<n where i is the no of total users; User_ID is user's ID have been identified. User_IP is the user's IP address. User_Url is the web page accessed. User_Time is the time at which user accessed. User_RefferPage is the last page that the user requested. User_Agent is the agent user used. By applying all those above rules, we can easily identify the individual users.

## 2.4 Session identification

The goal of session identification is to divide the page requests of each user into individual sessions i.e we find each user's access pattern and frequent path. The best method to identify a session is using a timeout mechanism.

We use the following rules[9] in our experiment to identify individual sessions.1.If there is a new user, then there is a new session 2.If the referrer page is null in a user session, then we can make sure that there is a new session.3.If the time between page requests exceeds certain limit (25 or 30 minutes), we can assume that user is starting a new session. Here we propose some notation notations which help us to identify user's sessions.

Sessionsi={User_ID,Sj,[urlj1,urlj2,…urljk]), 0<i<n where n is the total no of sessions. User_ID stands for user's ID that has been identified; Sj stands for one of the user's sessions; urljk stands for aggregate of web pages in session Sj. By applying the above stated rules, we can identify user's sessions.

## 2.5 Path completion

Because of local buffers existence, some requested pages will not be recorded in access log. The goal of path completion is to fill in all the missing references that are not recorded. The solution for path completion[9] is that if a requested page can be reachable by a hyperlink from any of the visited pages by the user, we assume that it should be added in the session. When there are two or more pages which have a super link to it in one session, then it should be placed before the latest visited page.

## 3. Experimental Results

In this paper we took the server log file of a e-commerce site dreamers.net.in whose size is around 7.8 MB. Various analyses have been done to identify the user web watching behavior and interests.

**Table1: format of the weblog taken**

| host | url | time | status | method | agent |
|---|---|---|---|---|---|

Here the results of extracted user profiles have been in the form of tables. Table2 table 3, table4 shows user's daily access statistics, top hosts, most requested images etc.

**Table2: user's daily access statistics**

| Date | Hits | Page Views | visitors |
|---|---|---|---|
| 1-1-2012 | 70 | 20 | 3 |
| 2-1-2012 | 307 | 38 | 7 |
| 3-1-2012 | 822 | 162 | 8 |
| 4-1-2012 | 393 | 78 | 3 |
| 5-1-2012 | 195 | 17 | 5 |
| 6-1-2012 | 88 | 3 | 1 |

**Table 3: Top Hosts**

| S no | Hosts | Hits | Bandwidth |
|---|---|---|---|
| 1 | 117.211.85.58 | 1559 | 9987 |
| 2 | 115.248.116.137 | 193 | 782 |
| 3 | 65.255.37.250 | 62 | 121 |
| 4 | 202.62.86.58 | 346 | 2395 |

| | | | |
|---|---|---|---|
| 5 | 59.93.118.237 | 51 | 224 |
| 6 | 202.133.58.89 | 106 | 401 |
| 7 | 203.6.213.145 | 18 | 158 |

**Table 4: Most requested images**

| Images | Hits |
|---|---|
| image/data/logo.png | 46 |
| image/data/cart.png | 40 |
| img-sys/headerbg.png | 47 |
| image/cache/data/imac_90*90.png | 51 |
| catalog/view/theme/default.png | 65 |
| image/cache/data/palm_logo.png | 54 |
| image/cache/data/ipod_shuffle.png | 48 |
| image/cache/data/ipod_touch.png | 46 |
| image/cache/data/hp_logo.png | 72 |
| catalog/view/theme/default/add.png | 34 |
| image/cache/data/hp_banner.png | 76 |
| image/cache/data/macbook_pr.png | 43 |
| cache/data/iphone_1-80*80.png | 56 |

