November.2013-Volume 1.No3
ISSN 2345-3397

User Behavior Analysis from Web Log using Log Analyzer Tool

A.Brijesh Bakariya, B.Ghanshyam Singh Thakur

Department of Computer Application, Maulana Azad National Institute of Technology, Bhopal, India
brijesh_scs@yahoo.co.in
Department of Computer Application, Maulana Azad National Institute of Technology, Bhopal, India
ganshyamthakur@gmail.com

Abstract

Now a day, internet plays a role of huge database in which many websites, information and search engines are available. But due to unstructured and semi-structured data in webpage, it has become a challenging task to extract relevant information. Its main reason is that traditional knowledge based technique are not correct to efficiently utilization the knowledge, because it consist of many discover pattern, contains a lots of noise and uncertainty. In this paper, analyzing of web usage mining has been made with the help if web log data for which web log analyzer tool, "Deep Log Analyzer" to find out abstract information from particular server and also tried to find out the user behavior and also developed an ontology which consist the relation among efficient web apart of web usage mining.

Keywords: Web Usage Mining, Preprocessing, Ontology, Web Log Analyzer, Web Log.

I. Introduction

Due to huge volume of web data, web data is not available in proper structure, due to which the searching result is irrelevant and also consist noise. It is a complex task of searching and retrieving data or information from log data. Ontology and web usage mining are interrelated (Bowman Vladan Devedzic, 2006). Using web usage mining technique, we can represent the provided information through ontology (Theint Aye, 2011). In web usage mining, web user profile is identified, for this work, different data mining methods can be used like classification, clustering, association rules (J. Han et al, 2006). For developing the ontology, Semantic web plays a vital role (Thomas B. Passin). The term of semantic web is a meaningful or intelligent web which is very efficient in real. Its main purpose is to convert data in machine understandable manner in the form of ontology (R.M. Suresh, 2007),(A.C.M. Fong et al, 2012). The Semantic web describes the relationship between things (For example: P is a part of Q and R is a member of S) and the property of things (Like size, weight, age and price). It depends upon efficient web data to make semantic web an intelligent web (Sanjay Kumar Malik et al, 2010). The web log contains sequence of URL's accessed by the user. Using web usage mining, frequent information can be accessed according to user (Jaideep Srivastava et al, 2000). Though web log data we can find out user behaviour maintain user session and extract hidden information of user and web. The term "Hidden Information" means for how much time user accessed any particular website, how many persons visited on any particular website, what is the interest ?,those information like this mentioned above can be provided using log data. For getting this information, we need to use efficient Log Analyzer Tool (Brijendra Singh, 2010). After analyzing the behavior of the user by using such type of tools we can flash advertisement to that website according to the interest of the user. The log of the website is maintained by the server. For this work, server needs such tool which can analyze that log data and can discover the Pattern. Server uses web mining technique for different types of web data (Etminani et al, 2009).

A. Web Mining

Web mining is the appliance of data mining functionality which is used to mine relevant information from log data (Wang Bin et al, 2003), (Mahendra Pratap Yadav et al, 2012). Whatever interesting data has to retrieve from Web, It is also possible through web mining (Yuefeng Li et al, 2007). Today huge amount of data is available on the web to extract data from such vast collection is a complex task. By applying some data mining method, we can find out useful pattern using web mining (R. Cooley et al, 1997). Web mining has been classified into three types:

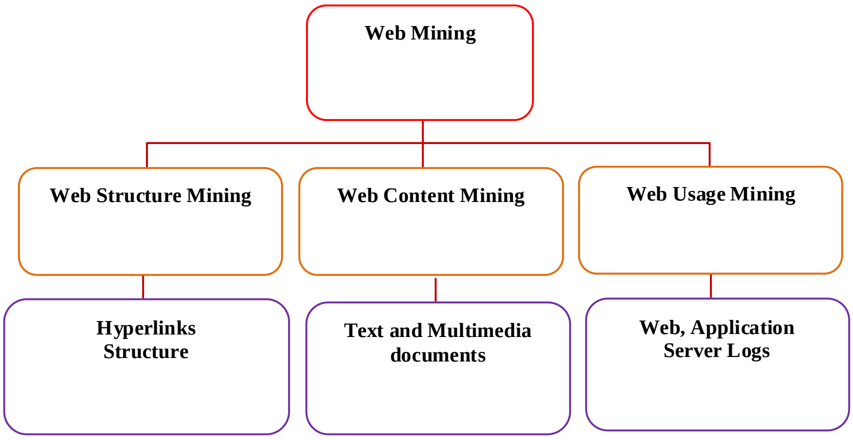


Figure. 1. Classification of web Mining

a) Web Structure Mining

Through Web Structure mining, useful information is extracted from hyperlink (Miguel Gomes et al, 2005). Which represent the structure of web, just like we can discover important page, we can analyze it. It is widely used in search engines. Through Web Structure mining we can discover community of such user who share common interest. In traditional Data mining technique such tasks are not performed because generally in relational table such link structure is not present (G.K. Gupta et al, 2011).

b) Web Content Mining

Web Content Mining extracts useful information from web page contents. Through web content mining, we can automatically classify and cluster according to web page topics. This task is similar to data mining techniques. In web content mining we can deal with document in the form of text, audio, video, image and other contents.

c) *Web Usage Mining*

Web usage Mining is the application of Data Mining technique to discover user access pattern from web usage log (Jaideep Srivastava et al, 2000). It is used to provide better services of the website through which every click done by the user is measured.

