COMPARISON ANALYSIS OF WEB USAGE MINING USING PATTERN RECOGNITION TECHNIQUES

Nanhay Singh[1], Achin Jain[1], Ram Shringar Raw[1]

Ambedkar Institute of Advanced communication Technologies & Research
Delhi, India
nsingh1973@gmail.com, achin_jain25@yahoo.com, rsrao08@yahoo.in

ABSTRACT

Web usage mining is the application of data mining techniques to better serve the needs of web-based applications on the web site. In this paper, we analyze the web usage mining by applying the pattern recognition techniques on web log data. Pattern recognition is defined as the act of taking in raw data and making an action based on the 'category' of the pattern. Web usage mining is divided into three parts- Preprocessing, Pattern discovery and Pattern analysis. Further, this paper intended with experimental work in which web log data is used. We have taken the web log data from the "NASA" web server which is analyzed with "Web Log Explorer". Web Log Explorer is a web usage mining tool which plays the vital role to carry out this work.

KEYWORDS

Web Usage Mining, Log Data, Pattern Recognition, Web log explorer, Data preprocessing

1. INTRODUCTION

Information on Internet is increasing rapidly day by day, and it is applied for different purposes in decision support system. Web mining is the application of data mining and is used for the same task. Web mining is classified into three categories: Web Usage Mining (WUM), Web Content Mining (WCM), and Web Structure Mining (WSM). Among these, WUM is applied on usage data and it is being used at large scale by the organizations to study the behaviour of their web users. In WUM the user's web log is collected for inferring the useful information by analyzing it. In present scenario every organization trusts on their websites for the growth of their business. The organizations collect the data from their web server to analyze the behaviour and investigating interest of the users. The ability to track user browsing behaviour down to individual mouse click has brought the vendor and end customer closer than ever before, it is now possible for a vendor to personalize his product message for individual customer [1].

This information is gathered automatically by web server and stored in web logs and access logs. The web logs is use to track the end user behaviour for WUM. Log files are those files that list the actions that have been occurred [2]. Web server creates and maintains log files for the purpose of getting feedback about activity & performance of the server and the problems occurring in the web server [3]. Log files plays very important role in pattern recognition as analyses of log files helps in identifying relationships and patterns between messages or request from the user.

Pattern recognition is the task of finding useful information from web server logs applying various techniques such as filtering, grouping etc. This extracted knowledge plays a very important role in formulation of important rules (decisions) regarding organization website

structure, making marketing and advertising more fruitful and effective. Before the process of pattern recognition the log file data has to go through three stages i.e., preprocessing, pattern discovery and pattern analysis.

This article mainly considers web usage mining which is the process of extracting useful information from server logs i.e. user's history. The rest of the paper is organized as follows: The work related to our work is given in section 2. Section 3 presents problem formulation. In section 4 description on web usage mining using pattern recognition is given in detail using log files which plays an important role in extraction of useful patterns. Experimental result analysis of the problem is carried out in section 5. Finally section 6 concludes the paper.

2. LITERATURE REVIEW

The WUM mainly focuses on the study of user browsing patterns. Yang Bin et al. in [4] used negative association rules in discovery of web visitor's patterns. Negative association rules have been deployed to solve the deficiencies in which positive rules are referred to. In [5] many researchers carried out numerous surveys on web usage mining. In [6] Resul Das and Ibrahim Turkoglu proposed a new approach for preprocessing of the web log data and then association rules are being employed to extract the useful patterns. Kobra Etminani et al. in [7] used concept of Ant based clustering method to preprocess web log data to extract frequent patterns for pattern discovery.

In [8] authors proposed two tier architecture for capturing user intuition in the form of recommendation list containing pages visited by users and pages visited by other users having similar usage profile. In [9] author introduced a novel approach growing neural gas which is a kind of neural network, in the process of web usage mining to detect user's patterns. In [10] Juan Julian Merelo Guervos et al. proposed an automatic extraction of association rules from the results of surveys to recommend readers about other about related topic weblogs.

3. PROBLEM FORMULATION

In today's world information on Internet is increasing day by day and web administrator's continuously trying to make their website more users friendly and efficient. Pattern extracted from web server log helps them in a big way to make decision about restructuring of websites and implementation of new applications which will increase their traffic and eventually business. In this paper the problem defined is the extraction of patterns from web server log file. I think it is an excellent way to define the usage mining using pattern recognition techniques.