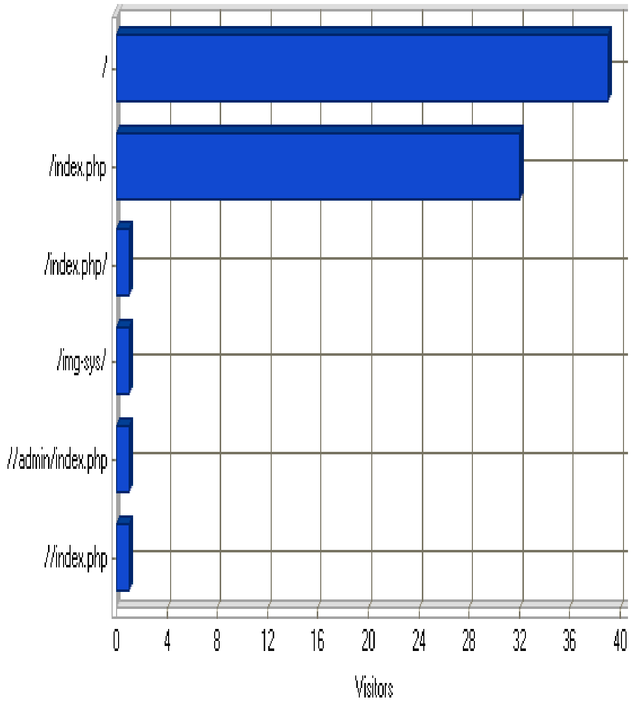


**Fig 4: Most popular pages**

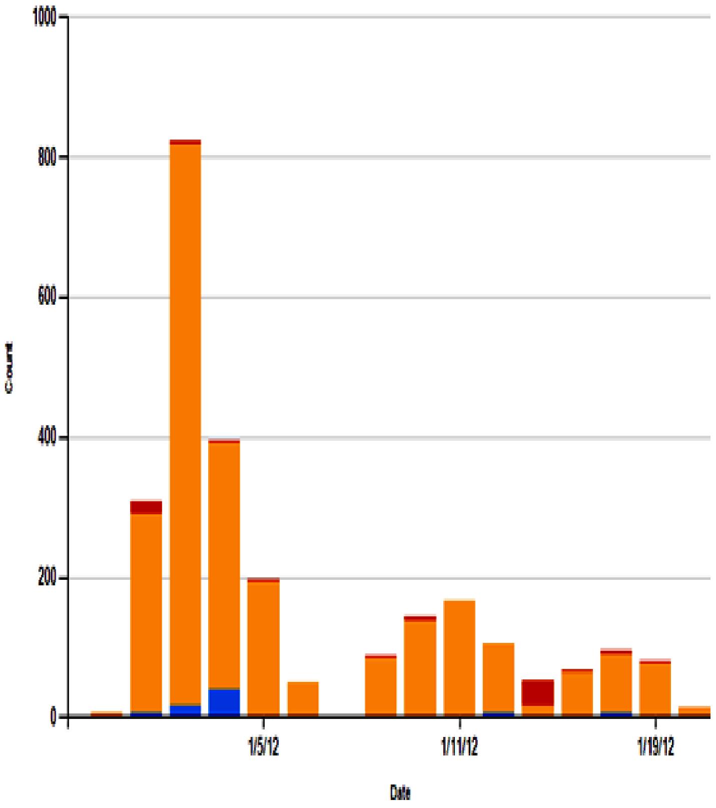


**Fig 5: Status codes**

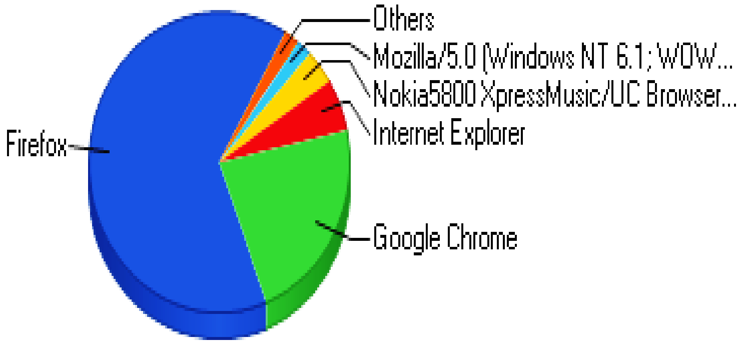


**Fig 6: Most popular browsers**

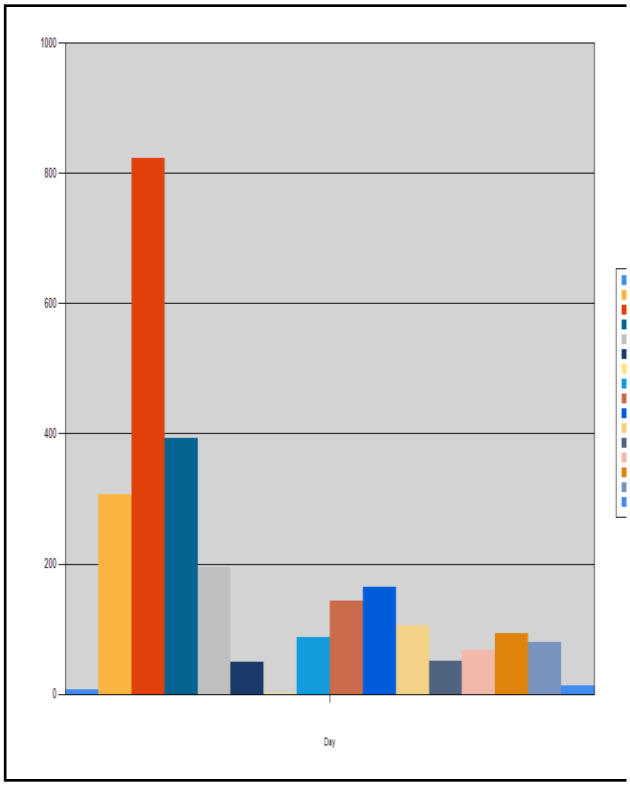


**Fig 7: Hits per each day in January**

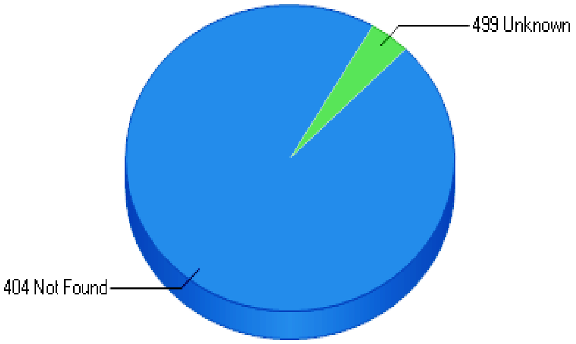


**Fig 8: Error Types**

## 4.Conclusion And Future work

In this paper we presented an efficient preprocessing methodology for web log data and we described some techniques to identify individual users and sessions. And we analyzed a server log file of an ecommerce site to determine some statistics like user's daily access statistics, top hosts, most popular requested pages, most popular images, browsers, hits per each days etc. Apart from that, status codes and frequent error types are also shown to help the system administrator and web designer in improving the performance of the web site. Similar studies can be done for any other web sites to improve their performance.

But this work can be extended by applying data mining techniques like association, classification, clustering to a group of regular users to find frequently accessed patterns that leads to high accuracy and performance.

## Acknowledgement

We are very thankful to the faculty of IT Department, KL University, Vijayawada, India for guiding us throughout this research.

## References

[1].K.R Sunnetha, R.krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009

[2].R. Cooley, B. Mobasher, and J. Srivastava" Web Mining: Information and Pattern Discovery on the World Wide Web.