B. *Semantic Web*

Semantic Web is an extension of World Wide Web (WWW). It is about collection and mining important information from knowledge which is embedded in web application. The basic idea of Semantic Web is to expand the people enable web by converting some semantics of resources in machine computational form (Bowman Vladan Devedzic, 2006). Due to which the computer is able to process, search, integrate and present content in a better way. Which is generally happens in an intelligent and meaningful manner (Neha Goel et al, 2013).

C. *Ontology*

Ontology is a term that can define in many ways. Genersereth and Nilsson defined ontology as an “explicit specification of a set of object, concepts and other entities that are presumed to exist in some area of interest and the relationship that holds them”. Through the above definition, we can say that ontology is domain dependent and it is designed for sharing and reusability. Ontology always tells abstract content properties and relationship for improving searching through web data, for making reusing capabilities in a more better way, In web personalization ,ontology based web usage mining is used in user profile learning (R.M. Suresh, 2007), (Mariangela Vanzin et al, 2005).

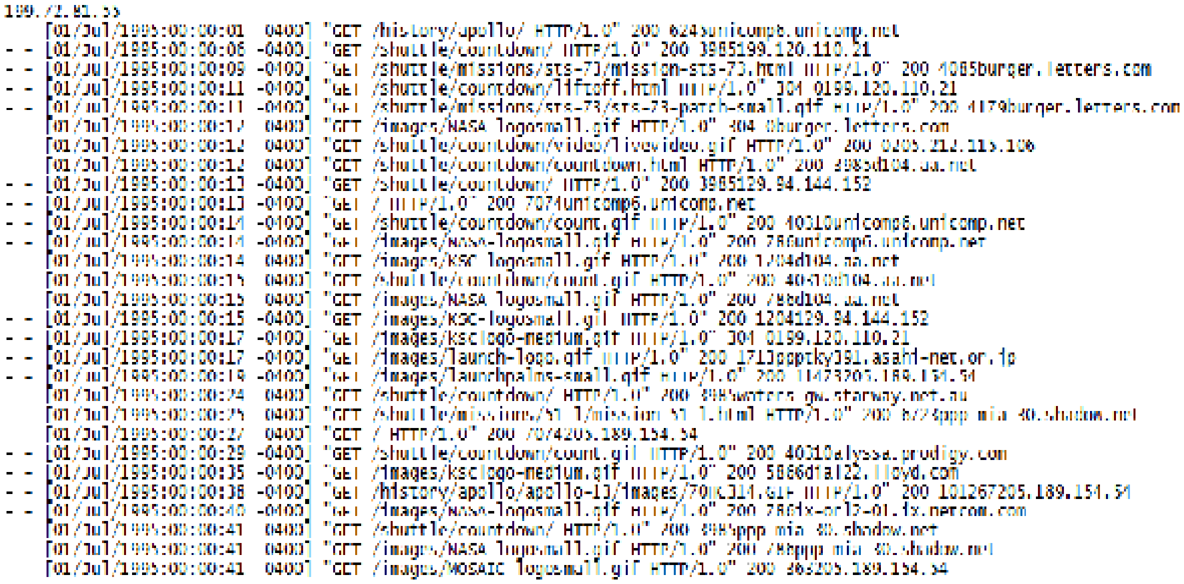
In web usage mining, whenever any user interacts with any website then ‘clickstream’ data needs to be maintained (S. Singh Anand et al, 2004) Clickstream is an aggregate sequence of any page visit which is formed through the page navigation of the user. In clickstream, logs cookies and other data is present which is used to transfer web page from server to the users.

II. **Web Server Log**

Statistics of any website is based on server log. Server log is a simple text file, which records every activity of visitors. The log file used or maintained on server log should be in a proper format which is called as file format. Now days, many file formats are available which are used by the server.W3C (World Wide Web Consortium) W3C extensible log file format (W3C log file, 2003). It is text based, customizable format for single site. This is the default file format. Microsoft IIS log file:- it is text based, fixed format for single site. NCSA (National Centre for Supercomputing Application):- It is common log file format. It is text based fixed format for single site.

In W3C file format different types of fields are present (W3C log file, 2003). For example- Date-The date on which the activity occurred., Time-The time at which the activity occurred, Client IP addresses-The IP address of the client that made the request, User name-The name of that particular user who accessed your server; anonymous user indicated by a “Hyphen” (-), Service Name- It tells the internet se vice name on which the log file entry was generated, Server IP address-The IP address of the server on which log file entry was generated, Server Port- The server port was configured for the service, Method- The requested action for ex: GET Method, URI item-The target of the action, for ex: Default.htm, URI Query- The query, if any that the client was trying to perform. A universal resource identifier (URL) query is necessary only for Dynamic pages, HTTP Status- The HTTP Status code, WIN 32 Status- The windows status code.

