Current Trends in Technology and Science
ISSN: 2279-0535. Volume: 3, Issue: 3 (Apr-May. 2014)

Use of Log Data for Predictive Analytics through Data Mining

Dr. Girish S. Katkar
Assist. Professor, Arts, Sci., Commerce College, Koradi-441111 ,Nagpur, girishkatkar2007@rediffmail.com

Amit Dipchandji Kasliwal
Lecturer, M.S.G. College, Malegaon-423203, Nasik, amitkasliwal16@gmail.com

Abstract — The availability of the data of web accessed is in human readable form generated by computer referred to as web log, provided by online sources, it make that data into day to day life of individual as well as for business operations for more dynamism and bring it closer to real time for the web administrator about what is happening with the web. With the help of such web log data helping the business organization before having to wait a week, or even a month for data through which those people will able to mine data and perform predictive analysis from multiple access made daily as well as in regular manner from users around the world. Data Mining is used for finding expected patterns from that large set of log data using Web Mining. When used together, predictive analytics and data mining can make the future prediction more efficiently with respect to web access.

Keyword — Data Mining, Predictive Analytics, Web Log Data, Web Usage Mining.

1. INTRODUCTION

Data mining is the process of discovering and modeling the hidden patterns in large volumes of raw data. Theapplication in which data mining techniques are used on Web data is referred to as Web data mining. Web data mining can be divided into three different processes: Web content mining, Web structure mining and Web usage mining. There are typically three uses for mining in this fashion.

The first is usage mining, used to complete pattern discovery. This first use is also the most difficult because only bits of information like IP addresses, user information and site clicks are available. With this minimal amount of information available, it is harder to track the user through a site, being that it does not follow the user throughout the pages of the site. The second use is content contenting, consisting of the conversion of Web information like text, images, scripts and others into useful forms. This helps with the clustering and categorization of Web page information based on the titles, specific content and images available. The third is structure mining. This consists of analysis of the structure of each page contained in a Web site. This structure process can prove to be difficult if resulting in a new structure having to be performed for each page.

Web usage mining is a process of extracting useful and meaningful information from web logs viz. server logs, client logs and proxy logs. This process is used for finding out how user uses the Internet and how better the web administrator can admin the web for data availability for the user. Web usage mining techniques can be used to search for patterns in the user behavior when surfing the Web. Web usage mining techniques are useful both to the Web Administrators and to the individual user. When understanding the user's preferences, characterized by the surfing log of users, the web administratorcan improve the site architecturewith respect to the need of users as well as their business objectives. Such that to personalize Web pages, introduce new links, to increase the average time a user spends in the site, page following mechanism on system or on proxy server. By creating links between pages of popular trails we can increase the average time a user spends on a site and even a new product can be given good focus by placing links with same sort of or related links. Web administrator then can identify the pages where users frequently terminate their sessions so that the contents and layout of such pages can be improved. On the other hand, if the browser is prepared up to record the user's navigation history, Web usage mining techniques can be useful to the individual user. In fact, by using such techniques it is possible to view out the user's preferred pages and navigation pattern from accessed log data.

2. DATA PREPARATION

When Web users interactwith a site, data recording their behavior is stored in Web logs, which in a specific sized site can amount to several megabytes per day. These include integrating of various data sources such as server access logs, referrer logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions from usage data, site architecture, and models of user behavior. Since the log data is collected in a raw format it is an ideal target for being analyzed by automated tools. For the research community there has been lots of personal studying Webusage mining techniques totake full advantage of information available in the log files.For the same process here we are using third party application developed in JAVA and WEKA application which are freely available over internet also here we will

representing the web usage mining process by building association rule build in RapidMiner Application which is also build with JAVA. Currently, several commercial log analysis tools are available.However,these tools have essential analysis capabilitiesproducing analytical results such as statistics including summary statistics, visits statistics, access statistics including web pages visits or hit, number of hosts and there hits to website and frequency counts of page visits.

