AN INTELLIGENT APPROACH OF WEB DATA MINING

Shakti Kundu

School of Computer Applications,
Lovely Professional University, Jalandhar-Delhi GT road, NH-1, Phagwara, Punjab, India.
shaktikundu@rediffmail.com

***Abstract* - With an explosive growth of the World Wide Web, websites are playing an important role in providing an information and knowledge to the end users. Web usage patterns are an important aspect to discover hidden and meaningful information. It will be big challenge in web mining when the volume of traffic is large and the volume of web data is still in the growing phase. To face the challenge an intelligent approach of web traffic analysis has been highlighted in this paper.**

Keywords* - *Analysis, Data, log, Mining, Web.

I. INTRODUCTION

Web Mining [17] refers to overall process of discovering potentially useful and previously unknown information from web document and services web mining could be viewed as an extension of standard data mining to web data.

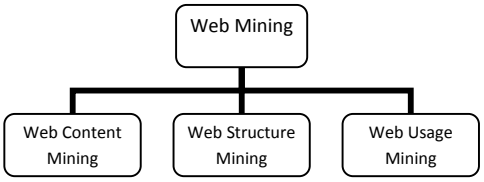

Figure 1. Taxonomy of Web Mining

The Web analysis relies on three general sets of information given: past usage patterns, degree of shared content [1] and inter-memory associative link structures [2] corresponding to the three subsets in Web mining namely: (i) Web usage mining [3], (ii) Web content mining and (iii) Web structure mining. In Web usage mining, the pattern discovery consists of several steps including statistical analysis, clustering [4-5], classification and so on [6-7-8]. For instance, in E-commerce, analyzing the Web usage data can help organizations to understand the customer Web browsing patterns [9] which in turn might help to facilitate E-commerce specific processing such as: Web structure management for designing a better Website, and promotional campaigns for building customized advertisements and for making better strategic marketing decisions [10]. Most of the current research is focusing on finding patterns but with little effort on the detailed pattern/trend analysis that varies with the Web environments and the intelligent paradigms considered [8-9-10-11-12].

The article by L.J. Haravu and A. Neelameghan offers information about the advances in text and data-mining in knowledge-based software. After research and experience using these technologies, both Haravu and Neelameghan suggest two essential methods of creating these platforms. First, using natural language processing software in text mining, and second, the planning, designing, and developing of a comprehensive multiple media product that would satisfy their target audiences' needs. They trust that these text and data-mining products can only become more useful if the features of a subject classification system are incorporated into text mining techniques and products. In other words, the specialized role of human language technologies in the library and information science venue has the potential to become standardized, and thus predicted [13].

The article, written by computer engineers, closely relates to Condé Nast's text mining software objectives by discussing web usage mining, user profiles, web analytics, and data streams. In other words, recently, current publishing organizations have started dedicating its resources to tracking various users' behavior on their online databases to better understand and satisfy their needs. As a direct result, web usage mining tools were developed to help them use web logs to discover usage patterns and profiles. Many publishing companies refer to this information as valuable evidence or case studies for usability. In addition, with this data, companies like Condé Nast are better able to generate accurate text-mining languages that will best satisfy their target audiences [14].

The article addresses lexicons and their connection to text mining. In linguistics, the lexicon of a language is its expressions, words, and vocabularies. In other words, lexicons, similar to text mining, are a language's inventory of lexemes, or combination pattern. Efficient computing is consistently advancing as a field, and allows new forms of human-computer interactions, in addition to the use of a standardized natural language. There is a common perception that the future of human-computer interaction lies in themes such as entertainment, aesthetics, and publishers, to name a few. This article helps with studying the relationship between natural language and effective information and dealing with its computational treatment, while valuing this practice as crucial to future development. Later in this article, the authors present another linguistic resource for a lexical representation of affective knowledge that competes with Condé Nast's text-mining software, called Wordnet-Affect [15].

The remainder of the paper is organized as follows: In the subsequent Section II, architecture of web usage mining has been discussed. In Section III, the web traffic analysis has been presented to learn and predict the short-term and long-term usage patterns. Finally, some conclusion and future works are given in Section IV.

II. ARCHITECTURE OF WEB USAGE MINING

Web usage mining normally contains four processing stages including data collection, data preprocessing, pattern discovery and pattern analysis [10]. The architecture of web usage mining is shown in figure 2.

A. Data Collection:

The data source selected for our approach is from the Web traffic data generated by the 'Webalizer' Web access log file analyzer. It is a usual practice to embed Web trackers or Web log analysis tools to analyze the Web server log files for providing useful information to Web administrators. After browsing through some of the features of the best trackers available on the market [8], it is easy to conclude that rather than generating basic statistical data they really cannot provide much meaningful information.

In order to overcome the drawbacks of available Web log analyzers, the hybrid approach is proposed to discover hidden information and usage pattern trends, which could aid the Web managers for improving the management, performance and controlling of the Web servers.