BYTE sent- The number of bytes that the server sent. BYTES received- The number of bytes that the server received, TIME taken- The length of time that the action took in millisecond, Protocol version- The protocol version-HTTP or FTP that the client used, HOST- The host header name, if any, User Agent- The Browser type that the client used, Cookie- The content of the cookie sent or received, If any, Reference- The site that the user last visited this site provided a link to the current site, Protocol Sub status- The sub status error code. In IIS file format contains different types of fields like: Client IP address- The IP address of the client, User Name- The name of the user, Date-The date of the log file entry, Time- The time of the log file entry, Service and instances-The Internet service name on which the log file entry is made, Server Name-The name of the server, Server IP- The IP address of the server, Time Taken-The time taken to fulfill the request, Client Byte sent-The number of bytes sent from the client to the server, Server Byte sent-The number of bytes sent from the server to the client, Request Type- The type of request generated by the client, Target of operation-What was the target of the client. Definitely all fields will not contain information. The hyphen is used for the fields that have no information. If a field contains non printable character, Default.sys replaces it with a plus (+) sign. In NCSA file format contains different fields. Remote Host Address-The IP address of the client, Remote LOG Name- The name of the server. Date and Time-The date and time on which the log file entry was created, Request and protocol version-The type of request made by the client and version of the protocol used, Service status Code-The status of the service provided by the server. Bytes Sent-The number of bytes sent. In this article, the data source which is in IIS file format, for the finding hidden information of visitor is collected by NASA-HTTP. The log file is available at The Internet Traffic Archive sponsored by ACM SIGCOMM. We use the part of the logs during the period of 1 July to 31 July1995. For session identification, set the maximum elapsed time to 30min, which is used in many commercial applications (Web Log Data, 1995), (Internet Traffic Archive, 2000)

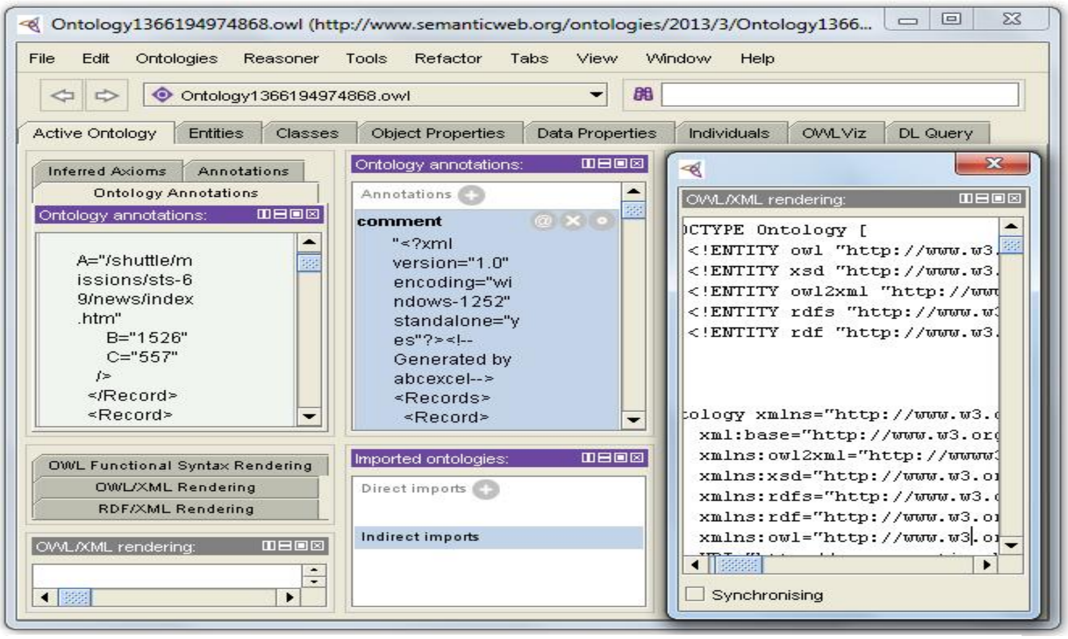


```
199.72.81.55
    [01/Jul/1995:00:00:01  0400] "GET /history/apollo/ HTTP/1.0" 200 6245unicomp6.unicomp.net
- - [01/Jul/1995:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985199.120.110.21
- - [01/Jul/1995:00:00:09 -0400] "GET /shuttle/missions/sts-73/mission-sts-73.html HTTP/1.0" 200 4085burger.letters.com
- - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/countdown/liftoff.html HTTP/1.0" 304 0199.120.110.21
- - [01/Jul/1995:00:00:11 -0400] "GET /shuttle/missions/sts-73/sts-73-patch-small.gif HTTP/1.0" 200 4179burger.letters.com
    [01/Jul/1995:00:00:12  0400] "GET /images/NASA logosmall.gif HTTP/1.0" 304 0burger.letters.com
    [01/Jul/1995:00:00:12  0400] "GET /shuttle/countdown/video/livevideo.gif HTTP/1.0" 200 0205.212.115.106
    [01/Jul/1995:00:00:12  0400] "GET /shuttle/countdown/countdown.html HTTP/1.0" 200 3985d104.aa.net
- - [01/Jul/1995:00:00:13 -0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985129.94.144.152
- - [01/Jul/1995:00:00:13 -0400] "GET / HTTP/1.0" 200 7074unicomp6.unicomp.net
- - [01/Jul/1995:00:00:14 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310unicomp6.unicomp.net
- - [01/Jul/1995:00:00:14 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786unicomp6.unicomp.net
    [01/Jul/1995:00:00:14  0400] "GET /images/KSC logosmall.gif HTTP/1.0" 200 1204d104.aa.net
    [01/Jul/1995:00:00:15  0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310d104.aa.net
    [01/Jul/1995:00:00:15  0400] "GET /images/NASA logosmall.gif HTTP/1.0" 200 786d104.aa.net
- - [01/Jul/1995:00:00:15 -0400] "GET /images/KSC-logosmall.gif HTTP/1.0" 200 1204129.94.144.152
- - [01/Jul/1995:00:00:17 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0199.120.110.21
- - [01/Jul/1995:00:00:17 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713ppptky391.asahi-net.or.jp
- - [01/Jul/1995:00:00:19 -0400] "GET /images/launchpalms-small.gif HTTP/1.0" 200 11473205.189.154.54
    [01/Jul/1995:00:00:24  0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985waters gw.starway.net.au
    [01/Jul/1995:00:00:25  0400] "GET /shuttle/missions/51 l/mission 51 l.html HTTP/1.0" 200 6724ppp mia 40.shadow.net
    [01/Jul/1995:00:00:27  0400] "GET / HTTP/1.0" 200 7074205.189.154.54
- - [01/Jul/1995:00:00:29 -0400] "GET /shuttle/countdown/count.gif HTTP/1.0" 200 40310alyssa.prodigy.com
- - [01/Jul/1995:00:00:35 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 200 5866dial22.lloyd.com
- - [01/Jul/1995:00:00:38 -0400] "GET /history/apollo/apollo-13/images/70K314.GIF HTTP/1.0" 200 101267205.189.154.54
- - [01/Jul/1995:00:00:40 -0400] "GET /images/NASA-logosmall.gif HTTP/1.0" 200 786ix-or13-01.ix.netcom.com
    [01/Jul/1995:00:00:41  0400] "GET /shuttle/countdown/ HTTP/1.0" 200 3985ppp mia 40.shadow.net
    [01/Jul/1995:00:00:41  0400] "GET /images/NASA logosmall.gif HTTP/1.0" 200 786ppp mia 40.shadow.net
    [01/Jul/1995:00:00:41  0400] "GET /images/MOSAIC logosmall.gif HTTP/1.0" 200 363205.189.154.54
```