Predictive analytics is the use of intelligence data for forecasting and modeling. It is a way to use predictive analysis data to predict future patterns. It is used widely in the business area such as insurance, medical and credit industries. Now using web log access data the process of predictive analysis can be used to improve architecture of web site so that any individual user can access web site very easily. These can be perform by administrator of that particular website by performing predictive analysis using some tools that works on web mining techniques. Using access data of the past, administrator are able to estimate the likelihood of future events and perhaps can make the availability of data. Data mining aids predictive analysis by providing a record of the past that can be analyzed and used to predict which access pattern is most likely to access later, can be followed later and services.Proper data mining techniques, algorithms and predictive modeling can cover the hidden pattern about website access and will allow tailoring ads to each online user as he or she navigates particular site.Predictive analytics can aid in choosing modeling methods more efficiently. In the best cases, predictive analytics can reduce the amount of money spent to provide the data by provide or sharing maximum bandwidth of the network.At its most effective,data mining can present data on demographics which may have been previously overlooked.

3. Application for Predictive Analytics

Anapplication developed for analyzing Web log data using Web Mining techniques collects the data from various sources mainly server side log, proxy server log and access log, such as mining for association rules, are invoked. This application for Web Usage mining may have following approaches, Preprocessing Tasks, Data Cleaning, Transaction Identification Discovery Techniques on Web Transactions using Path Analysis, Association Rules, Sequential Patterns, Clustering and Classification. WEKA and WebLog Expert system covers all these task together for predictive analytics by providing GUI so that the beginner can also study and view the data and can performed the analysis on that for future work. For performing web usage mining, the data set we are having the web log access data of www.syncsystem.com collected from KDNuggetes (www.kdnuggets.com) in WEKA as shown in figure given bellow.

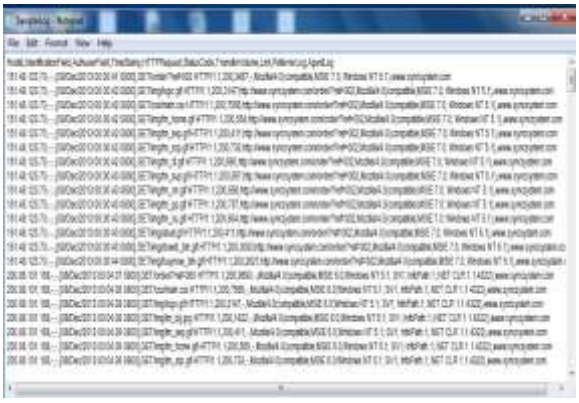

Fig. 1 SampleLog File

The sample web log file has recorded information for each access as:

a) Remote host : Remote hostname or IPaddress number (if DNS hostname is not available or was not provided)
b) RFC931: The remote login name of the user. (If not available a minus sign is typically used)
c) Authuser: The username as which the user has authenticated himself. This is available when using password protected WWW pages.
d) (If not available aminus sign is typically placed)
e) Date: Date and time of the request(Timestamp).
f) Request: The request line exactly as it came from the client. (i.e.,name and the method used to retrieve it, typically GET)
g) Statuscode: The HTTP response code returned to the client. Indicates whether or not the successfully retrieved, and if not, what error message was returned.
h) Bytes: The number of bytes transferred.
i) Referer: The url, the client was on before requesting your url. (If it could not be determined a minus sign will be placed)
j) User agent: The software the client claims to be using. (If it could not be determined a minus sign will be placed)

The WEKA is an academic based application, very popular provides suite of machine learning for analytics written in JAVA, developed at the University ofWaikato, New Zealand. The WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. It provides many different algorithms for data mining and machine learning. WEKA is open source and freely available and platform-independent. Figure 2 Shows the WEKA GUI Chooser.

Current Trends in Technology and Science
ISSN: 2279-0535. Volume: 3, Issue: 3 (Apr-May. 2014)

Fig. 2 WEKA GUI

This Java-based version is used in many different application areas, in particular for analytical and educational purposes and research. The WEKA GUI consists of four options as. Explorer: An environment for exploring data with WEKA.

a) Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
b) Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag and drop
c) interface. One advantage is that it supports incremental learning.
d) Simple CLI: Provides a simple commandline interface that allows direct execution of WEKA commands for operating System systems that do not provide their own command line interface.

