International Journal of Advanced Science and Technology
Vol.105 (2017), pp.25-36
http://dx.doi.org/10.14257/ijast.2017.105.03

Web Data Mining and Analysis: An Intelligent Perspective

Shakti Kundu and Madan Lal Garg

Department of Computer Science & Engineering,
DIT University, Dehradun - 248009, Uttarakhand, India
shaktikundu@gmail.com, mlgarg2000@yahoo.com

Abstract

Due to extensive growth of the online information, websites are providing the relevant information to the users which are one of the significant tasks. The essential aspect to discover hidden and interesting information is web data mining architecture. When quantity of web information is in growing phase then web data mining becomes challenging. The web data analysis will remove the non interesting rules or patterns that were generated. They head to extract the interesting rules or patterns from the outcome of pattern discovery and pattern analysis process. An intelligent perspective of web data mining and analysis has been discussed in this paper to face the challenge of huge web data.

Keywords: *Analysis, Data, Intelligent, Mining, Web*

1. Introduction and Motivation for Research

Web data is mined with various web log analyzer tool into the documents on the World Wide Web. Web mining is the procedure which helps in extracting the interesting information from the World Wide Web via accepting the standards of the data mining and incorporating it into the features of website.

Transaction log analysis is a broad category of methods used for macro and micro analysis of transaction logs. It refers to electronic records as an interface that happened between system and users. Transaction Log Analysis consist of study of web based system logs, web blogs and search engine logs [1] which has been highlighted in Figure 1.

Web Log Analysis: Active running website is a technique to track and weight the visitor's traffic and performance. Essential task of this technique is to study the visitor's performance with the probable performance.

Blog Analysis: A blog (short for Web log) is part of the network of social media designed and popularized by participants to exchange information and expressed opinions.

Search Log Analysis: The web data stored in search engines and websites provides important insights into understanding the information searching tactics of online content. This activity can help system designers and interface developers in gaining information [1].

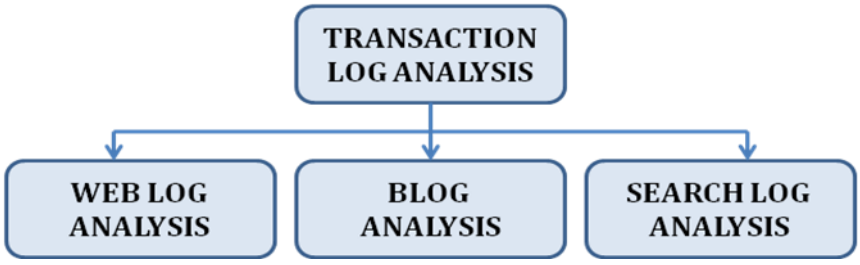

Figure 1. Methods of Transaction Log Analysis

ISSN: 2005-4238 IJAST

The globe of Web has become the nature where the masses lead their day to day digital lives. Web users in each place are bounded by framework of machines, networks and activities. Web log expert through the help of data analytical tool collect the facts or information from a variety of web log data via subsequent analysis. Researchers also gain fruitful information from aggressively growing source of personal communication blogs [2, 4].

Web-traffic measurement is the analysis of data between client and server computers. It provides insight into how people use computers and is commonly used in research. The two dominant forms of Web traffic measurement is log file analysis and ASP-based tools which are generally used in practice.

Mobasher *et al*. (1999) suggested the web personalization system based on the direct clustering of URLs instead of clustering the user sessions. This research comprises of two phases: offline and online. The usage data were mined via offline approach. The web page customization based on the knowledge discovery was done via online process. To provide automatic filtering capabilities to capture the relationship among items based on their patterns, association rule hypergraph partitioning technique is used in the offline part. These items are known as frequent item sets which further used as hyerredges to form a hypergraph, and then partitioned into set of clusters [4].

The relational OLAP method for raw log data and mined logs (association rules and clusters discovered from logs) was proposed by Joshi *et al*. (1999). To clean the data, pre-processing phase was used and log data was transformed into Oracle SQL loader format. CGI was used for the warehousing and Perl CGI scripts were used for web interface [5]. To discover association rules, Agrawal and Srikant (1994) implements a variation of the apriori algorithm via SGI's 'Mineset' [6]. Krishnapuram *et al*. (1999) developed fuzzy C-medoids algorithm to cluster and capture a graded notion of the similarity between sessions. It was possible to analyze both the Web logs and traversal patterns using an online query [7]. For finding usage patterns, Zaı¨ane *et al*. (1998) combined OLAP and data mining techniques [8].

Association-rule mining and clustering had been used in many research projects (Berkan and Trubatch, 2002; Chen and Kuo, 2000) to discover the access patterns and sequential navigational patterns [9, 10]. Some of the popular projects for general access pattern tracking are 'WebLogMining' (Zaı¨ane, 2001), 'WUM' (Spiliopoulou and Faulstich, 1998) and 'WebSIFT' (Cooley *et al*., 1999) [11, 13]. The research related to customized usage tracking research includes 'Adaptive Web sites' (Perkowitz and Etzioni, 1998) and WebWatcher (Joachims *et al*.,1997) [14,15]. The detailed overview of research related to Web Usage Mining could be obtained from (Srivastava *et al*., 2000; Wang *et al*., 2002; Cheung *et al*., 1997, 2001) [16, 17, 18].