In this paper, our main aim is to carry out experimental work on web log data collected from NASA web server to find out useful browsing patterns. Results extracted from this work will play an important role in improving the web server performance. There are a number of web usage mining tools available in the market but here Web Log Explorer (WLE) tool is used for the implementation of our work [11]. WLE is used to determine the number of accesses to the server and to individual files, the times of visits and the domain names, and URLs of users. The input of this tool is web log files collected from the web servers. In this work, we have also carried out comparative analysis of JPG and GIF image file types using results generated through MATLAB. In comparative work, we analyzed the effect of image file types for bandwidth usage per hit as parameter.

4. WEB USAGE MINING WITH PATTERN RECOGNITION

Web usage mining is the process to automatic discovery of user access patterns from web servers. On the daily operation several organizations collect large volumes of data, generated automatically by web servers and collected in server access logs. Web usage mining focuses on

techniques that predict user behaviour while the user interacts with the web. Other sources of user information include referred logs which contain information about the referring pages for each page references and user registration or survey data.

To analyze such data can help the organizations to determine the value of customers, marketing strategies across products and effectiveness of promotional campaigns. Web usage mining can also provide information on how to restructure and modify a web site to create a more presentable, easy to access and shed light on management of workgroup communication and organizational infrastructure. Web log files play an important role in recognition of web usage browsing patterns. Web log data contains various parameters related to web server activity which are analyzed to extract useful information. Log files are explained in detail in next section.

4.1 Log Files

In this paper, log file data of NASA web server is used to extract useful patterns. Before the application of pattern recognition, initially log file data is being preprocessed to remove any unwanted entries so that the patterns extracted are useful and relevant. Log files can be classified into three categories depending on the location of their storage.

- **Web Server Log Files:** These log files resides in web server and notes activity of the user browsing website. There are four types of web server logs i.e., transfer logs, agent logs, error logs and referrer logs.
- **Web Proxy Server Log Files:** These log files contains information about the proxy server from which user request came to the web server.
- **Client browser Log Files:** These log files resides in client's browser and to store them special software are used.

4.2 Log Files Parameters

Log files contain various parameters which are very useful in recognizing user browsing patterns. Below is the list of some of the parameters.

- **User Name:** Identifies the user who has visited the website and this identification normally is IP address.
- **Visiting Path:** It is the path taken by the user while visiting the website.
- **Path Traversed:** It is the path taken by the user within the website.
- **Time Stamp:** It is the time spent by user on each page and is normally known as session.
- **Page Last Visited:** It is the page last visited by the user while leaving the website.
- **Success Rate:** It is measured by downloads and copying activity carried out on the website.
- **User Agent:** It is the browser that user uses to send the request to the server.
- **URL:** It is the resource that is accessed by the user and it may be of any format like HTML, CGI etc.
- **Request Type:** It is the method that is used by the user to send the request to the server and it can be either GET or POST method.

4.3 Types of Log File Format

There are mainly three types of log file formats that are used by majority of the servers.

- **Common Log File Format:** It is the standardized text file format that is used by most of the web servers to generate the log files. The configuration of common log file format is given below in the box.

"%h %l %u %t \"%r\" %>s %b" common CustomLog logs/access_log common
eg: *127.0.0.1 RFC 1413 frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326*

- **Combined Log Format:** It is same as the common log file format but with three additional fields i.e., referral field, the user_agent field, and the cookie field. The configuration of combined log format is given below in the box.

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-agent}i\"" combined CustomLog log/access_log combined
eg**:** *127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"*

- **Multiple Access Logs:** It is the combination of common log format and combined log file format but in this format multiple directories can be created for access logs. Configuration of multiple access logs is given below in the box.

LogFormat "%h %l %u %t \"%r\" %>s %b" common
CustomLog logs/access_log common
CustomLog logs/referer_log "%{Referer}i -> %U"
CustomLog logs/agent_log "%{User-agent}i"

5. EXPERIMENTAL RESULT ANALYSIS

Pattern recognition is defined as the act of taking in raw data and making an action based on the "category" of the pattern [12] or it can be defined as the process for observing patterns of interest (e.g. most used file type) from entire data (e.g. web server log data). Pattern recognition can also be used to make important decisions about the patterns. Finding or recognizing patterns from web logs requires the log data to go through three stages i.e., preprocessing, pattern discovery and pattern analysis to investigate the pattern of the file types accessed by the users during the browsing of the NASA website. The results can be analyzed in terms of the following browsing patterns. Details of the user access log files that are used to analyze the user browsing patterns are shown in the table 1.