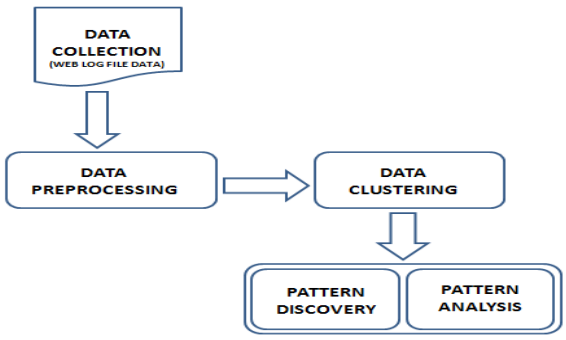

Figure 2. Web Usage Mining Architecture

B. Data Preprocessing

The web data used from February 2011 to January 2012 for the cluster analysis process. Selecting useful data is an important task in the data pre-processing stage. After some preliminary analysis, the statistical data were selected comprising the traffic data on the monthly, hourly and daily basis including request volume and page volume in each data type to generate the cluster models for finding Web user access and server usage patterns.

To build up a precise model and to obtain more accurate analysis, it is also important to remove irrelevant and noisy data as an initial step in pre-processing task. The most recently accessed data were indexed with higher value of 'time index' while the least recently accessed data were placed at the bottom with lowest value [16]. This is a very critical step to obtain more precise analysis result due to time dependence characteristic of Web usage data itself.

C. Data Clustering

With the increasing popularity of Internet, crores of requests (with different interests from different countries) are received by Web servers of large organizations, Guru Jambheshwar University is a truly national university with its main campus located in Hisar, Haryana, India. The university has plans to extend its educational services around the globe. Therefore, the huge traffic volume and the dynamic nature of the data require the necessity to implement efficient and intelligent Web mining framework. In Web usage mining research, the method of clustering is broadly used in different projects by researchers for finding the usage patterns or user profiles [8].
Among all the popular clustering algorithms, In websites, Logs Cluster Analysis is one of the best mining method. The cluster objects include user groups and web pages. The cluster of user groups plays an important role in providing personalized services in websites. In order to cluster user groups, we must describe their browser action.

D. Pattern Discovery and Pattern Analysis

The pattern discovery and pattern analysis have been done based on data analysis and graphing workspace tool such as Origin version 8. With the help of pattern discovery and analysis, it becomes easier to predict the relevant useful information and knowledge.

III. WEB TRAFFIC ANALYSIS

Web user access and server usage patterns had been analyzed from Guru Jambheshwar University of Science & Technology website's main web server located at http://www.gjust.ac.in/ [18]. Statistical/ text log file data have been used for experimentation provided by webalizer, which is one of the popular web server analysis tools.

The typical web traffic patterns of Guru Jambheshwar University of Science & Technology website in Figure 3 and Figure 4 are showing summary by month from February 2011 to January 2012 in terms of daily average (number of visit, pages, files, and hits) and monthly average (number of visit, pages, files, hits and sites).

Generally, in a month, Guru Jambheshwar University of Science & Technology website's main web server in terms of monthly average receives over maximum to 79215 hits in the month of June 2011 and minimum to 19885 hits in the month of October 2011. Similarly in terms of daily average, it receives over maximum to 118700 hits in the month of June 2011 and minimum to 22664 hits in the month of October 2011. It becomes a real challenge to discover hidden information or to extract usage patterns from such data sets which are too large in volume but also cover different aspects (domains, files accessed, daily and hourly access volume, page requests, etc.). Moreover, the volume of traffic keeps on growing due to the growth of the organization itself.

The University's monthly average size in terms of Kbytes from Febrauary2011 to January 2012 is shown in figure 5. The maximum average size in the month of February 2011 is 365115897 kbytes.

The University's daily and hourly web data patterns month wise from February 2011 to January 2012 are shown in Figures 4 and 5 respectively. The daily and hourly patterns nevertheless tend to follow a similar trend (as evident from the figures 6 and 7) the differences tend to increase during high traffic days (Monday – Friday) and during the peak hours (10:00-17:00 Hrs). Due to the enormous traffic volume and chaotic access behavior, the prediction of the user access patterns becomes more difficult and complex.

The daily web data for the month of February 2011 to January 2012 (Figure 6 (a-l)) basically tells about the number of Hits occurs, which Files, Pages have been visited. For the month of January 2012, the maximum Hits per Day were 74736, the maximum Files per Day were 52419, the maximum Pages per Day were 12039, the maximum visits per Day were 4703 and the maximum Kbytes per Day were 19333530.

The hourly web data for the month of February 2011 to January 2012 (Figure 7 (a-l)) basically tells about the number of Hits, Files and pages. The maximum Hits per hour in the month of January 2012 were found to be 110014. In our case while visiting the university website, an analysis is done on user access patterns which help in targeting particular section of website to specific group of users.