Wang *et al*. (2002); Cheung *et al*. (1997, 2001); Han *et al*. (2000); Jespersen *et al*. (2002) proposed that the current research are focusing on finding patterns but with little effort was made on the detailed pattern / trend analysis that varies with the web environments and the intelligent paradigms [17, 18, 19, 20].

1.1 Roadmap

In the subsequent section, Section 2 describes Web Data Mining Architecture. Section 3 describes Analysis of Web Log Data. Section 4 refers to Conclusion and future work. Finally acknowledgements and References are provided.

2. Web Data Mining Architecture

Web data mining normally contains five processing stages including data gathering, data preprocessing, extraction / selection, discovery and analysis and decision making / prediction. The architecture of web data mining is shown in Figure 2.

2.1 Data Gathering

The web log data gathered is generated through webalizer which is one of the web log analyzer tool. By searching the best data analyzer tool, it becomes easier to conclude rather than generating basic statistical data which was not capable of providing semantic information.

2.2 Data Preprocessing

To get correct analysis, it's necessary to eliminate irrelevant and non-interesting web data as one of the first step in preprocessing of data. The data with most recently accessed were indexed with higher value of 'time index' while the data with least recently accessed were kept in lower hierarchy.

2.3 Extraction / Selection

Due to the explosion of the internet world in the past few years, the billions of requests are received by the online servers. The dynamic nature of the data requires an intelligent web mining paradigm which helps in extracting the valid profiles or patterns [17].

2.4 Discovery and Analysis

The discovery and analysis have been done via data analytical and graphing workspace tool such as Origin. To predict the relevant information and knowledge, this phase make the process easier.

2.5 Decision Making / Prediction

It is the thought process of selecting a logical choice from the existing cases. A person must weight the positives and negatives of each case, and consider all the alternatives in making a good decision for particular situation.

Figure 2. Web Data Mining Architecture

3. Analysis of Web Log Data

In the current research, web log data taken from Guru Jambheshwar University of Science & Technology website from February 2014 to January 2015. Web log data of university website is generated through Webalizer but it is having one major limitation. It cannot differentiate between the human user visit and robot visit. So to overcome this, Origin tool have been used for analysis of web log data which is one of the data analytic tool.

Web data patterns of university website is shown in Figure 3 and Figure 4 while highlighting the daily average summary and monthly average summary including particular such as number of sites, visit, pages, files and hits from February 2014 to January 2015.

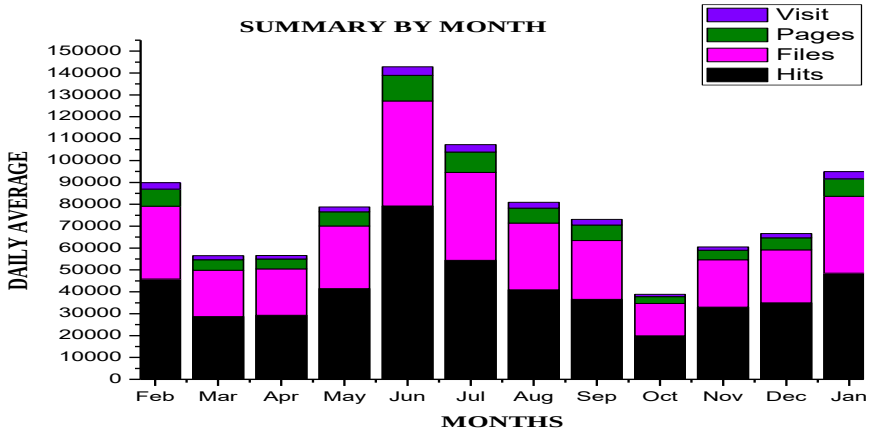

Figure 3. Daily Average Summary from February 2014 to January 2015

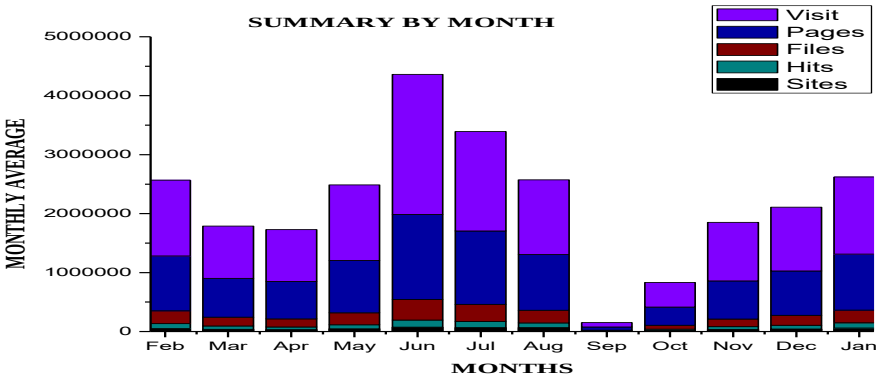

Figure 4. Monthly Average Summary from February 2014 to January 2015

University website's main web server receives highest 4479215 hits in June 2014 and lowest 190885 hits in September 2014 in terms of monthly average. Further, it obtain highest 77685 hits in June 2014 and lowest 20032 hits in October 2014 in terms of daily average. The biggest challenge is to extract relevant information from such data sets which is too huge and chaotic. As organization is growing itself, so parallel there is growth of web traffic also.

