Rummaging Around Workload Portrayal for Web Servers

Vishal Srivastava
AZAD IET, Lucknow
UP, INDIA

Raj Gaurang Tiwari
AZAD IET, Lucknow
UP, INDIA

Dr. R. A. Khan
Babasaheb Bhimrao Ambedkar
University, Lucknow,
UP, INDIA

Dr. Mohd. Husain
AZAD IET, Lucknow
UP, INDIA

ABSTRACT

Use of the Internet and the World Wide Web has improved swiftly over the past few years. Personals deploying Web servers crave to appreciate how their servers are being used by Internet users, how those patterns of use are changing over time, and what steps they should take to ensure adequate server response to the incoming requests today and in the future. This requires an evaluation of the requests offered to the Web server and the characteristics of the server's response to those requests over a suitably long time interval.

In this paper we present the results of a study of Web server system of http://www.lawetalnews.com for a six month period. In this duration the traffic to the website raised drastically in terms of incoming requests and outgoing bytes. We study the request and response types, and portray the traffic distribution on the basis of request size, response time, as well as some other factors. We wind up with system performance recommendations and identify future directions for our Web research.

General Terms

Measurement, Performance, Design, Experimentation

Keywords

E-commerce, performance evaluation, workload characterization

1. INTRODUCTION

The performance of any type of system cannot be determined without knowing the workload, that is, the requests being processed. This paper presents the workload characteristics of a busy World Wide Web Server. The purpose of this research is to obtain a better understanding of today's WWW traffic patterns and to set the stage for analysis of system resource utilization as a function of Web Server workload. By workload we mean both the request stream presented by clients (the work) as well as the server response to the requests (the load). Characterization involves determining and describing the fundamental character of the workload as presented over time[1]. Thus this paper presents the results of the Web workload characterization characteristics such as observed file types, file size distribution, the popularity of Web objects. One of the benefits of workload characterization is that it allows construction of analytical models of the workload and simulators that emulate client behavior in order to study the performance of similar systems in a controlled lab test environment. In this controlled environment it is possible to measure with greater accuracy the effects of the various types of user requests and thereby construct accurate models of server system resource utilization as a result of a given workload. A model of system resource utilization supports capacity planning for future deployments where the workload is known and helps developers assess trade-offs in application development based upon application and system workload data from the field. Developers and system managers can optimize system architecture and design in such environments to achieve optimal performance.

2. COLLECTION AND REDUCTION OF DATA

The data set used in this workload characterization study is composed of the access logs collected from the server of http://www.lawetalnews.com. Law et al. news is dedicated to dissemination of legal news and information with the mission to provide a quality reading experience to all lawyers, allied professionals and others interested to know the latest developments in the field of law - be it courts, bar, firms, corporate or academia. The entries of a web log file consist of several fields which represent the date and the time of the request, the IP number of the visitor's computer (client), the URI requested, the HTTP status code returned to the client, and so on. The web logs' file format is based on the so called "extended" log format, proposed by W3C [14]. The anatomy of a Log File is as below:

1. Internet provider IP address: This can be either lawetalnews.com or 69.167.154.158.
2. Identification field: This generally appears as a dash, "-"
3. AuthUser: This is an ID or password for accessing a protected area
4. Date, time, and GMT (Greenwich Mean Time): Thu July 17 12:38:09 1999
5. Transaction: Generally "GET" filename such as /index.html/products.htm
6. Status or error code of transaction: Usually 200 (success)
7. Size in bytes of transaction (file size): 3234 Additional Fields in the Extended Log Format
8. Referrer: search engine and keyword used to find your Web site, such as http://search.yahoo.com/bin/search?p=data+miningý/index.html
9. Agent: browser used by visitor, such as Mozilla/2.0 (Win95; I)
10. Cookie: .snap.com TRUE / FALSE 946684799 u_vid_0_000ed7085

In our research we used the access log of this website from 5/27/2010 20:25:00 - 9/11/2010 05:29:28. The web log is analyzed by Weblog Expert 7.0 Professional Edition. The general statistics is shown in table 1.