To perform preprocess we followed the process as to Run WEKA application, launch the explorer window and select the Preprocess tab. Then Open the data-set and enter what information do you have about the data set e.g. number of instances, attributes, classes? What type of attributes does this data-set contain (nominal or numeric)? What are the classes in this data-set? Which attribute has the greatest standard deviation? What does this tell you about that attribute? After entered the data set under ―Filter‖ choose the Standardize filter and apply it to all attributes. To understanding the data and now apply the Normalize filter and apply it to all the attributes. What does it do? Howdoes it affect the attributes' statistics? How does it differ from Standardize? At the bottom right of the window there should be a graph which visualizes the data-set, making sure Class: class (Nom) is selected in the drop-down box click Visualize All. Under Filter choose the AttributeSelection‖ filter. The attributes it selects the same as the ones we chose as above. How does its behavior change as we alter its parameters? Data cleaning is a process of removing irrelevant or unwanted items from server or proxy log which doesn't have any importance for any type of Web log analysis.

After performing preprocess on log file we got the following statistics table

Summary	
Hits	
Total Hits	30,474
Visitor Hits	29,191
Spider Hits	1,283
Average Hits per Day	3,809
Average Hits per Visitor	8.18
Cached Requests	3,979
Failed Requests	239
Page Views	
Total Page Views	4,436
Average Page Views per Day	554
Average Page Views per Visitor	1.24
Visitors	
Total Visitors	3,577
Average Visitors per Day	447
Total Unique IPs	3,039
Bandwidth	
Total Bandwidth	567.49 MB
Visitor Bandwidth	549.91 MB
Spider Bandwidth	18.67 MB
Average Bandwidth per Day	70.94 MB
Average Bandwidth per Hit	19.07 KB
Average Bandwidth per Visitor	157.11 KB

Fig.3 Statistical Summery

This summery is prepared with WebLog Expert application. It is a fast and powerful web access log analyzer tool used for predictive analytics. It also generate information about site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. It helps to produce easy to read reports including text information and charts.

The figure 4 shows the number of host served by the web site for that particular time period as mentioned in web log file in timestamp field.

Hosts				
	Host	Hits	Visitors	Bandwidth (KB)
1	213.186.34.25	78	39	24
2	83.246.159.200	20	20	142
3	78.110.215.109	20	20	84
4	134.174.21.2	26	19	109
5	83.246.159.197	17	17	120
6	70.87.179.90	16	14	187
7	82.146.53.28	12	12	6
8	62.2.98.66	12	12	171
9	83.246.159.196	12	12	85
10	64.233.166.136	11	11	133
11	87.98.216.168	10	10	3
12	69.9.31.40	9	9	37
13	81.57.109.132	229	9	150
14	58.5.250.41	11	9	46
15	72.185.164.0	14	9	59
16	71.114.21.3	9	9	37
17	216.197.96.153	29	7	113
18	71.120.25.4	7	7	29
19	12.178.36.25	14	7	22
20	62.246.107.162	14	7	163
21	69.181.74.38	34	7	101
22	203.177.240.23	7	7	85
23	64.92.172.180	7	7	81
24	74.53.46.178	7	7	4
25	83.246.159.198	8	7	57
26	76.108.231.245	6	6	25
27	121.9.209.17	25	6	955
28	70.84.183.105	6	6	3
29	65.214.44.29	55	6	126
30	71.191.243.123	6	6	25
31	193.131.2.148	7	6	11
32	213.138.192.220	6	6	42
33	80.80.111.129	60	6	988
34	69.98.31.144	6	6	95
35	87.99.79.90	6	6	59
36	87.99.79.93	6	6	59
37	87.99.79.94	6	6	59
38	87.99.79.91	6	6	59
39	87.99.79.92	6	6	59
40	8.10.179.182	12	6	4
41	200.204.77.46	192	6	4,526
42	222.126.107.93	74	6	116
43	38.113.234.181	6	6	1
44	76.90.74.223	6	6	21
45	58.142.116.135	6	6	36
46	209.128.93.229	5	5	1
47	169.132.18.248	10	5	19
48	170.149.100.10	6	5	25
49	202.163.208.50	40	5	53
50	62.112.192.11	5	5	74
	Subtotal	1,199	437	9,500
	Total	29,191	3,577	561,980

Fig. 4 Host Summery

Current Trends in Technology and Science
ISSN: 2279-0535. Volume: 3, Issue: 3 (Apr-May. 2014)

WebLog Expert can analyze logs of Apache and IIS web servers. It can even read GZ and ZIP compressed log files. The WebLog Expert used to view intelligent data such as number of host visiting the site and the pages or links that users are accessing as shown in figure 4, it is clear that what kind of hosts are accessing that web site and how much of data had been used for making that data available to each and every user. From the summery, if we point out towards number of access from various host then there are total 29191 hits are done from 50 host and for that access total 561980 Kb of data had been used. Now for making prediction over summery the web administrator can control the data size by providing same access pattern for the same kind of host and thus helping the network bandwidth always available to each and every user.