Monthly average size in Kbytes from Feb 2014 to Jan 2015 is shown in Figure 5, where the maximum average size in Feb 2014 is 28076187 kbytes. The Layout of Origin Pro 8 Data Analysis Tool is shown in Figure 6.

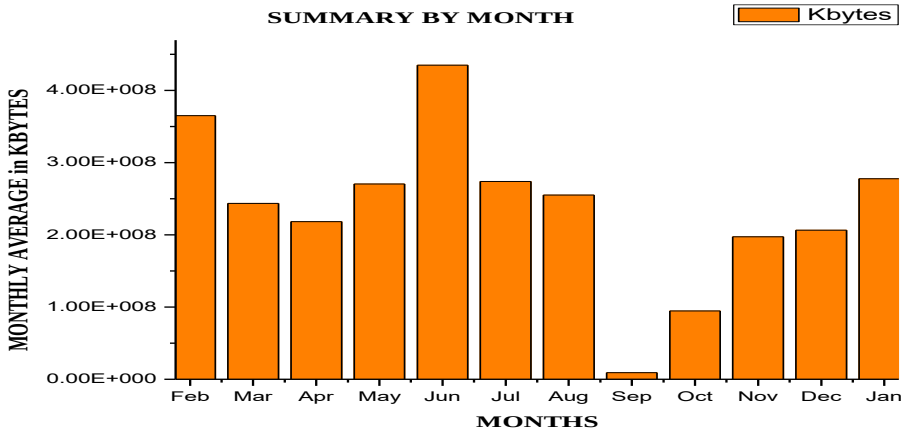

Figure 5. Monthly Average Size in Kbytes

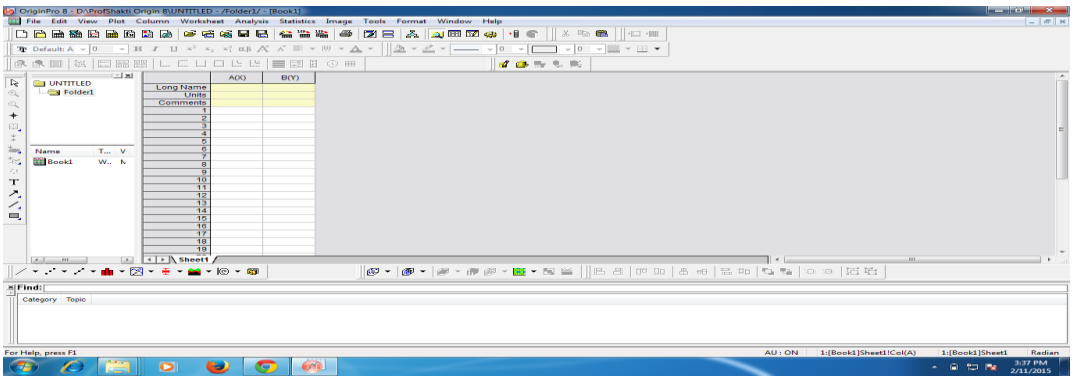

Figure 6. Layout of Origin Pro 8 Data Analysis Tool

The daily and hourly web data patterns from February 2014 to January 2015 are shown in Figure 7(7.1-7.12) and Figure 8(8.1-8.12). The University's daily and hourly patterns result in a similar trend but during the peak hours (10:00-18:00 Hrs), it becomes more difficult and complex when volume of web traffic is high and chaotic.

For the month of January 2015, the daily web data highlight the aggregate Hits occurs and which Files, Pages have been visited. The highest Hits per Day were 75934, the highest Files per Day were 56131, the highest Pages per Day were 12698, the highest visits per Day were 4635 and the highest Kbytes per Day were 22903987.

The hourly web data illuminating the number of Hits, Files and Pages. The highest Hits per hour were found to be 117977 in the month of January 2015.