Table 1. Details of Log File used

Log File Details	
Log File Name	NASA-HTTP Logs
Log Duration and Data Range	Two months - Jul 01 to Jul 31(1995)
Size of the Log File	20.7 MB gzip compressed, 205.2 MB uncompressed
URL	ftp://ita.ee.lbl.gov/traces/NASA_access_log_Jul95.gz

Since the size of considered data is very huge and it is very difficult to represent in this paper, therefore we have taken very few data from the NASA database. The sample of NASA web log data is presented in the fig.1.

The stages used to carry out the experimental work on log file from NASA web server using WLE tool to recognize important and useful patterns are explained as follows.

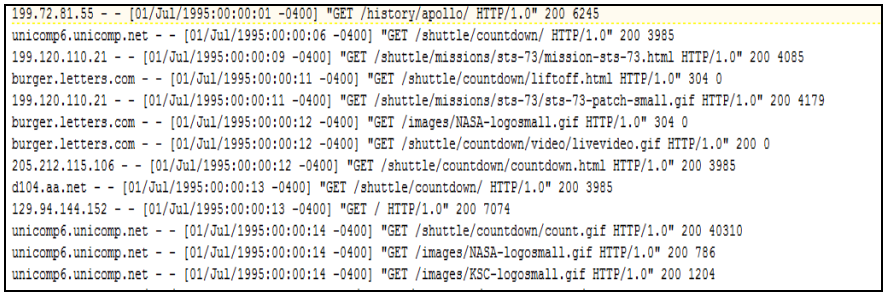

Figure 1. Sample of NASA Web Log Data

5.1 Preprocessing

This step is necessary to restore user's activities in the web server log in readable and consistent way. There are three types of preprocessing techniques: usage preprocessing, content preprocessing and structure preprocessing. In this work, preprocessing of log data is applied to remove entries having status (response code) other than 200.

- Total entries present in log file = 1891697
- After removing entries which are not having status code 200 = 1701534
- Entries filtered out = 190163

Summary for general profiles and activities of the users are listed in Table 2.

Table 2. Summary for General Profile: Activity Statistics

Summary of Activity	
Unique IP	80018
Visitors	80018
Hits	1559519
Bandwidth	23.68 GB
Pages/Files	4050
Countries	83
Entry Points	774
Page Views	562179
Average Page Views per Visitor	7.03

5.2.Pattern Discovery

Pattern Discovery is used to extract patterns of usage from web data [13]. This method uses data mining techniques and algorithms to find out useful information. Knowledge extracted can be represented in many ways such as graphs, charts, tables, forms etc. Techniques used in pattern discovery are given as follows:

5.2.1. Converting IP Address to domain name

This Technique is useful in discovering important facts about the visitor such as Country of Visitor by looking domain name extension such as ".in" domain specifies that the user is from India. Finding such information about visitors helps in customizing website according to visitor's interest. In this study the log file is analyzed and results extracted are as follows.

- IP address (199.72.81.55) map to user from USA
- IP address (205.189.154.54) maps to user from Canada

5.2.2. Grouping

This technique is used to extract high level in formation by grouping similar information about the visitor's usage pattern.For e.g. grouping all visitors receiving Unsuccessful Response Code (URC) form server shows how many users are not able to get their request fulfilled. This kind of information helps in website performance as webmaster is able to filter out user's requesting non available web pages by looking out response code. In this paper, log file data is grouped according to the response code sent by the server. For grouping of the visitors based on the response code, log file used in not preprocessed. See Fig. 2 for the grouping of visitors based on response code.

- Total Number of Visitors with response code '200 – OK' = 80018
- Total Number of Visitors with response code '304 - Not Modified' = 11229
- Total Number of Visitors with response code '404 - Not Found' = 3487
- Total Number of Visitors with response code '403 – Forbidden' = 22
- Total Number of Visitors with response code '501 - Not Implemented' = 8
- Total Number of Visitors with response code '400 - Bad Request' = 1