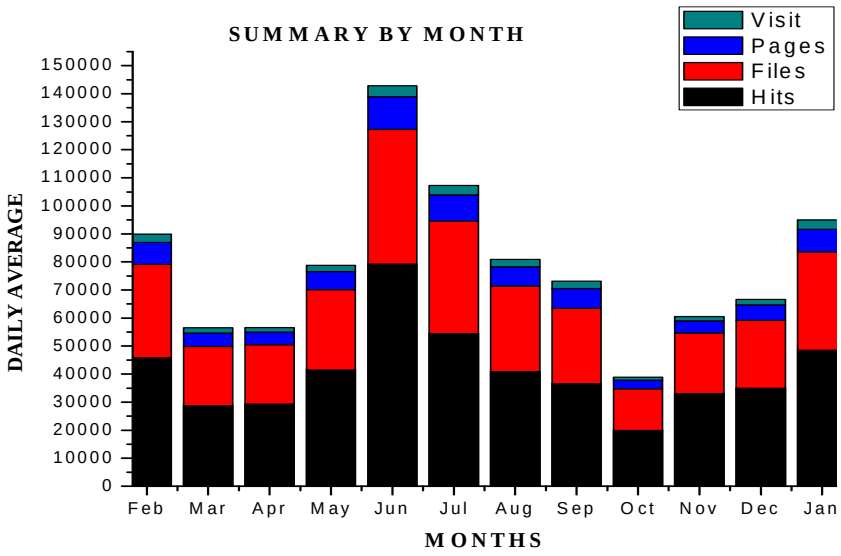

Figure 3. Guru Jambheshwar University of Science & Technology website's summary by month from February 2011 to January 2012 in terms of daily average (number of visit, pages, files, and hits)

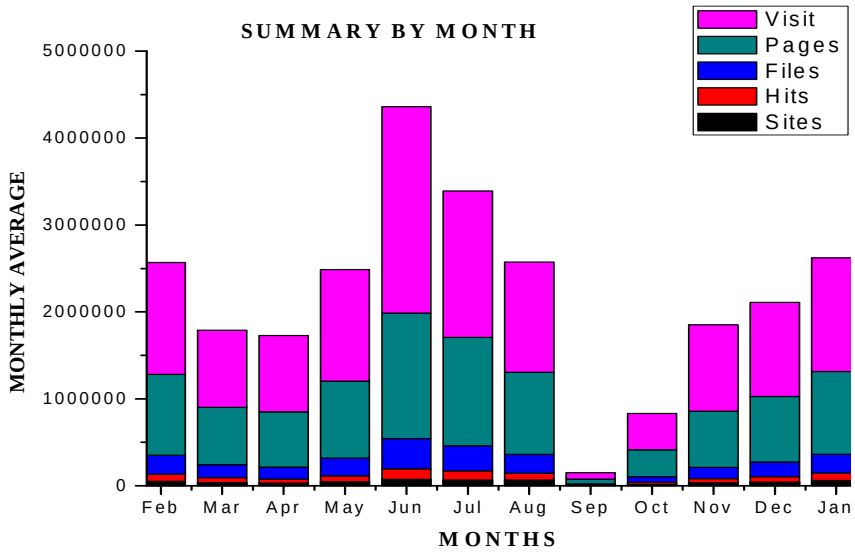

Figure 4. Guru Jambheshwar University of Science & Technology website's summary by month from February 2011 to January 2012 in terms of monthly average (number of visit, pages, files, hits and sites)