The use of WebLog Expert in predictive analysis is that from preprocessing the web log data we can find the patterns having the access to which page from which page. The figure 5 shows the access to the different pages from different or same host.

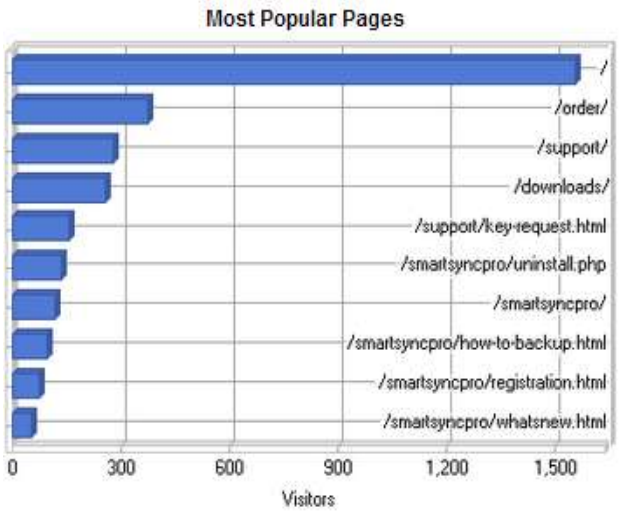

Fig. 5 Most popular pages

After viewing the figure 5 we can say how many visits are done on which page and this is what the administrator can lookout for and set the patterns depending upon the visits made for particular page.

For using web usage mining for predictive analysis, the association rule technique can be used to define the rule for web pages so that by making the pattern available to visitor, the access to website can be increased. For preparing the association rule on collected data from web log here we are using WEKA's Knowledge Flow option. The figure 6 shows screen of designed association rule from knowledge flow option from WEKA.

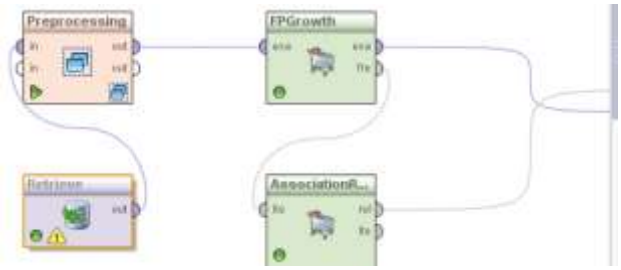

Fig. 6 Knowledge Flow of WEKA

By using these design, we got the result for our sample size that we got after performing preprocessing in graph visualizer of WEKA as shown in figure 7.

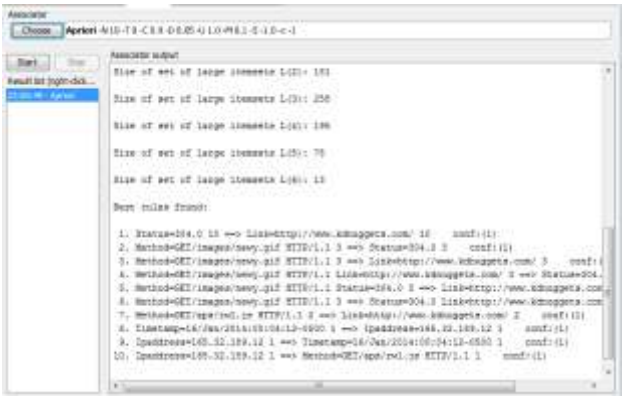

Fig. 7 Association Rule

As we can see in the above figure, after preprocessing on out of 30474 records, only 439 records are considered as in preprocessing the error log, referral log, hits to images are removed. After performing preprocessing, in the web mining the next process is visualization of expected pattern and it is done with the help of WebLog Expert Application. The WebLog Expert provides the actual data related to website by providing web log data as an input. The challenge is when it comes to doing predictive analytics with web data using web mining is to realize that web data for the most part is completely anonymous, usually incomplete and sometimes unstructured. When we want to do traditional data mining (and not just analysis) and predictive analytics all of these things are tedious may take lots of time to perform.