Figure. 2. Sample of Web log data

III. Example of Ontology Development

Protege version 4.0.2 tool has been used to give an idea of developing ontology, basically ontology based software provide web ontology language. This is the snap for creation ontology for a web log data.



```
A="/shuttle/m
issions/sts-6
9/news/index
.htm"
    B="1526"
    C="557"
  />
 </Record>
 <Record>
```

```
"<?xml
version="1.0"
encoding="wi
ndows-1252"
standalone="y
es"?><!--
Generated by
abcexcel-->
<Records>
 <Record>
```

```
)CTYPE Ontology [
 <!ENTITY owl "http://www.w3.
 <!ENTITY xsd "http://www.w3.
 <!ENTITY owl2xml "http://www
 <!ENTITY rdfs "http://www.w3
 <!ENTITY rdf "http://www.w3.


tology xmlns="http://www.w3.
  xml:base="http://www.w3.org
  xmlns:owl2xml="http://wwww3
  xmlns:xsd="http://www.w3.or
  xmlns:rdfs="http://www.w3.o
  xmlns:rdf="http://www.w3.or
  xmlns:owl="http://www.w3.or
```

Figure. 3. Tool to develop ontology

A. XML Code Snippet

Extensible Markup Language (XML) allows you to explain and arrange information in ways that are effortlessly understandable by both computers and human. We can then share that information and its description with others over the internet, network, or in other ways, ontology can be represent in the form of ontology. Following code snippet of XML is said in the ontology.

```
<?xml version="1.0" encoding="windows-1252" standalone="yes"?><!-- Generated by
abcexcel-->
<Records>
 <Record>
  <Row
   A="FileName"
   B="Page Views"
   C="Data Transferred (Kb)"
  />
 </Record>
 <Record>
  <Row
   A="/shuttle/countdown/index.htm"
```

```
    B="40255"
    C="147714"
  />
</Record>
<Record>
  <Row
    A="/ksc.html"
    B="40090"
    C="264054"
  />
</Record>
<Record>
  <Row
    A="/index.htm"
    B="32674"

  <Row
    A="/shuttle/missions/missions.html"
```

IV. Web Log Analyzer

Web log analyzer software a type of web statistics software that divide a log file from a web server, and based on the data contained in the log file derives indicators about how, when and by whom a server is visited. Generally reports are generated from the log files directly, but the log files can alternatively be parsed to a database and reports generated on demand.
System administrators and webmaster for the purposes of “how much traffic they are getting, how many requests successes or unsuccessful, and what kind of problems are being generated”, or such same kinds of information. Analyzing and discovering log could be of help in improvement website and user interactions and performance, detecting user visits and strategy of navigation to trace quality of service and so on (S. Singh Anand et al, 2004). There may be different types of Web Log Analyzer like “Deep Log Analyzer”, which has been illustrated below.

A. Deep Log Analyzer

Deep Log Analyzer is highly developed and inexpensive web statistical solution for small and medium size websites. We can examine web site user’s behavior, when they interact with the websites and obtain complete website usage statistics in various ways with this analytics software. The software knows exactly where your customers came from, how they moved through your site and where they left it. This complete information will help us. When more users interact with website then this software convert these visitors to satisfied customers. With Deep Log Analyzer we can analysis reports from log data source, visitors (users) activity and pattern of navigation, sites that refer web traffic, search queries, visitor’s browsers and operating systems, web server errors and much more. Website statistics software makes it easy to analysis how statistics changes over period of time and compare it. We can examine deeper your website statistics data with the unique hierarchical reports which Deep Log Analyzer tool, It can analyze logs of IIS web servers. It can read Zip and Gz compressed log files so that we won't need to extract them manually .For implementing the Web Usage Mining, Web Server Logs may be required.