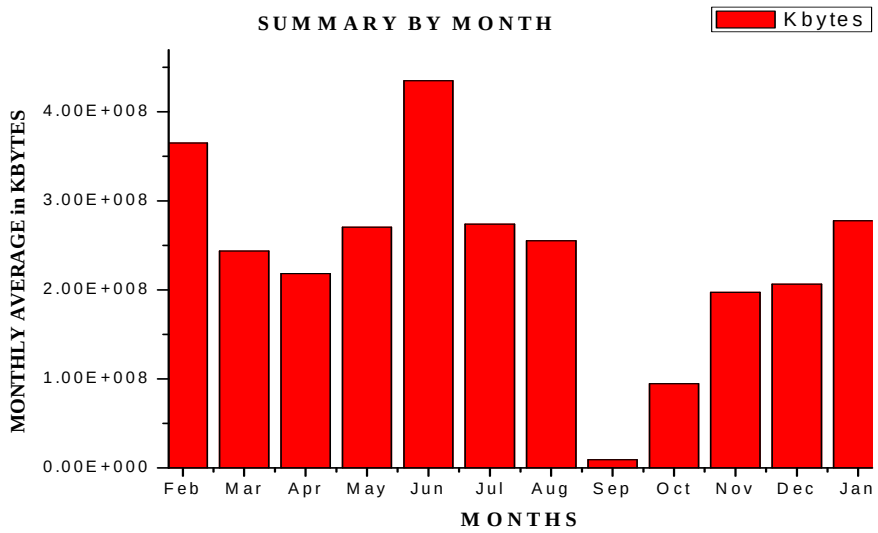

Figure 5. Guru Jambheshwar University's Monthly Average Size in Kbytes

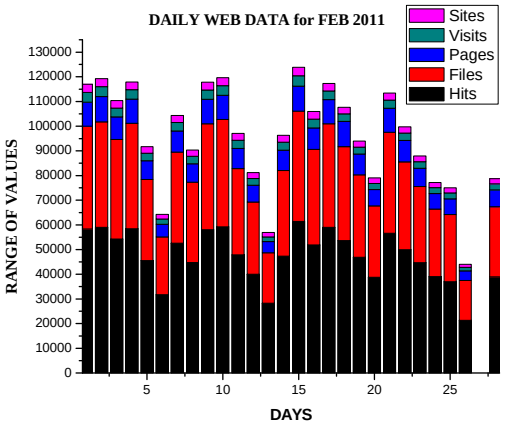

Fig. 6. (a) Daily Web Data for February 2011

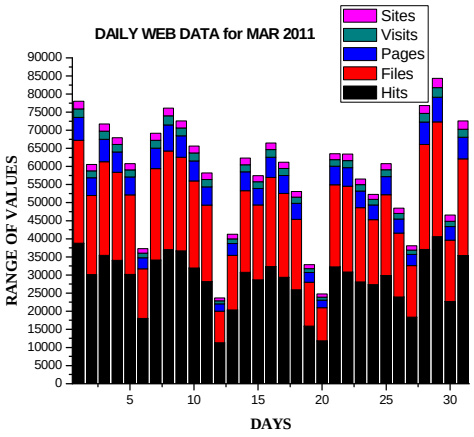

Fig. 6. (b) Daily Web Data for March 2011

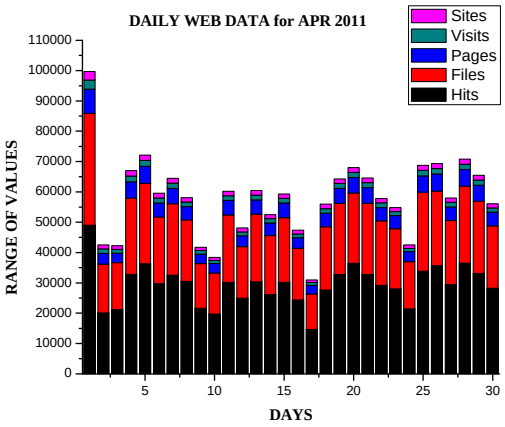

Fig. 6. (c) Daily Web Data for April 2011

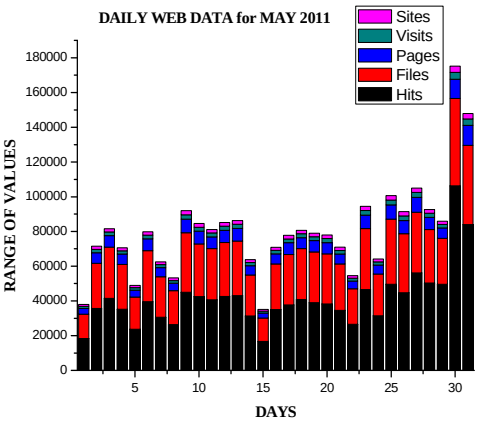

Fig. 6. (d) Daily Web Data for May 2011

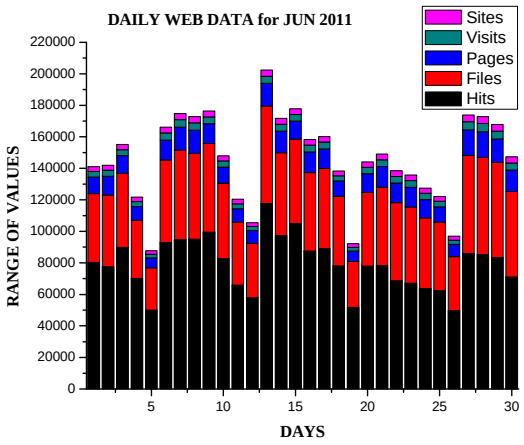

Fig. 6. (e) Daily Web Data for June 2011

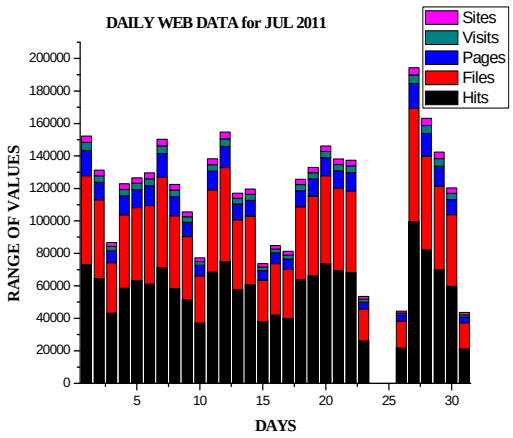

Fig. 6. (f) Daily Web Data for July 2011