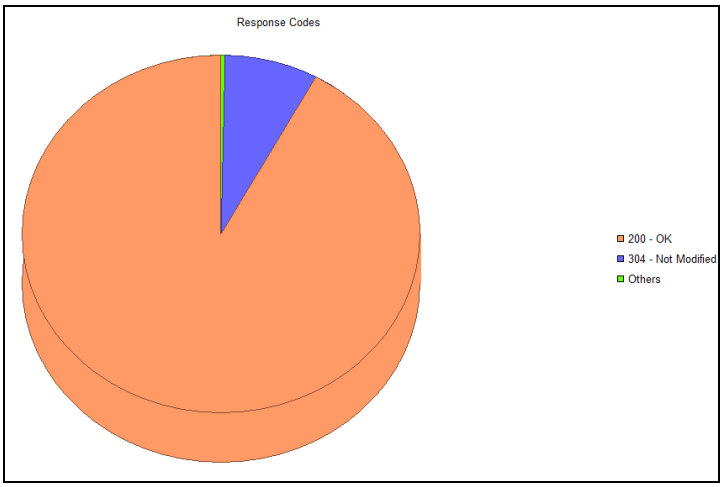

Figure 2. Grouping of Visitors based on Response Code

5.2.3. Filtering

Filtering allows web administrator to answer specific questions about web such as "How many visitors came from India this week referred from Google" or "Which file type is most accessed by the visitors". The usage of most file types is shown in fig. 3. From the analysis of fig. 3 it is found that most file used on NASA web server is the image file and above all file types '.gif' files are accessed most. The usage of .gif files by day of week is shown in fig. 4 which shows that on Thursday image files with extension '.gif' are accessed mostly. The usage of .gif files by hour of day is shown in fig. 5 which helps in recognizing very important pattern about the usage of image (.gif) files which is these files are mostly accessed at hour 12 of the day.