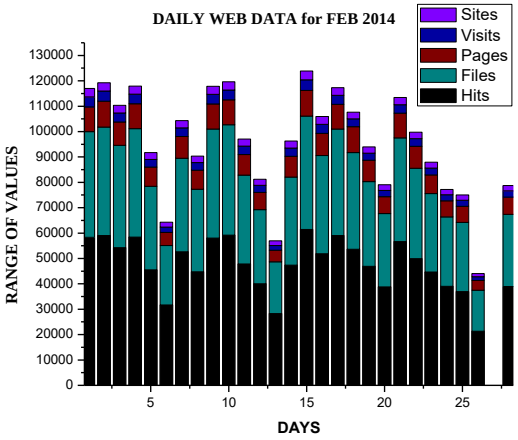

Figure 7.1. Daily Web Data for February 2014

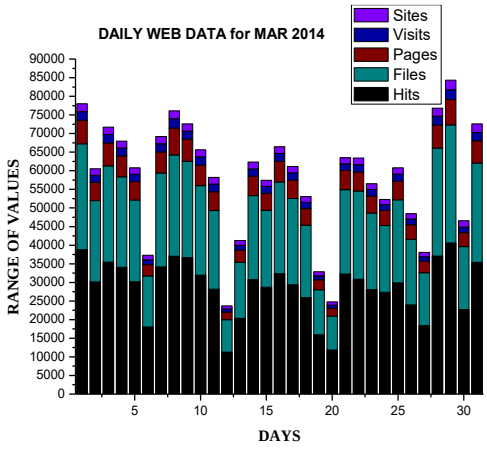

Figure 7.2. Daily Web Data for March 2014

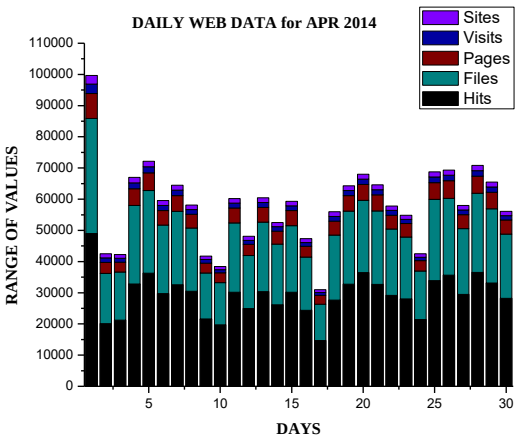

Figure 7.3. Daily Web Data for April 2014

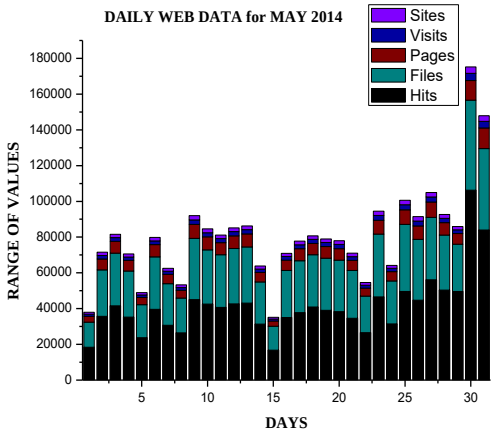

Figure 7.4. Daily Web Data for May 2014

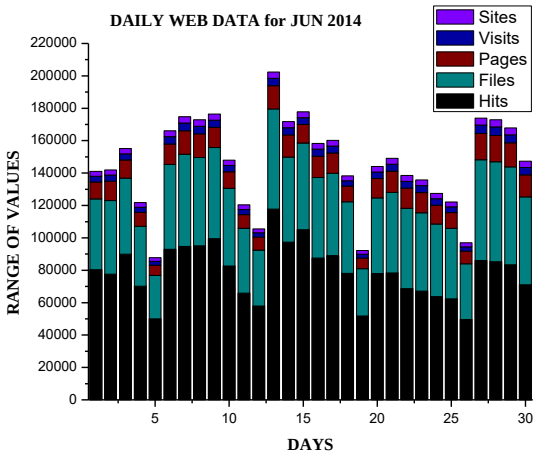

Figure 7.5. Daily Web Data for June 2014

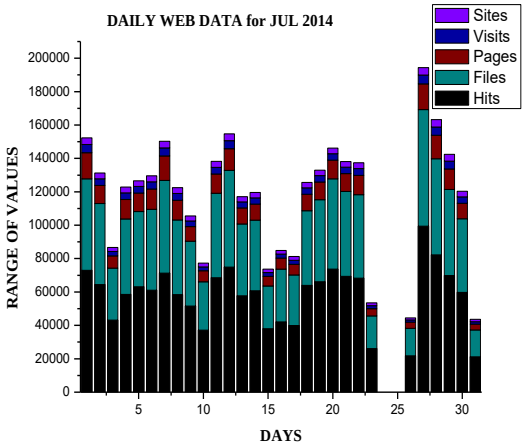

Figure 7.6. Daily Web Data for July 2014

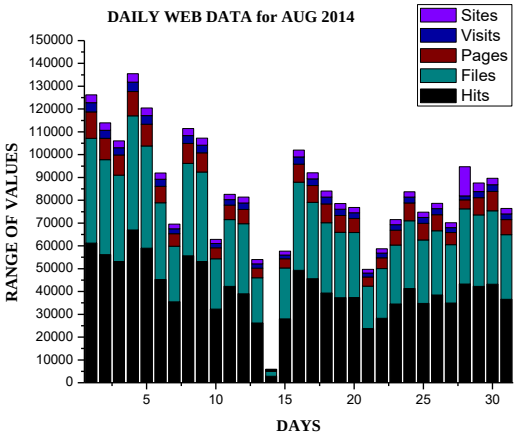