Here is another thing that lots of researcher fails to cover about to find the which are top referring site are there or we say provided by web administrator and how many hits are done to which referral site. In the figure 8, the chart shows the top referral site and the number of visitors accessing that site. It is easier to Mine and then predict when there is a certain amount of expected existence. So again if we think about how searching is performed or handle by web administrator, then we can say that there are some top search engines helps to improve this ability of website and hence also making the website and the data on their website to be available for everyone.

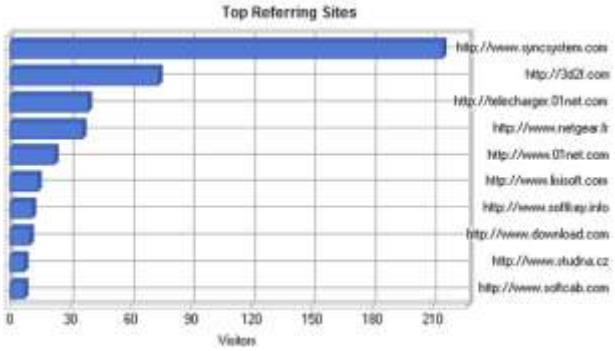

Fig. 8 Number of Visitor to Top Site

Web usage mining is valuable not only for researchersor businesses using online process, but also to electronic businesses whose business is based solely on the traffic provided through search engines. The use of this type of web mining helps to gather the important information from customers visiting the site. Figure 9 shows the how many user visits the search engine from www.syncsystem.com site and how web admin view that access data.

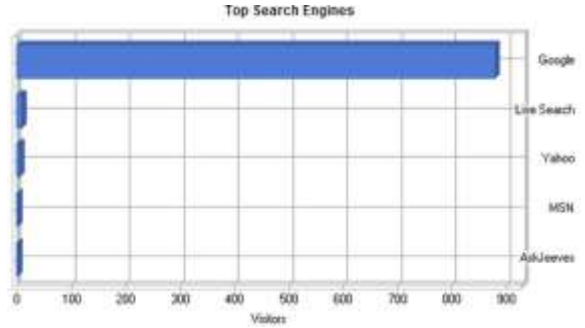

Fig. 9 TopSearch Engine

The Primary Purposes of predictive analysis considered, this issue becomescomplicated. We are trying to predict the outcome of website, a complex being that exists to do lots (even things that your website was not created for). For predictive analytics, it enables an in-depth log to complete analysis of website's productivity flow. In web based businesses, this information organization to the most effective Web server for promotion of their product or service.

The web mining also enables Web based businesses to provide the best access routes to services or other advertisements. When a company advertises for services provided by other companies, the usage mining data allows for the most effective access paths to these portals. This can be possible only by performing mining on the access data to website collected in log file.

Analysis of this usage data will provide the companies with the information needed to provide an effective presence to their customers. This collection of information may include user registration, access logs and information leading to better Web site structure, proving to be most valuable to company online marketing. Again the data that we neglect during analytics is the errors. Most of the time while modifying structure of website, user find them on the error page. For that part using, WebLog Expert we covered all errors separated with respect to their code and the numbers hits faced the error while they accessed the website.

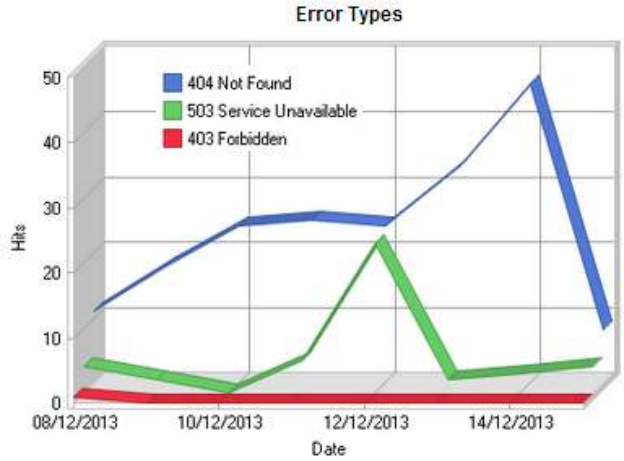

Fig. 10 Error Types

Internally, for predictive analytics web usage mining effectively provides information to improvement of communication via web communications. Developing strategies through this type of mining will allow for online based company databases to be more effective through the provision of easier access paths and hence improving future trends or setting future trends.