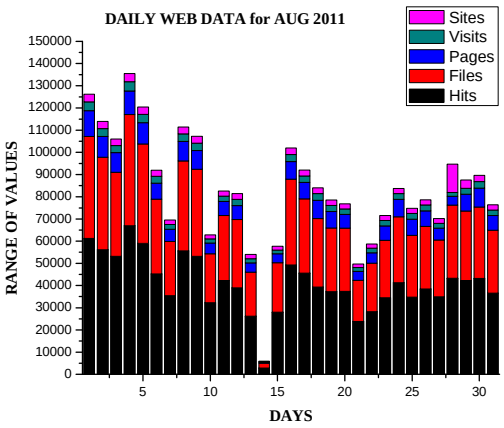

Fig. 6. (g) Daily Web Data for August 2011

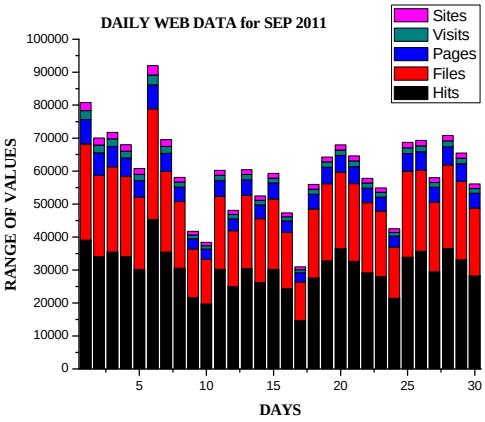

Fig. 6. (h) Daily Web Data for September 2011

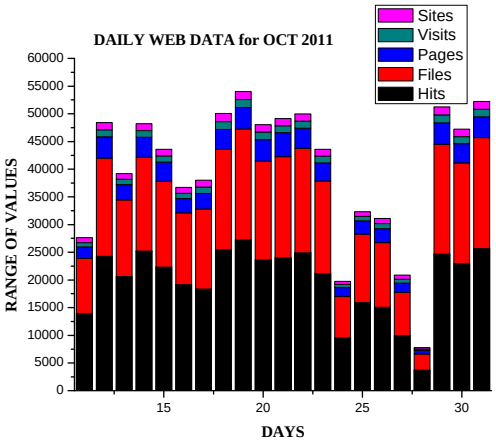

Fig. 6. (i) Daily Web Data for October 2011

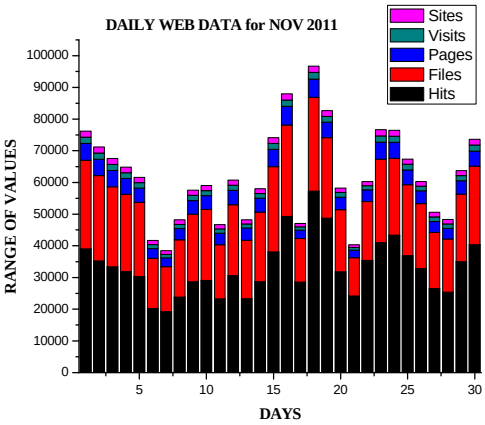

Fig. 6. (j) Daily Web Data for November 2011

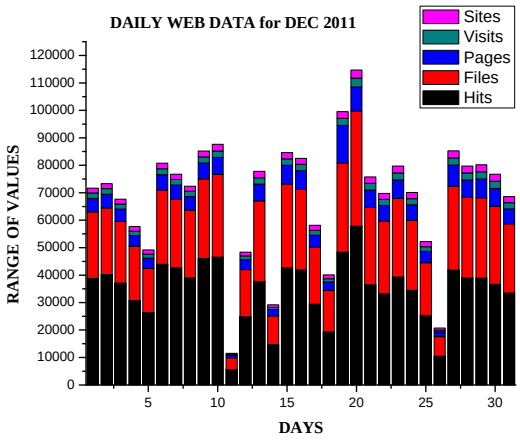

Fig. 6. (k) Daily Web Data for December 2011

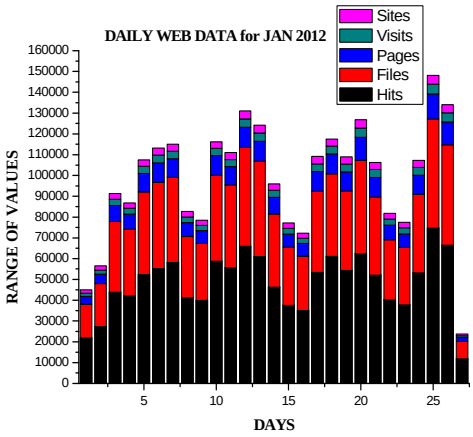

Fig. 6. (l) Daily Web Data for January 2011

Figure 6. Guru Jambheshwar University's daily web data patterns month wise from February 2011 to January 2012 (Fig. 6 (a-l))