Figure 7.7. Daily Web Data for August 2014

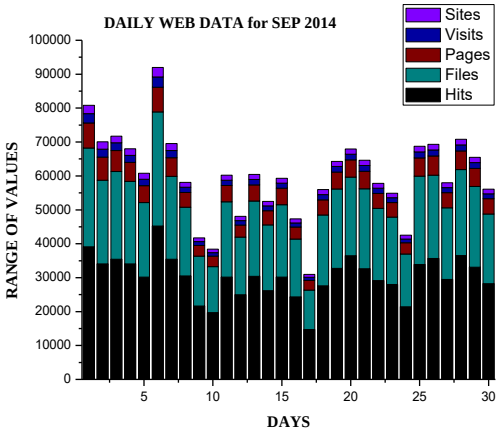

Figure 7.8. Daily Web Data for September 2014

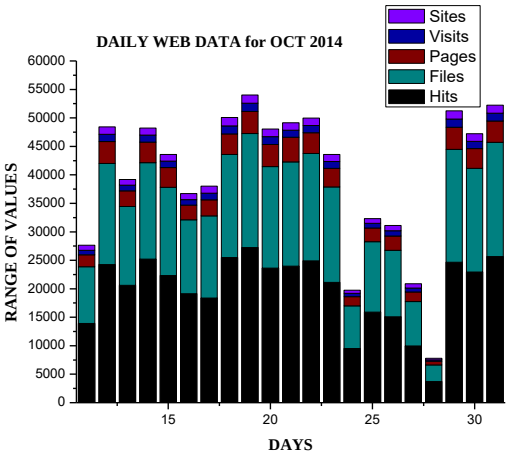

Figure 7.9. Daily Web Data for October 2014

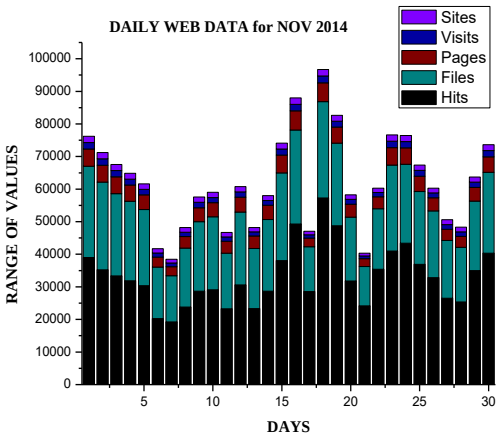

Figure 7.10. Daily Web Data for November 2014

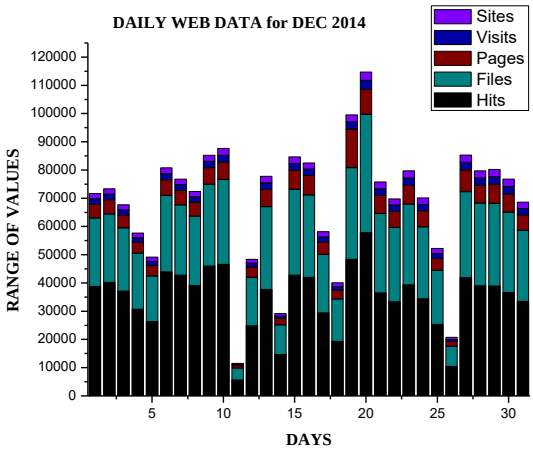

Figure 7.11. Daily Web Data for December 2014

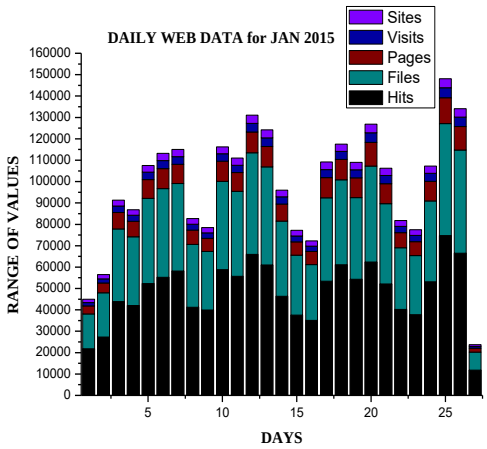

Figure 7.12. Daily Web Data for January 2015

Figure 7. Daily Web Data Patterns from February 2014 to January 2015 (Figure 7.1-7.12)