V. Experimental Results

In the current research web access logs were taken from an NASA-HTTP website and log file is available at The Internet Traffic Archive sponsored by ACM SIGCOMM for the time period 1 July to 31 July1995 and the following results were obtained (Web Log Data,1995),(Internet Traffic Archive, 2000).

A. *General Statistics*:

In this part we get common information related to the website like how many times the website was hit, an average of hits in a day, bandwidth, page views etc. It enlists all the common information which one should know related to a website.

B. *Activity Statistics*:

The feature of "Deep Log Analyzer" is provides the statistics on daily and hourly basis. Apart from the report in tabular form, it also gives a graphical chart which helps in ascertaining at which no. Of visit on a date, hour of the day along with the days on which the website was hit the maximum etc.. By utilizing this information special schemes can be initiated which might help in increasing the people who access the website.

TABLE I
GENERAL STATISTICS OBTAINED AFTER ANALYZING WEB LOGS

Hits	
Number of Hits:	1,887,881
Number of Successful Hits:	1,877,037 (99%)
Visitors	
Number of Unique visitors	81,893
Visitors who visited once	58,497 (71%)
Repeat visitors:	23,396 (29%)
Visits	
Number of Visits	161,918
Average Number of visit per day	5,783
Page Visews	
Total Page Views	688,639
Most popular page (/shuttle/countdown/index.htm)	40,255
Most popular download (/payloads/schedule.../fawgman.pdf)	20
Most popular Entry Page (/index.htm)	24,422
Most popular Exit Page (/ksc.html)	16,081
Technical Summary	
Error Hits	10,844 (1%)

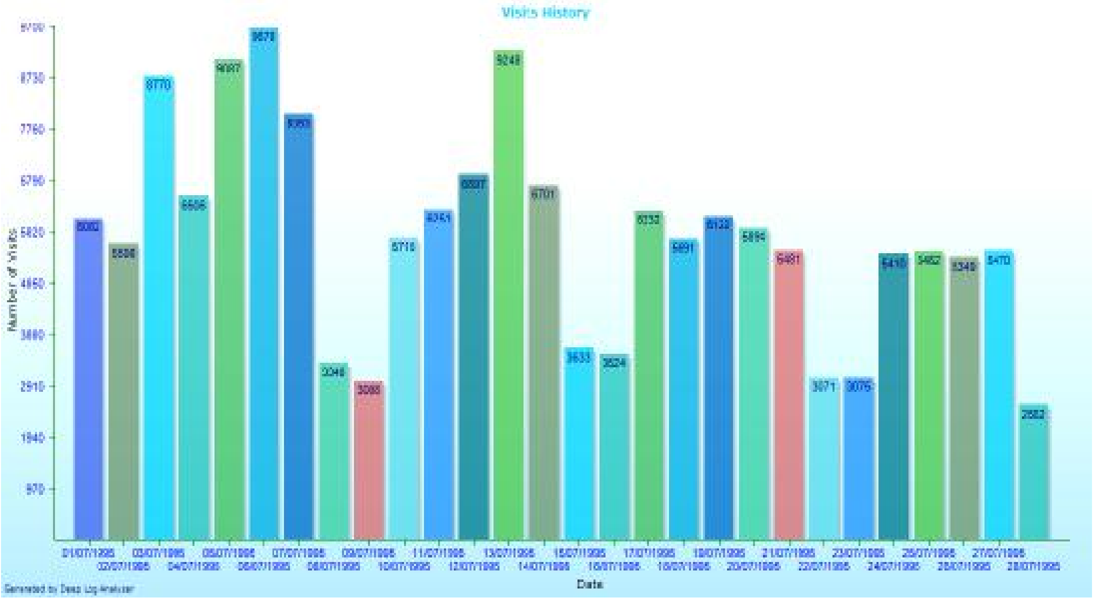


Figure. 4. Activity statistics showing at no. of visit on particular date

C. *Access Statistics*

This part of the analysis can be considered the most important as it not only ascertains which page is hit the maximum number of times according to country wise rather and thus help in recreating the website to suit the need of customer. It states that which maximum number of hits amongst all the country.