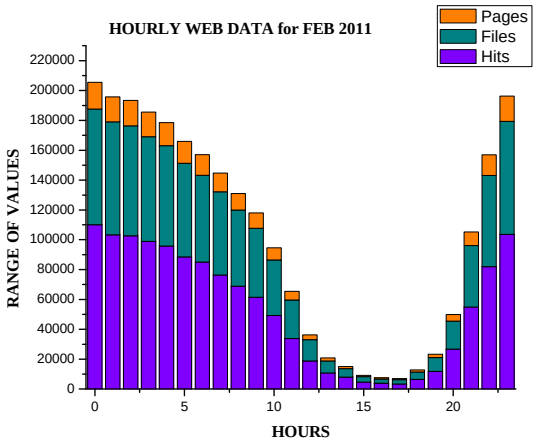

Fig. 7. (a) Hourly Web Data for February 2011

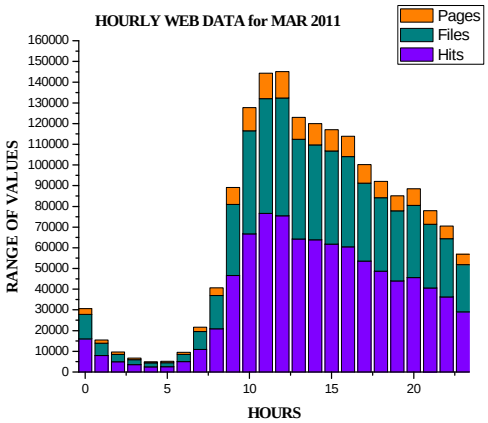

Fig. 7. (b) Hourly Web Data for March 2011

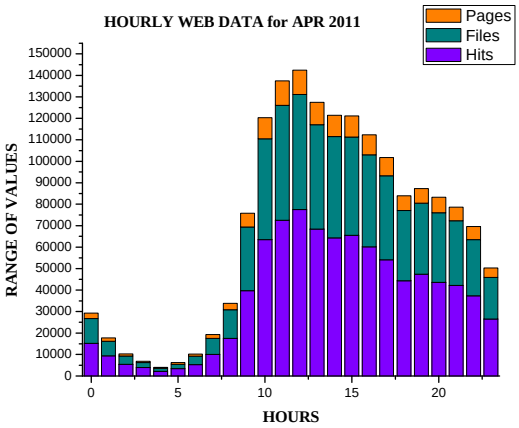

Fig. 7. (c) Hourly Web Data for April 2011

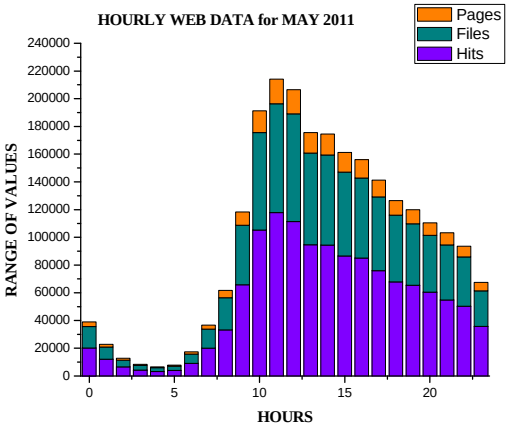

Fig. 7. (d) Hourly Web Data for May 2011

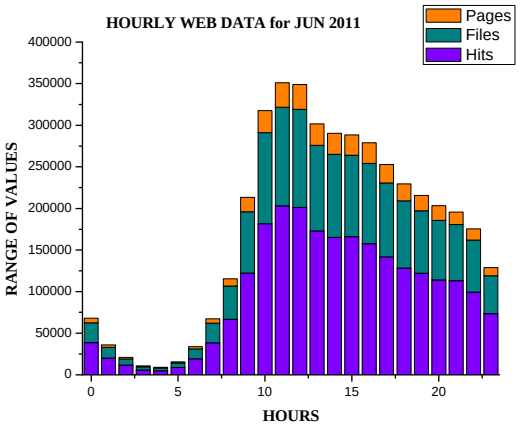

Fig. 7. (e) Hourly Web Data for June 2011

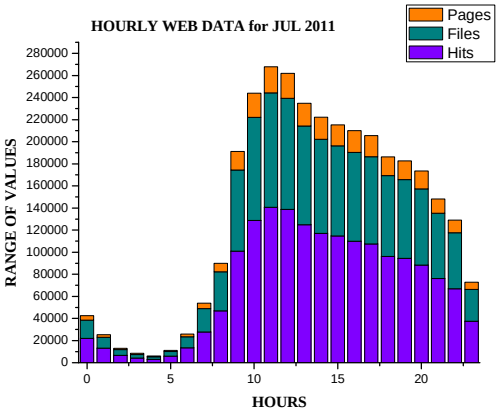

Fig. 7. (f) Hourly Web Data for July 2011

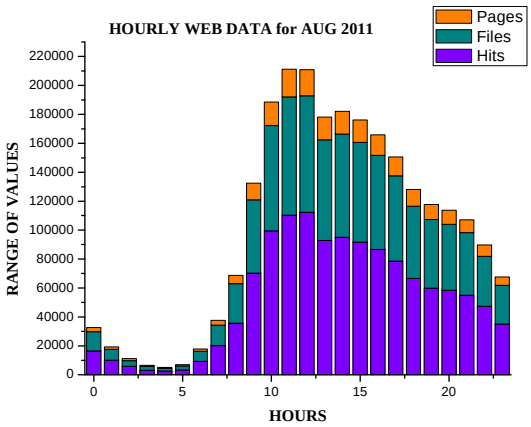

Fig. 7. (g) Hourly Web Data for August 2011

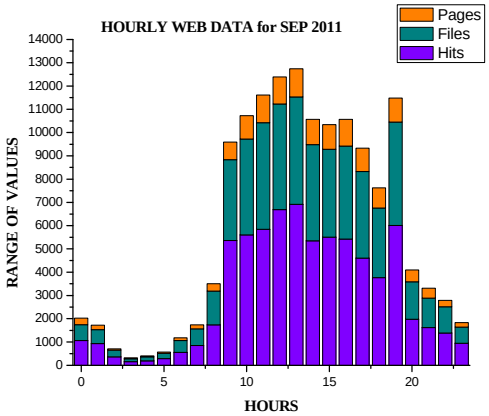

Fig. 7. (h) Hourly Web Data for September 2011

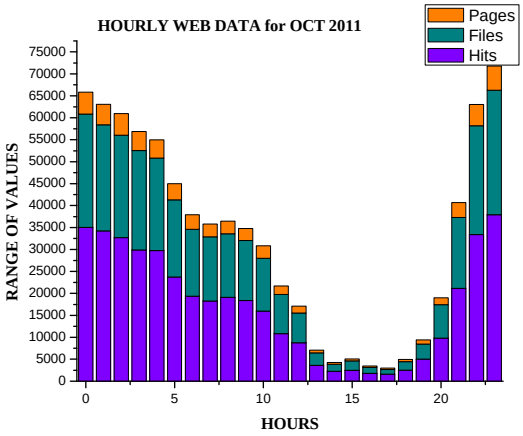

Fig. 7. (i) Hourly Web Data for October 2011

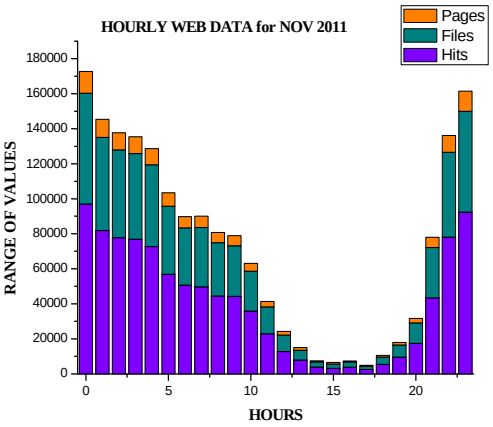