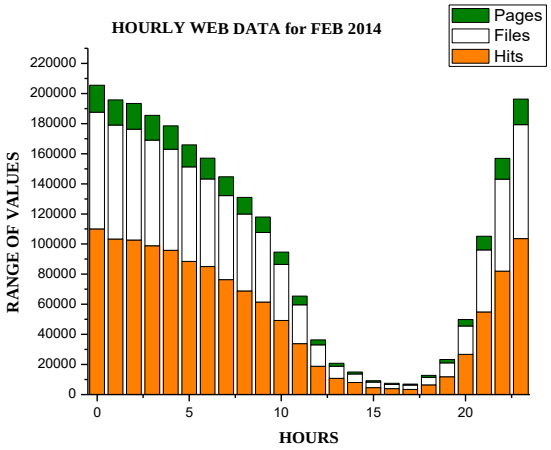

Figure 8.1. Hourly Web Data for February 2014

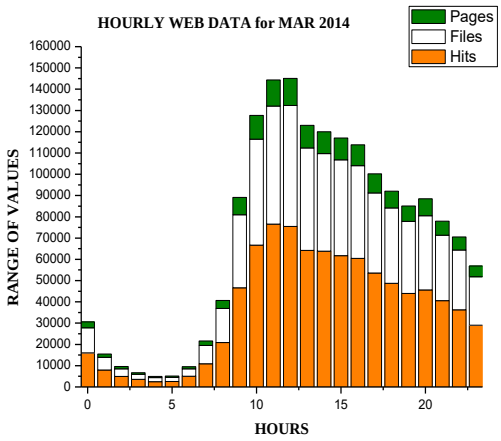

Figure 8.2. Hourly Web Data for March 2014

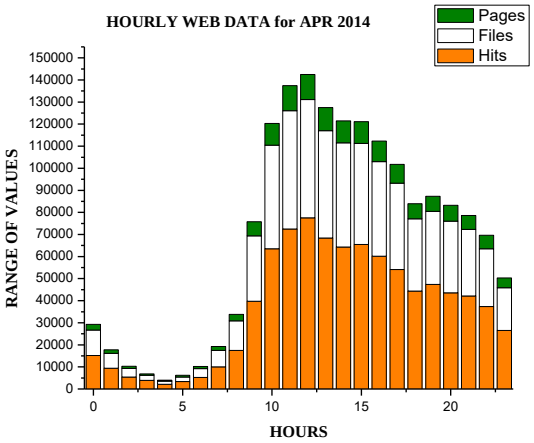

Figure 8.3. Hourly Web Data for April 2014

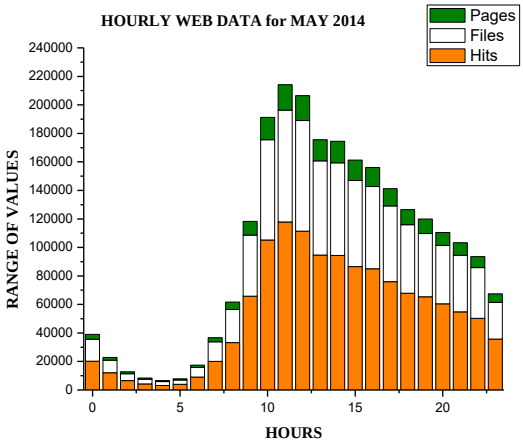

Figure 8.4. Hourly Web Data for May 2014

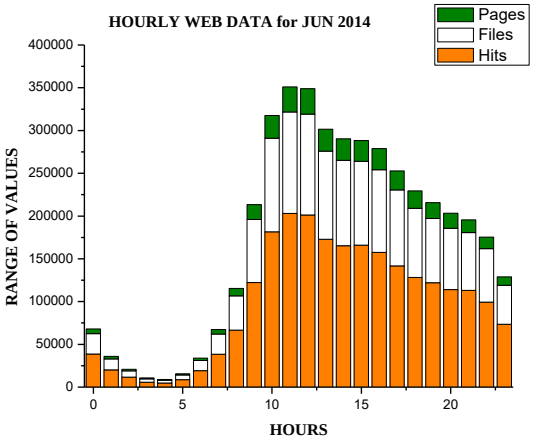

Figure 8.5. Hourly Web Data for June 2014

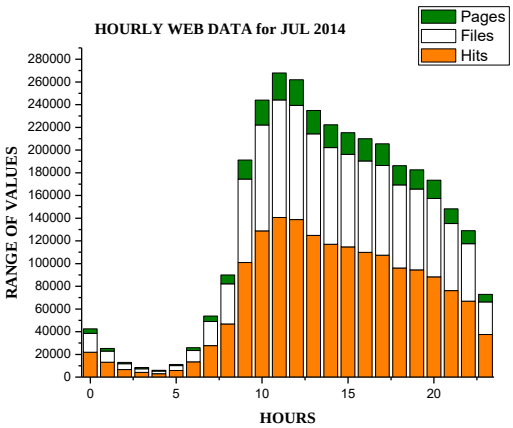

Figure 8.6. Hourly Web Data for July 2014

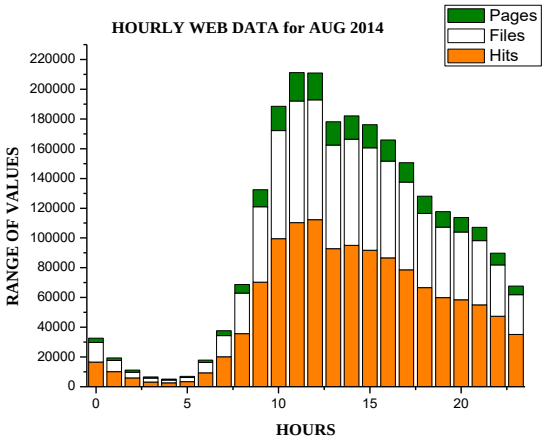

Figure 8.7. Hourly Web Data for August 2014

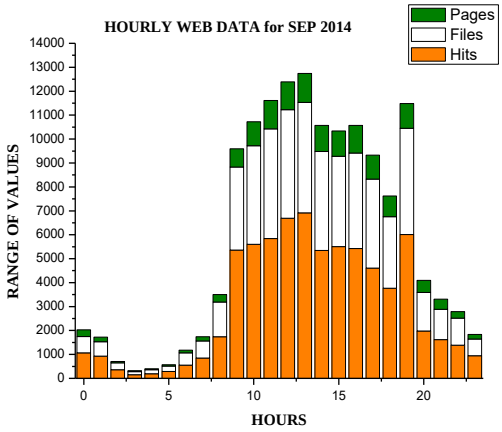

Figure 8.8. Hourly Web Data for September 2014

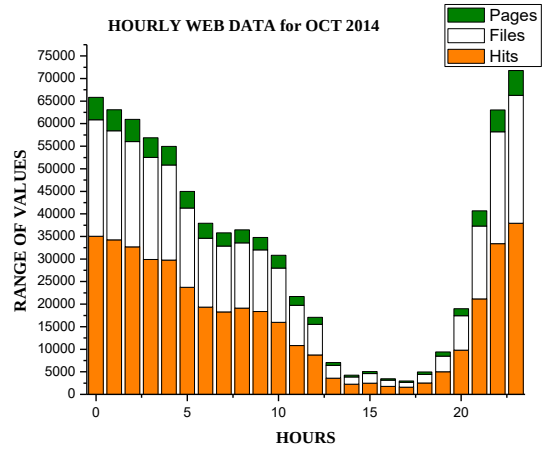

Figure 8.9. Hourly Web Data for October 2014

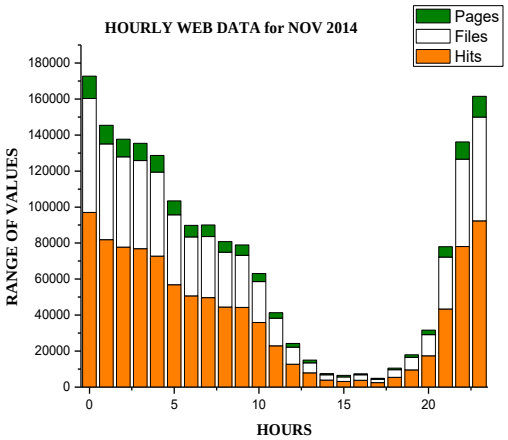

Figure 8.10. Hourly Web Data for November 2014

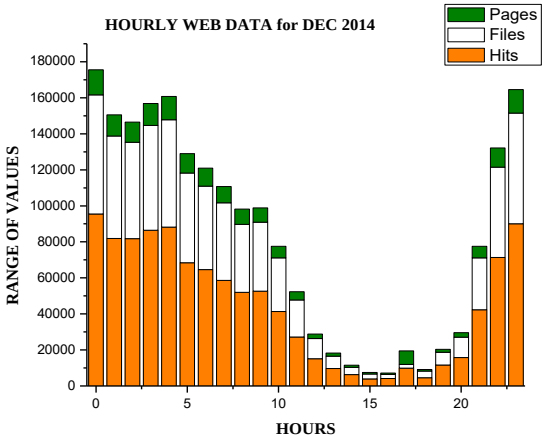

Figure 8.11. Hourly Web Data for December 2014

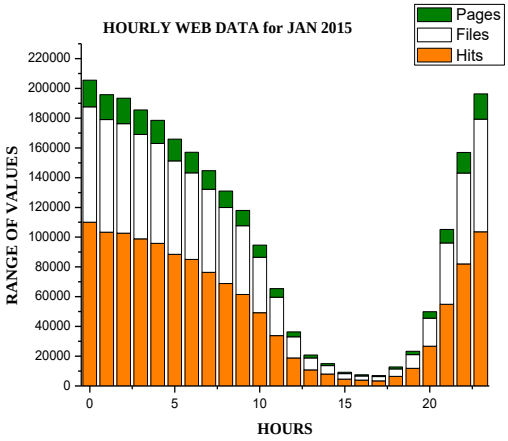

Figure 8.12. Hourly Web Data for January 2015

Figure 8. Hourly Web Data Patterns from February 2014 to January 2015 (Figure 8.1-8.12)

4. Conclusion and Future Work

The research on university website's various web data patterns reveals the necessity to incorporate more intelligent techniques and methods for mining useful information and predicting web traffic analysis. Through the data analytical and graph workspace tool, we are feasible in providing relevant information related to the user access.

In this exploration, the web traffic data during the university's peak working time is considered. Our forthcoming study will to incorporate agent communication language which will helps in making the web mining system not only efficient but also an intelligent system.

Acknowledgements

Thanks to Mr. Mukesh Kumar, Head, University Computer & Informatics Centre, Guru Jambheshwar University of Science & Technology (GJUS&T), Hisar, Haryana, India for his invaluable contribution to this analysis. Any belief, result and assumption stated in this research paper are those of the author and do not necessarily reflect those of GJUS&T, Hisar.