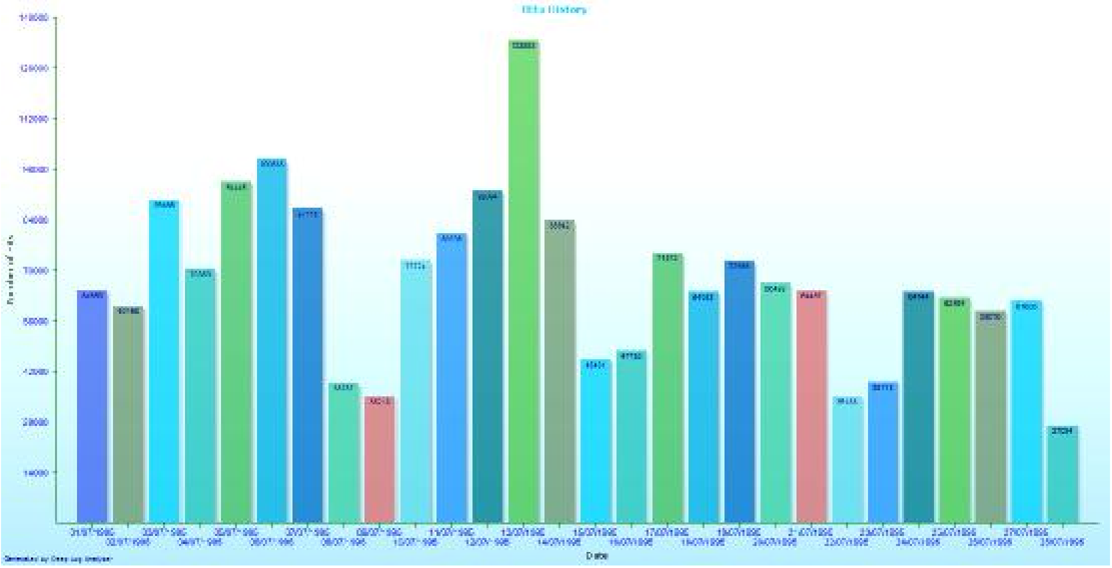


Figure. 5. Activity statistics showing at no. of hits on particular date

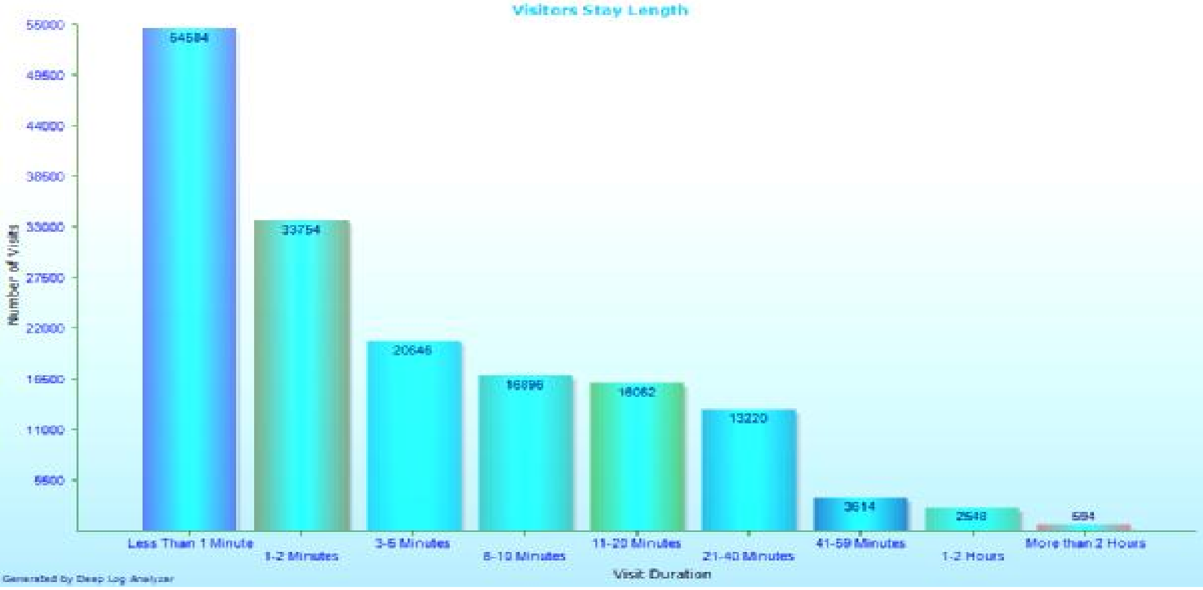


Figure. 6. Activity statistics showing at no. of hits on particular date

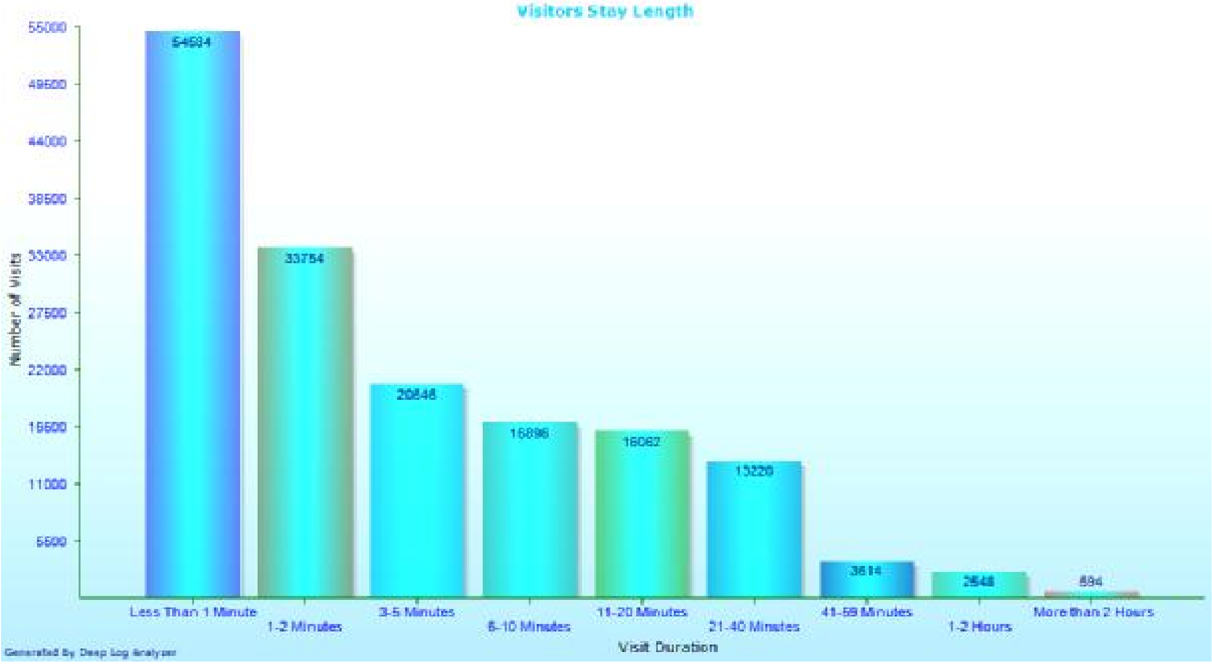


Figure. 7. Activity statistics showing at no. of visits on particular time

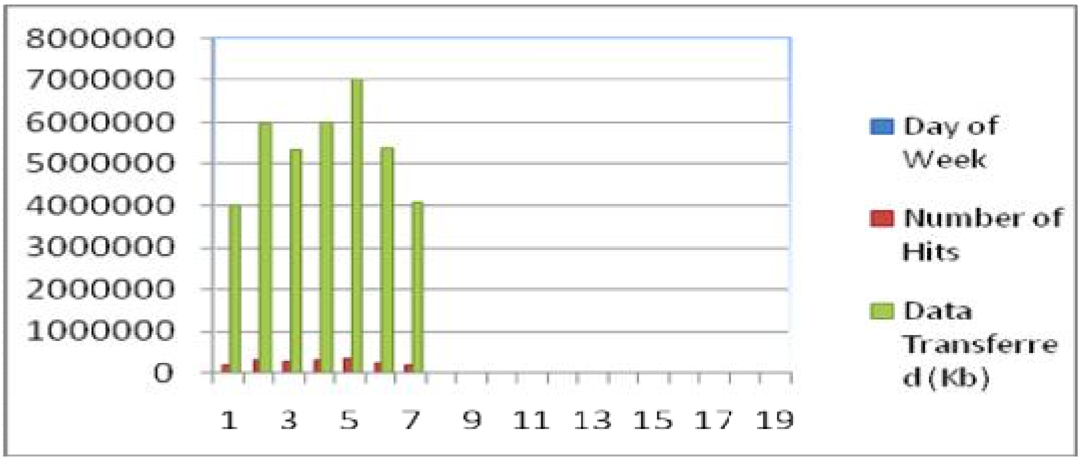


Figure. 8. Statistics showing no. of hits on day of week and data transferred

D. *Errors*

The last feature provided by Deep Log Analyzer tool is finding out what kind of errors user face when they access the website. For the error feature both a tabular and a graphical form of representation are available.

TABLE II

GENERAL STATISTICS OBTAINED SOME TECHNICAL ERROR BY SERVER