Table 1. General Statistics

Hits	
Total Hits	3,854,789
Visitor Hits	3,761,651
Spider Hits	93,138
Average Hits per Day	35,692
Average Hits per Visitor	189.34
Cached Requests	143,255
Failed Requests	2,278,187
Page Views	
Total Page Views	661,624

Average Page Views per Day	6,126
Average Page Views per Visitor	33.30
Visitors	
Total Visitors	19,867
Average Visitors per Day	183
Total Unique IPs	12,313
Bandwidth	
Total Bandwidth	17.60 GB
Visitor Bandwidth	13.78 GB
Spider Bandwidth	3.81 GB
Average Bandwidth per Day	166.83 MB
Average Bandwidth per Hit	4.79 KB
Average Bandwidth per Visitor	727.57 KB

3. TRAFFIC PATTERN ANALYSIS

The volume of traffic grows enormously. This marks the start of an extended flash-crowd. That is, the site suddenly became very popular, remained popular for a period of time, and then quickly faded back into obscurity. Although the daily traffic volume is quite bursty during the access of website, the traffic volume remains higher than it was at any time prior to the start of the event. During the analysis of weblog the impact of web spamming [7] has also been considered. Understanding spamming techniques is essential for evaluating the potency and flaw of work load portrayal.

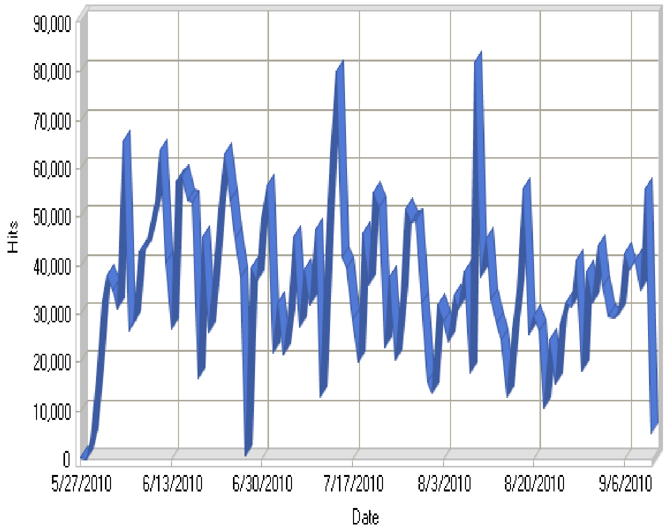

Figure 1. General Statistics of Daily hits

In order to better understand the causes of this burstiness we analyzed the traffic in more detail. Figure 1 shows the Daily traffic volume of the Web site. The blue curve depicts the number of hits per day. Analysis regarding file category and characterization is helpful in understanding the composition of Web workload. Some other factors like different browsers used, platforms used, different errors related to dynamic or static and web objects, daily url's access may be effect the workloads of Web servers[5]. Figure 3-8 depicts these facts for the www.lawetalnews.com.