References

[1] B. Hawwash and O. Nasraoui, "Mining and Tracking Evolving Web User Trends from Large Web Server Logs", Statistical Analysis and Data Mining, Wiley Periodicals, Inc: MA, vol. 3, no. 2, **(2010)**, pp. 106-125.
[2] N. C. Romano, C. Donovan, H. C. Chen, J. F. Nunamaker, "A Methodology for Analyzing Web-based Qualitative Data", Journal of Management Information Systems, vol. 19, no. 4, **(2003)**, pp. 213-246.
[3] P. Rossler, "Content Analysis in Online Communication: A challenge for traditional methodology", Available: Batinic B, Reips UD, Bosnjak M (Eds.) Online social sciences Seattle WA: Hogrefe & Huber, **(2002)**.
[4] B. Mobasher, R. Cooley, J. Srivastava, "Creating Adaptive Web sites through Usage-based Clustering of URLs", Proceedings of 1999 Workshop on Knowledge and Data Engineering Exchange, Chicago Illinois USA, **(1999)**, pp. 19-25
[5] K. P. Joshi, A. Joshi, Y. Yesha, R. Krishnapuram, "Warehousing and Mining Web Logs", Proceedings of the 2nd ACM CIKM Workshop on Web Information and Data Management, Kansas City Missouri USA, **(1999)**, pp. 63-68.
[6] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules", Proceedings of the 20th International Conference on Very Large Databases, Santiago Chile, **(1994)**, pp. 487-499.
[7] R. Krishnapuram, A. Joshi, L. Yi, "A Fuzzy relative of the k-medoids Algorithm with Application to Document and Snippet Clustering", Proceedings of IEEE International Conference on Fuzzy Systems, FUZZIEEE(99), Seoul, Korea, **(1999)**, pp. 1281-1286.
[8] O. R. Zaı¨ane, , M. Xin, J. Han, "Discovering Web Access Patterns and Trends by applying OLAP and Data Mining Technology on Web Logs", Proceedings of Advances in Digital Libraries Conference, Santa Barbara California USA, **(1998)**, pp. 19-29.
[9] R. C. Berkan, S. L. Trubatch, "Fuzzy Logic and Hybrid Approaches to Web Intelligence Gathering and Iinformation Management", Proceedings of 2002 World Congress on Computational Intelligence, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE(02), Special Session on Computational Web Intelligence (CWI) Honolulu Hawaii USA, **(2002)**, pp. 1033-1038.
[10] P. M. Chen, F. C. Kuo, "An Information Retrieval System based on an User Profile", Journal of System Software, Vol. 54, **(2000)**, pp. 3-8.
[11] O. R. Zaı¨ane, "Building Virtual Web Views", Journal of Data Knowledge Engineering, Vol. 39(2), **(1998)**, pp.143-163.
[12] M. Spiliopoulou, L. C. Faulstich, "WUM: A Web Utilization Miner", Proceedings of Workshop on the Web and Data Bases (WebDB(98), Valencia Spain, **(1998)**, pp. 109-115.
[13] R. Cooley, P. N. Tan, J. Srivastava, "WebSIFT: The Website Information Filter System", Proceedings of the Web Usage Analysis and User Profiling (WebKDD'99), Workshop on Web Mining, an Diego CA USA, **(1999)**, pp. 163-182.
[14] M. Perkowitz, O. Etzioni, "Adaptive Websites: Automatically Synthesizing Web Pages", Proceedings of the 15th National Conference on Artificial Intelligence and 20th Innovative Applications of Artificial Intelligence Conference (AAAI(98, IAAI(98)) Madison Wisconsin USA, **(1998)**, pp. 727–732.
[15] T. Joachims, D. Freitag, T. Mitchell, "WebWatcher: A Tour Guide for the World Wide Web", Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI(97)) Nagoya Japan, **(1997)**, pp. 770-775.

[16] J. Srivastava, R. Cooley, M. Deshpande, T. N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", In SIGKDD Explorations, Vol. 1(2), **(2000)**, pp. 12-23.
[17] X. Wang , A. Abraham, K. A. Smith, "Web Traffic Mining using a Concurrent Neuro-Fuzzy Approach", Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, Computing Systems: Design, Management and Applications, Santiago Chile, **(2002)**, pp. 853-862.
[18] D. Cheung, B. Kao, J. Lee, "Discovering User Access Patterns on the World Wide Web", Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD(97)), Vol. 10, **(1997)**, pp. 463-470.
[19] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proceedings of the ACMSIGMOD International Conference on Management of Data (SIGMOD(00)), Dallas, TX, USA, **(2000)**, pp. 1-12.
[20] S. E. Jespersen, J. Thorhauge, T. B. Pedersen, "A Hybrid Approach to Web Usage Mining", Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery DaWaK(02), Aix-en-Provence, France, **(2002)**, pp. 73-82.
[21] D. Boley, M. L. Gini, R. Gross, E. H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, J. Moore, "Document Categorization and Query Generation on the World Wide Web using WebACE", Journal of Artificial Intelligence, Vol. 13(5-6), **(1999)**, pp. 365-391.
[22] F. Masseglia, P. Poncelet, R. Cicchetti, "An efficient Algorithm for Web Usage Mining", Journal of Networking Inf. Systems (NIS), Vol. 2(5-6), **(1999)**, pp. 571-603.
[23] M. Kitsuregawa, M. Toyoda, I. Pramudiono, "Web Community Mining and Web Log Mining: Commodity Cluster based Execution", Proceedings of the 13th Australasian Database Conference (ADC(02)), Melbourne Australia, Vol. 5, **(2002)**, pp. 3-10.
[24] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", IEEE Computer Society, **(1997)**, pp. 558-567.
[25] T. Y. Jing, "Supporting Research with Weblogs: A Study on Web-Based Research Support Systems", Proceedings of Web Intelligence and Intelligent Agent Technology Workshops IEEE/WIC/ACM International Conference, **(2006)**, pp. 161-164.

Authors

Shakti Kundu received his M.Tech. in Computer Science & Engineering from Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India in 2010, M.Phil. in Computer Science from Chaudhary Devi Lal University, Sirsa, Haryana, India in 2008, M.C.A. from Kurukshetra University Kurukshetra, Haryana, India in 2006. Presently he is pursuing his Ph.D in CSE from DIT University, Dehradun, India. The author current research interests are Web Mining and Web Testing. He is life member of CSI, ISTE, IAENG, AIRCC and IAEME.

M L Garg, presently Professor and Head, Department of Computer Science & Engineering at DIT University, Dehradun, India, has obtained his Ph.D degree in Computer Science & Engineering from Thapar Institute of Engineering & Technology (Deemed University), Patiala, India with collaborative research work at IIT Delhi, in the year 1992. His area of research includes Fuzzy Logic Genetic Algorithms & Knowledge Representation and Reasoning. He has published about thirty research articles in peer reviewed Int'l and Nat'l Jr.'s and Conferences.