Error	Number of Hits
404 - File Not Found	10,714
500 - Internal Server Error	62
403 - Forbidden	54
501 - Not Implemented	14

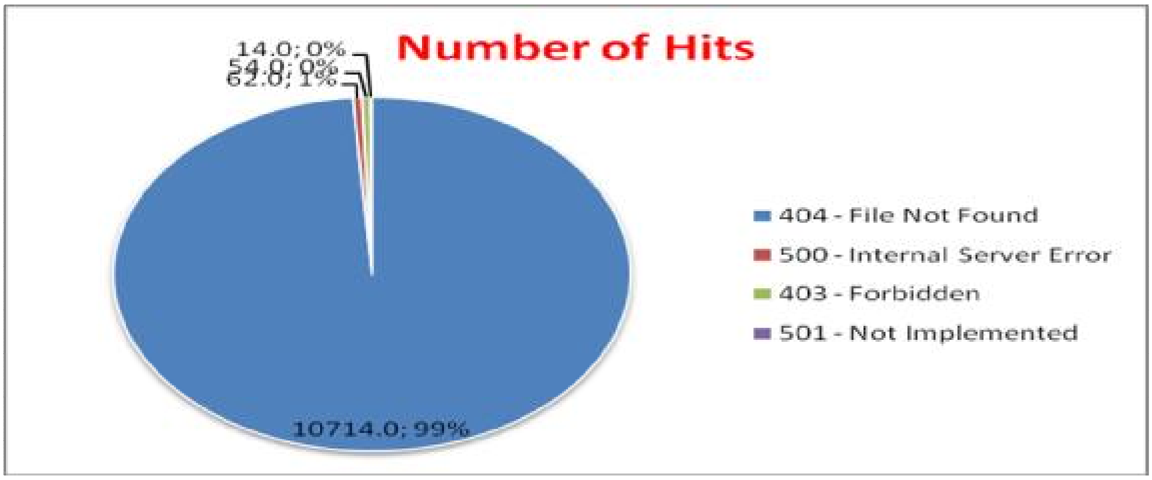


Figure. 9. Various errors due to services of server

VI. Ontology and Web Usage Mining

Web Usage Mining refers to the mine the knowledge from Server which contained Web pages, hyperlinks, and Web log data. Figure 11 relates Semantic Web, Ontology and Web Usage Mining. The traditional topics covered by Web usage mining includes Web clustering, Web page classification and Web extraction where Ontology may be applicable as background semantic structures for Web mining (Wang Bin et al, 2003). In this figure we are trying to relate ontology with web usage mining