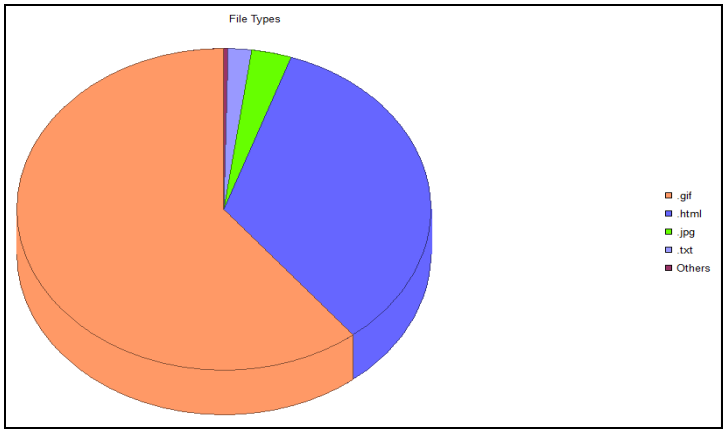

Figure 3. Top File Types Accessed

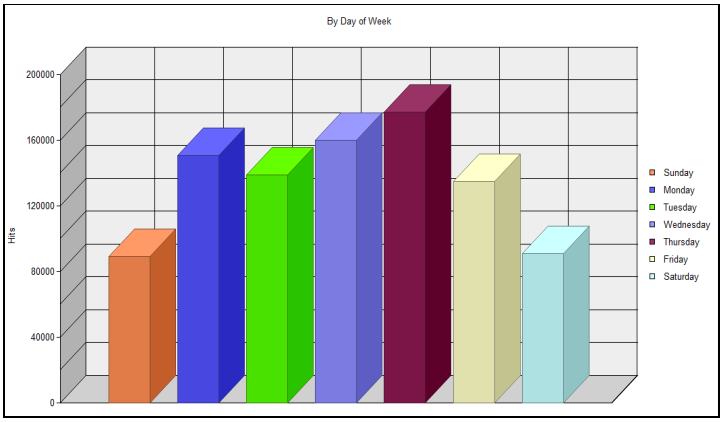

Figure 4. By Day of Week – Usage of .gif Files

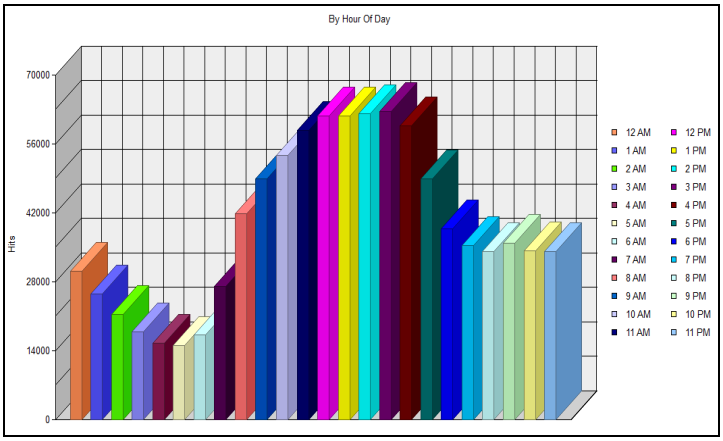

Figure 5. By Hour of Day – Usage of .gif File

Analysis of web server log file from NASA server resulted in recognition of various patterns. Technique "Converting IP address to domain name" helps in identification of visitor from the country they are sending request to the web server. Pattern recognized from grouping of visitors

based on response code is helpful in identifying the visitors causing unnecessary traffic by requesting the web pages that are not available.

7. COMPARISON OF BANDWIDTH/HIT USAGE FOR IMAGE FILES

In this section, we have carried out the comparative analysis between bandwidth usages per hits for the two different types of image files. As we have seen in the earlier section, image files are mostly used among all the files. To analyze the effect of different types of image files we have done the comparison between the bandwidth usage per hit for GIF and JPG Files.

Table 3.Bandwidth/Hits for "JPG" Files

Session(Days)	Bandwidth/Hits (KB) for "JPG" Files
1	63.55006061
2	64.58440011
3	81.38309762
4	95.33802687
5	95.61975889
6	91.79453199
7	82.8913834
8	88.35584634

For the comparison session of 3 days is considered in this work and in the first step, we have done the analysis for JPG files. Table 3 shows the Bandwidth/Hits (KB) for different sessions and Figure 6 gives graphical view of the data.

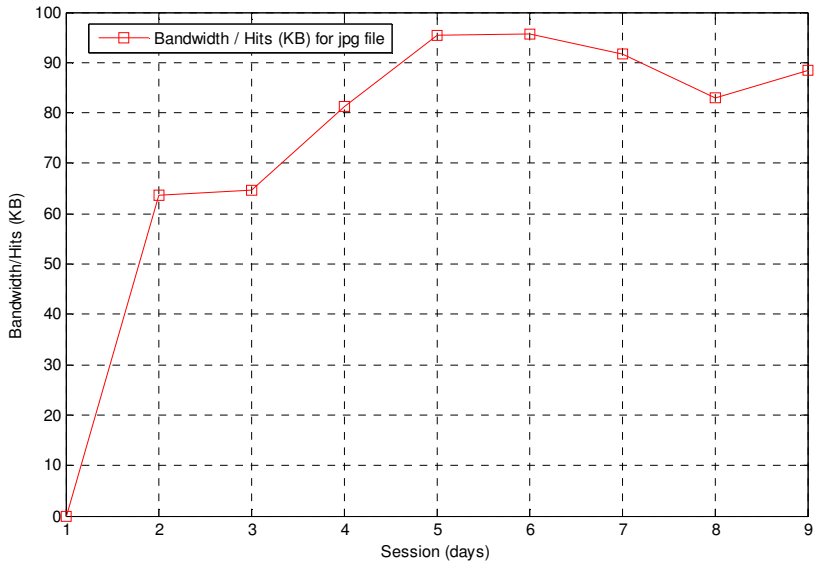

Figure 7. Bandwidth/Hits (KB) for JPG Files

In the second step, we have carry out the analysis for "GIF" image files and the files details taken from NASA web server log data has shown in table 4. Figure 7 shows the same details through graphical view.

Table 4. Bandwidth/Hits for "GIF" Files

Session(Days)	Bandwidth/Hits (KB) for "GIF" Files
1	16.03948522
2	13.29840249
3	12.18141264
4	13.72603617
5	11.60833916
6	11.61766552
7	10.80628987
8	9.754800298