Therefore, it is easily determined that usage mining has valuable uses to the predictive analytics for businesses and a direct impact to the success of their plans, strategies and website access traffic. This information is gathered on a daily basis and continues to be analyzed consistently. Analysis of this valuable information will help companies to develop plans that are more effective, internet accessibility, inter-company communication and structure, and productive skills through predictive analytics when used with web usage mining.

4. CONCLUSION

The information obtained with our experiments show the effectiveness of the web usage mining application for predictive analytics in the businesses, not only in reducing considerably the size of Web log files but also in grouping Web requests into a number of user which can encode the user browsing behavior in a significant manner. After having information about access the web site architecture can be improved thus helping the company to take future decision by doing predictions. It all can solve the queries like How to describe the preferences of users on the basis of their navigational access. Several measures and/or heuristics can be applied to obtain the degree of interest for a Web resource. A possibility is to consider the degree of interest for predictive analysis to a resource as strictly related to the frequency of accesses to that resource (number of accesses to that resource / total number of accesses during the session) and to the time the user spends on the same one. Analysis of this valuable information will help companies to develop plans that are more effective and

impactful, easy access to website, in between company communication and productive skills through predictive analytics when used with web usage mining.

5. REFERENCES

[1] R. Srikant and R. Agrawal., "Mining generalized association rules.", 21st VLDB Conf., 1995.
[2] Pitkow, J. and Bharat, K., "WebViz: A Tool for World Wide Web Access Log Analysis", Proceedings of Int. World-Wide Web Conference. pp. 271-277, 1994.
[3] Pierrakos D., "Web usage mining as a tool for personalization: a survey.", User Modeling and User-Adapted Interaction, Vol. 13 Issue(4), pp. 311-372, 2003.
[4] G. Castellano, A. M. Fanelli, M. A. Torsello., "Log Data Preparation For Mining Web Usage Patterns", Int. Conf.on Applied Computing, 2007.
[5] RanieriBaraglia and FabrizioSilvestri, "An OnlineRecommender System for LargeWeb Sites", Web Intelligence, Proceedings of IEEE/WIC/ACM Int. Conf., Sept. 2004.
[6] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. InkeriVerkamo., "Fast discovery of association rules.", Advances in Knowledge Discovery andData Mining, AAAI Press,pp307–328, CA, 1996.
[7] R. Cooley, B. Mobasher, and J. Srivastava. "Data preparation for mining world wide webbrowsing pattern.",Knowledge and Information Systems, Vol. 1Isusue(1), 1999.
[8] Rana, Chhavi, "A Study of Web Usage Mining Research Tools", Int. Journal of Advanced Networking &Applications, Vol. 3 Issue 6, p1422, Jun2012.
[9] UjwalPatil, SachinPardeshi, "A Survey on User Future Request Prediction: Web Usage Mining", Int. Journal of Emerging Technology and Advanced Engineering,Vol.2, Issue 3, March 2012.
[10] SawanBhawsar, KshitijPathak, SourabhMariya, Sunil Parihar, "Extraction of Business Rules from Web logs to Improve Web Usage Mining", Int. Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 8, August 2012.
[11] Agrawal, R., Srikant, R., "Fast algorithms for mining association rules in largedatabases.",20th Int. Conf. on Very Large Data Bases, pp487-499,2004.
[12] Witten, I.H., Frank, E., "Data Mining: Practical machine learning tools and techniques.", 2Edn. Morgan Kaufmann, San Francisco (2005).
[13] Pohle, C., Spiliopoulou, M., "Building and exploiting ad hoc concept hierarchies forweb log analysis.",Data Warehousing and Knowledge Discovery, Proceedingsof Int. Conf., DaWaK 2002.
[14] N. Labroche, M. J. Lesot, L. Yaffi, "A New Web Usage Mining and Visualization Tool", Int. Conf. on Tools with Artificial Intelligence,Vol.1,pp321-328, 2007.
[15] Y. Tao, T. Hong, and Y. Su, "Web usage mining with intentional browsing data", Expert Systems with Applications, Vol. 34, Issue 3,pp1893-1904,2008.
[16] www.passionned.com
[17] www.weblogexpert.com
[18] www.waikitouni.edu
[19] www.kduggets.com
[20] http://analytics.google.com/dashboard