Fig. 7. (j) Hourly Web Data for November 2011

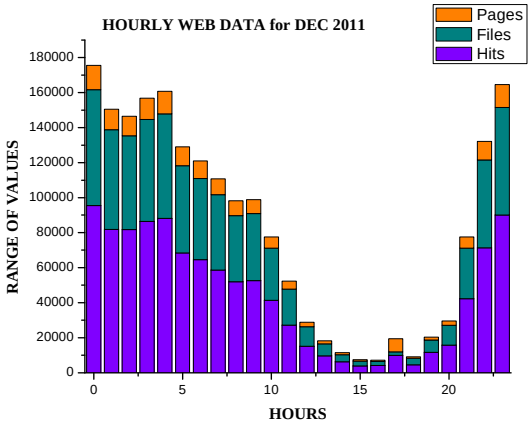

Fig. 7. (k) Hourly Web Data for December 2011

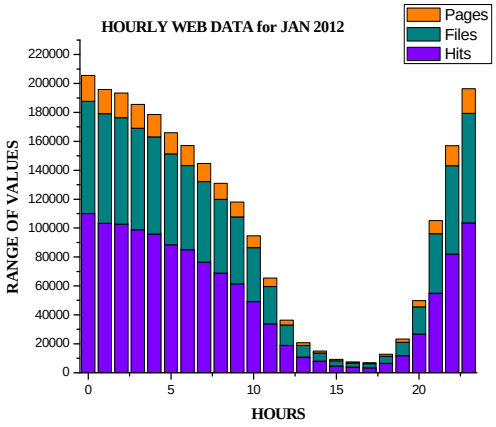

Fig. 7. (l) Hourly Web Data for January 2012

Figure 7. Guru Jambheshwar University's hourly web data patterns month wise from February 2011 to January 2012 (Fig. 7 (a-l))

IV. CONCLUSION AND FUTURE WORK

The study on Guru Jambheshwar University of Science and Technology website's various web data patterns reveals the necessity to incorporate more intelligent techniques and methods for mining useful information and predicting trend analysis. From the data analysis and graphing workspace tool, we are helpful in providing useful information related to the user access patterns, which could not be possible by using traditional approaches.