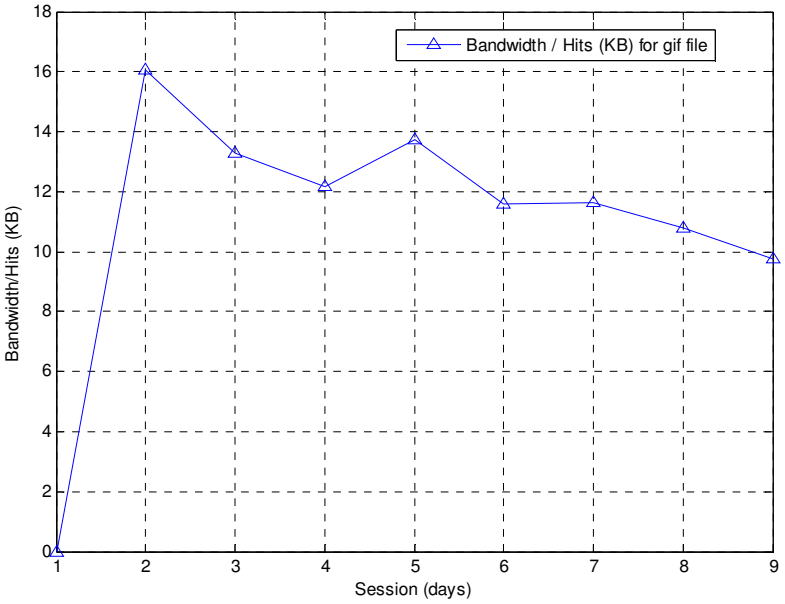

Figure 8. Bandwidth/Hits (KB) for GIF Files

For the comparative analysis, we have analyzed both the image types together and the result shown in figure 9. In the figure, it is clear that impact of GIF files is less on the server than JPG files. Bandwidth is the comparison parameter used for every hit on particular types of image file.

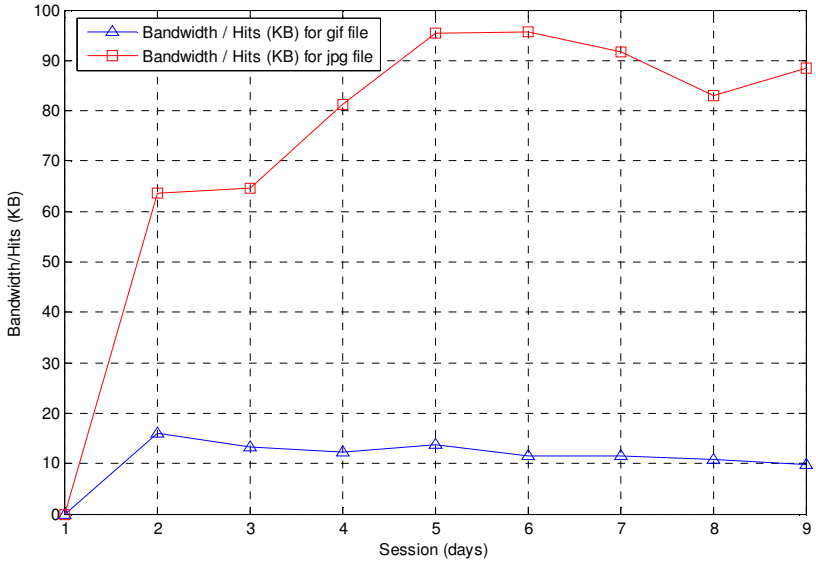

Figure 9. Comparison between Bandwidth/Hits (KB) for JPG and GIF Files

After analysis of Bandwidth/Hits for both GIF and JPG image files, results shows that the use of JPG files increase the bandwidth use of the server which is not desirable. After analysis of the image file types, it is clear that impact of using JPG files on the bandwidth of the server is much severe than the GIF files. To use the server bandwidth efficient, care should be taken to use the GIF files than JPG files.

8. CONCLUSION

In this paper, we study the web usage mining with pattern recognition techniques and carried out experimental work on web log data collected from NASA web server to find out useful browsing patterns Pattern recognition can be used by the web administrator to optimize web site performance by blocking such requests. From the patterns extracted from filtering technique it is clear that mostly used file type by NASA website visitors is image file with extension ".gif" and on Thursday at hour 12. From the comparison between JPG and GIF image files it is clear that if the web administrator uses GIF files for the image media than bandwidth of the server will be saved. This extracted usage pattern will help administrators managing the website resources in better way. The results or findings from this study are surely useful for web administrator in order to improve web site performance through the improvement contents, structure, presentation and delivery.

REFERENCES

[1] J. Srivastava et al., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," ACM SIGKDD Explorations, vol. 1, no. 2, 2000, pp. 12-23.

[2] L.K Joshila Grace, V. Maheswari, Dhinaharan Nagamalai, "Analysis of Weblogs and Web User in Web Mining," International Journal of Network Security & Its Applications (IJNSA), Vol. 3, No. 1, January 2011.