[3].Li Chaofeng "Research and Development of Data Preprocessing in Web Usage Mining"

[4].Suneetha K.R, Dr. R. Krishnamoorthi "Data Preprocessing and Easy Access Retrieval of Data through Data Ware House" Proceedings of the World Congress on Engineering and Computer Science 2009.

[5].Mr. Sanjay Bapu Thakare, Prof. Sangrarn. Z. Gawali "A Effective and Complete Preprocessing for Web Usage Mining" International Journal on Computer Science and Engineering.

[6].Jaideep Srivastava,Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data"

[7].G T Raju and P S Satyanarayana "Knowledge Discovery from Web Usage Data: CompletePreprocessing Methodology" International Journal of Computer Science and Network Security.

[8] Berendt, B., B. Mobasher, M. Nakagawa, and M. Spiliopoulou, "The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis," Lecture Notes in Computer Science, Vol. 2703, pp. 159-179, September, 2003.

[9].Fang yuan"', Li-Juan wang', ge yu' "Study on data preprocessing algorithm in web log mining" Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003.

## About Authors

Ms.T Revathi is currently working as Assistant Professor in the department of IT, KL University, Vijayawada, AP, and India. She received her master's degree from JNTU Kakinada. She published many papers in international journals. Her area of interests includes web usage mining, advanced computer architectures.

M.Praveenkumar,R.RavindraBabu,Md.Khallelur Rahaman,B.Aditya Reddy, all are pursuing B.tech final year in the Department of InfromationTechnology, KL University, Vijayawada,AP,India. Their area of interests is data mining, android and cloud computing.