In this research, only the Web traffic data during the university's peak working time is considered. Our future research will to incorporate cluster diagnostics and verification tool with the help of agent communication language which will helps in making the web mining system not only efficient but also an intelligent system.

REFERENCES

[1] Boley D, Gini ML, Gross R, Han EH, Hastings K, Karypis G, Kumar V, Mobasher B, Moore J. Document categorization and query generation on the world wide web using WebACE. J Artif Intell Rev 1999;13(5-6): 365–91.

[2] Pirolli P, Pitkow J, Rao R. Silk from a sow's ear: extracting usable structures from the web. In Proceedings of Conference on Human Factors in Computing Systems (CHI(96), Vancouver, British Columbia, Canada 1996;1996:118–25.

[3] Masseglia F, Poncelet P, Cicchetti R. An efficient algorithm for web usage mining. J Networking Inf Syst (NIS) 1999; 2(5-6):571–603.

[4] Kitsuregawa M, Toyoda M, Pramudiono I. Web community mining and web log mining: commodity cluster based execution. In Proceedings of the 13th Australasian Database Conference (ADC(02), Melbourne, Australia 2002;5:3–10.

[5] Lingras P. Rough set clustering for web mining. In 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;2002:1039–44.

[6] Ng A, Smith KA. Web usage mining by a self-organizing map. In Proceedings of International Conference on Artificial Neural Networks In Engineering (ANNIE(00) Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining and Complex Systems 2000;10: 495–500.

[7] Srivastava J, Cooley R, Deshpande M, Tan PN. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations 2000;1(2):12–23.

[8] Wang X, Abraham A, Smith KA. Web traffic mining using a concurrent neuro-fuzzy approach. In Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, Computing Systems: Design, Management and Applications, Santiago, Chile 2002;2002:853–62.

[9] Cheung D, Kao B, Lee J. Discovering user access patterns on the world wide web. In Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD(97), vol. 10.; 1997. pp. 463–70.

[10] Chang G, Healey MJ, McHugh JAM, Wang JTL. Web minig. In Mining the World Wide Web—An Information Search Approach, Dordetch: Kluwer; 2001.

[11] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACMSIGMOD International Conference on Management of Data (SIGMOD(00), Dallas, TX, USA 2000;2000: 1–12.

[12] Jespersen SE, Thorhauge J, Pedersen TB. A hybrid approach to web usage mining. In Proceedings of the 4th International Conference (DaWaK(02) on Data Warehousing and Knowledge Discovery, Aix-en-Provence, France 2002;2002:73–82.

[13] Haravu, L.J. and A. Neelameghan. "Text Mining and Data Mining in Knowledge Organization and Discovery: The Making of Knowledge-Based Products." *Knowledge Organization and Classification in International Information Retrieval*. Ed. Nancy J. Williamson and Clare Beghtol. Binghamton, NY: Haworth, 2003.

[14] Hawwash, Basheer and Olfa Nasraoui. "Mining and Tracking Evolving Web User Trends from Large Web Server Logs." *Statistical Analysis and Data Mining*. Vol. 3 (2). Wiley Periodicals, Inc: MA. 03/11/ 2010. Pg. 106-125.
[15] Valitutti, Alessandro, and Carlo Strapparava. "Developing Affective Lexical Resources." *Psychology Journal* 2 (1): 2004. Pg. 61-83.
[16] Aggarwal C, Wolf JL, Yu PS. Caching on the world wide web. IEEE Trans Knowledge Data Engg 1999;11(1): 94–107.
[17] Margaret H.Dunham, S.Sridhar, "Data Mining: Introductory and Advanced Topics", *Pearson Education*
[18] Guru Jambheshwar University of Science & Technology website's available at http://www.gjust.ac.in/

AUTHOR's PROFILE

Shakti Kundu received his MCA from Kurukshetra University Kurukshetra, India in 2006, MPhil in Computer Science from Chaudhary Devi Lal University, Sirsa, Haryana, India in 2008, MTech in Computer Science & Engineering from Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India in 2010. Currently he is pursuing his Ph.D in Computer Science from Manav Bharti University, Solan, Himachal Pradesh, India. The author current research interests are web mining and web testing. He is individual member of Computer Society of India and life member of Indian Society for Technical Education (I.S.T.E), New Delhi.