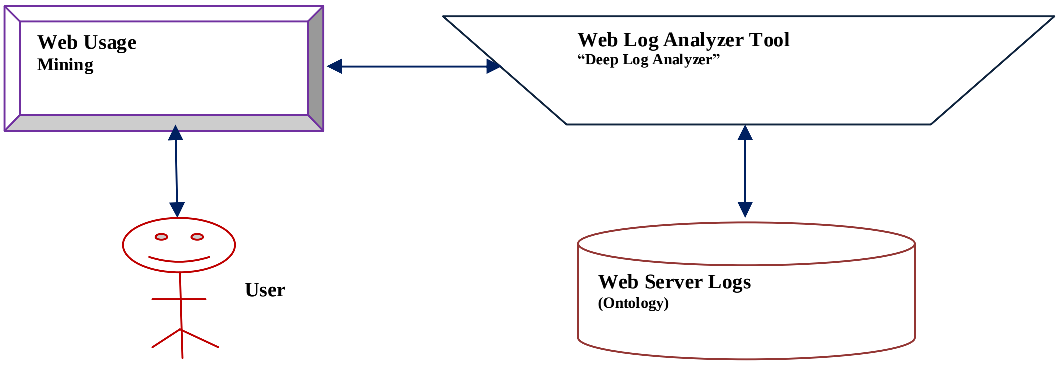


Figure. 10. Relation between Ontology and Web Usage Mining

VII. Conclusion

Deep Log Analyzer tools are a part of Web Analytics Software. They take input in terms of web log file, analyze it and generate results. A various kinds of tools are available which propose huge capabilities in preparing and reporting the results of analysis. A study was done and various types of tools were studied. Every tool offered some or the other feature which was better than the rest. Finally the tool, Web Log Analyzer was taken to analyze the web server logs of the website as it provided wide features. The results were checked and are being tried to integrate in the website of the user. Such log analyzer tools should be widely used as they help a lot in understanding the user behavior to the analysts, and also analyze of the data being machine understandable in the form of Ontology and Web Usage Mining for exploring the interesting knowledge abstract in the data and the usage of a Deep Log Analyzer for pattern discovery and analysis.

Acknowledgements

We wish to thank the anonymous Reviewers and the authors of references.

References

i. Bowman Vladan Devedzic (2006), University of Belgrade, Serbia and Montenegro, "Semantic Web and Education", Springer, pp 38-43.

ii. Thomas B. Passin, "Explorer's Guide to the Semantic Web", pp-161, pp-33, manning publications.

iii. R.M. Suresh (2007), "A Study on the Ontology Based Web Mining For Digital Library", IET-UK International Conference on Information and Communication Technology in Electrical Sciences, ICTES .

iv. Wang Bin, LiuZhijing (2003), "Web Mining Research ", Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA), IEEE.

v. R. Cooley, B. Mobasher and J. Srivastava (1997), "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence ICTAI.

vi. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM, SIGKDD.

vii. Mahendra Pratap Yadav,Pankaj Kumar Keserwani and Shefalika Ghosh Samaddar (2012),"An Efficient Web Mining Algorithm for Web Log Analysis: E-Web Miner",IEEE,.

viii. J. Han and M. Kamber (2006), "Data Mining: Concepts and Techniques". A. Stephan. San Francisco, Morgan Kaufmann Publishers is an imprint of Elsevier.

ix. Khasawneh N.,Shatnawi, M.,Fraiwan (2010), "Converting Web Applications into Standard XML Web Services "The Tenth International Conference on Intelligent System Design and Applications.

x. Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani (2009), "Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM", First International Conference on Networked Digital Technologies.

xi. S. Singh Anand, M. Mulvenna, and K. Chevalier (2004), "On the Deployment of Web Usage Mining", EWMF 2003, LNAI 3209, Springer-Verlag Berlin Heidelberg.

xii. Miguel Gomes, da Costa Junior, Zhiguo Gong (2005), "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, June 27 - July 3, Hong Kong and Macau, China.

xiii. Theint Aye (2011) Web Log Cleaning for Mining of Web Usage Patterns, IEEE.

xiv. Brijendra Singh, Hemant Kumar Singh (2010), Web Data Mining Research: A Survey, IEEE.

xv. G.K. Gupta, Introduction to Data Mining with Case Studies (2011), Web Data Mining, PHI Learning Private Limited, pp. 231-233.

xvi. Sanjay Kumar Malik, Nupur Prakash and S.A.M. Rizvi (2010)," Ontology and Web Usage Mining towards an Intelligent Web focusing web logs", International Conference on Computational Intelligence and Communication Systems International Conference on Computational Intelligence and Communication Networks.

xvii. Neha Goel and C.K. Jha (2013) "Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool", International Journal of Computer Applications , Volume 62– No.2, January.

xviii. Web Log Data (1995)," http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html".

xix. The Internet Traffic Archive (2000), "http://ita.ee.lbl.gov/".

xx. A.C.M. Fong, Zhou, Siu C. Hui,Jie Tang and Guan Y. Hong (2012)," Generation of Personalized Ontology Based on Consumer Emotion and Behavior Analysis" IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 3, NO. 2, APRIL-JUNE.

xxi. Mariangela Vanzin, Karin Becker, and Duncan Dubugras Alcoba Ruiz (2005)," Ontology-Based Filtering Mechanisms for Web Usage Patterns Retrieval", EC-Web,LNCS pp. 267 – 277, Springer.

xxii. Yuefeng Li and Ning Zhong (2007),"Ontology Based Web Mining for Information Gathering" Springer.

xxiii. W3C log file format (2003)," http://www.microsoft.com/technet/prodtechnol/ WindowsServer2003/Library/ IIS/676400bc-8969-4aa7-851a-9319490a9bbb.mspx?mfr=true".