[3] Rajni Pamnani, Pramila Chawan: "Web Usage Mining: A Research Area in Web Mining" Department of computer technology, VJTI University, Mumbai.

[4] Yang Bin, Dong Xianguin, Shi Fufu, "Research of Web Usage Mining based on Negative Association Rules" International Forum on Computer Science-Technology and Applications, 2009.

[5] M Eirinaki and M Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3, no. 1, 2003, pp. 1-27.

[6] Resul Das, Ibrahim TURKOGLU, "Extraction of Interesting Patterns Through Association Rule Mining for Improvement of Website Usability," Journal of Electrical & Electronics Engineering, Vol. 9, No. 2, 2009.
[7] Kobra Etminani, Mohammad-R. Akbarzadeh-T., Noorali Raeeji Yanehsari, "Web Usage Mining: user's navigational patterns extraction from web logs using Ant-based Clustering Method," IFSA-EUSFLAT, 2009.
[8] Ms. Dipa Dixit, Mr. Jayant Gadge, "Automatic Recommendation for Online Users Using Web Usage Mining," IJMIT, Vol. 2, No. 3, August 2010.
[9] Anshuman Sharma, "Web Usage Mining Using Neural Network," International Journal of Reviews in Computing (IJRIC), Vol. 9, 2012.
[10] Juan Julian Merelo Guervos et al., "Weblog Recommendation Using Association Rules," IADIS International Conference on Web Based Communities 2006.
[11] Web Log Explorer Tool - http://www.exacttrend.com/WebLogExplorer/
[12] Richard O. Duda, Peter E. Hart, David G. Stork 2001 Pattern classification (2nd edition), Wiley, New York, ISBN 0-471-05669-3.
[13] Resul Das, Ibrahim TURKOGLU, Mustafa POYRAZ, "Analyzing of System Errors for Increasing a Web Server Performance by using Web Usage Mining," Journal of Electrical & Electronics Engineering, Vol. 7, No. 2, 2007

Authors

Dr. Nanhay Singh, working as Associate Professor in Ambedkar Institute of Advanced Communication Technologies & Research, Govt. of NCT, Delhi-110031 (Affiliated to Guru Gobind Singh Indraprastha University, Delhi) in the Department of Computer Science & Engineering. He received his Ph.D (Computer Science and Technology) & M.Tech. (Computer Science & Engineering) from the Kurukshetra, University, Kurukshetra, Haryana. He has rich experience in teaching the classes of Graduate and Post-Graduate in India. He has contributed to numerous International journal & conference publications in various areas of Computer Science. He published more than 15 Research Paper in International Journals and Conferences. He has also written an International book Titled as "Electrical Load Forecasting Using Artificial Neural Networks and Genetic Algorithm", in Global Research Publications New Delhi (India). His area of interest includes Distributed System, Parallel Computing, Information Theory & Coding, and Cyber Law, Computer Organization.

Achin Jain, received B.Tech. Degree in Computer Science & Engineering from the G.G.S.I.P. University. He has Industrial Work experience of more than 2 years and Pursuing M.Tech. Degree in Information Security from Ambedkar Institute of Advanced Communication Technologies & Research, Delhi, India. He published more than 5 Research Papers in International Journals and Conferences. His area of interest includes Web Usage Mining, Web Attacks.

Dr. R. S. Raw, received his Ph.D (Computer Science and Technology) from the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India in 2011. He has obtained his B. E. (Computer Science and Engineering) from G. B. Pant Engineering College (HNB Srinagar-Garhwal Central University), Pauri-Garhwal, UK, India and M. Tech (Information Technology) from Sam Higginbottom Institute of Agriculture, Technology, and Sciences, Allahabad (UP), India in 2000 and 2005, respectively. He has worked as an Assistant Professor at Computer Science and Engineering Department, G. B. Pant Engineering College, Uttarakhand Technical University, since March 2001 to June 2003 and March 2005 to July 2011. Currently he is working as an Assistant Professor in the department of Computer Science and Engineering of Ambedkar Institute of Advanced Communication Technologies & Research, Delhi, India since August, 2011. Dr. Raw has published research papers in International Journals and Conferences including IEEE, Elsevier, Springer, Inderscience, American Institute of Physics, IERI Communications Letters, AIRCC, etc. His current research area are Mobile Ad Hoc Network and Vehicular Ad Hoc Network.