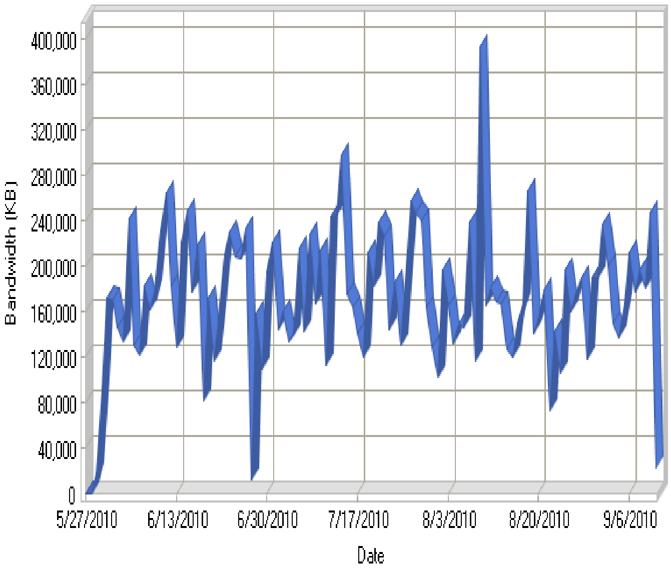

Figure 2. General Statistics of Daily Bandwidth

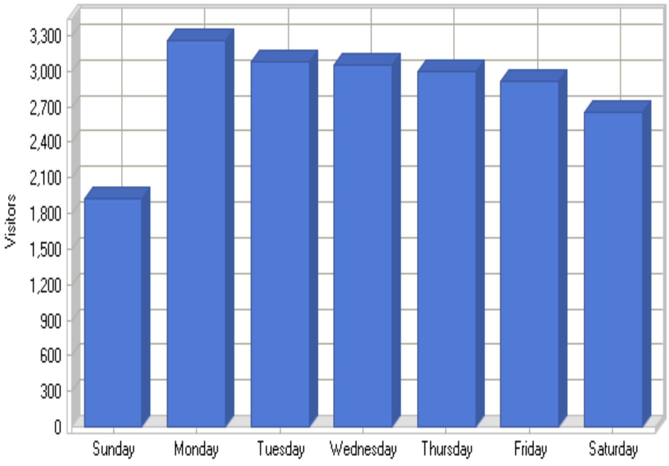

Figure 3. Visits By Day of Week

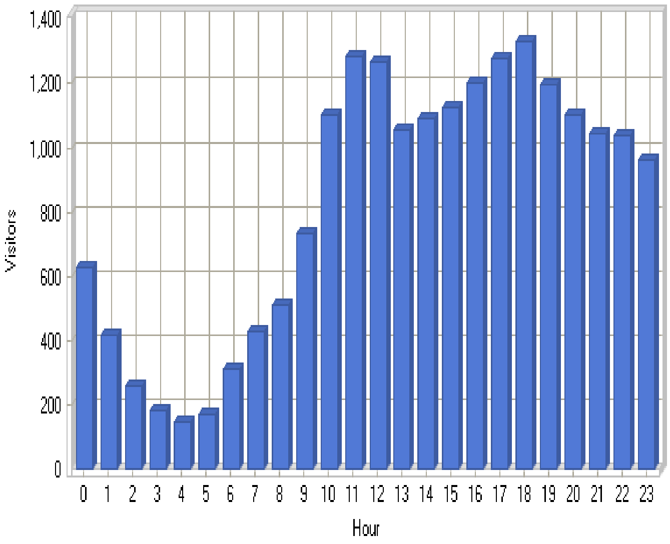

Figure 4. Visits By Hour of Day

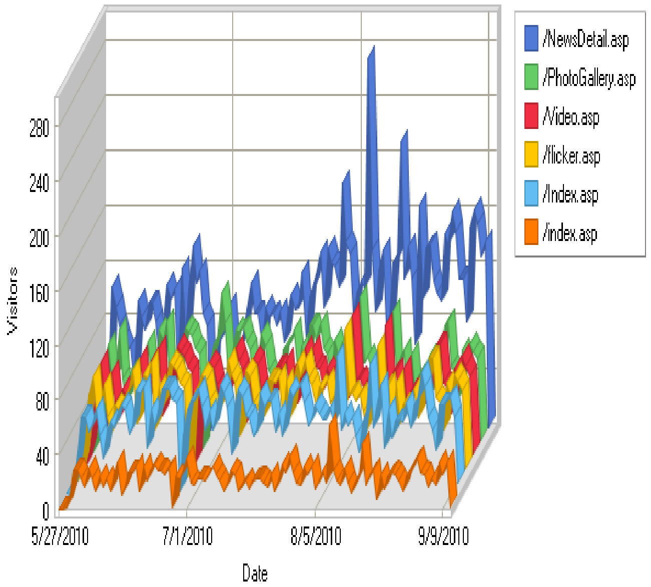

Figure 5. Daily Page Access

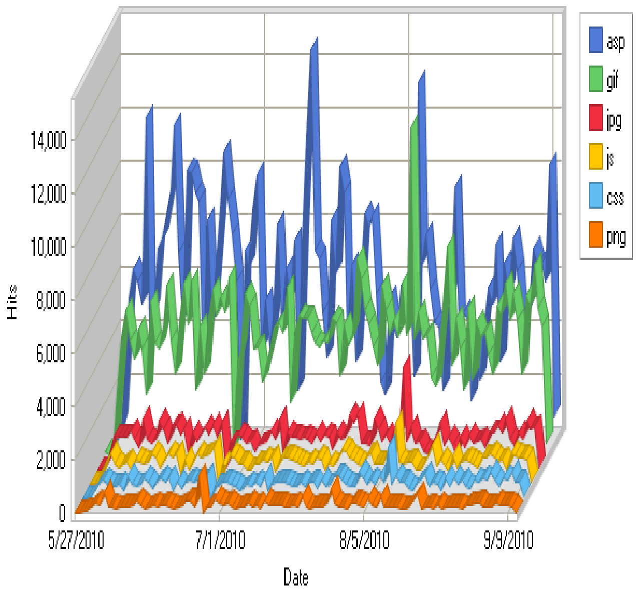

Figure 6. Daily File Type Access

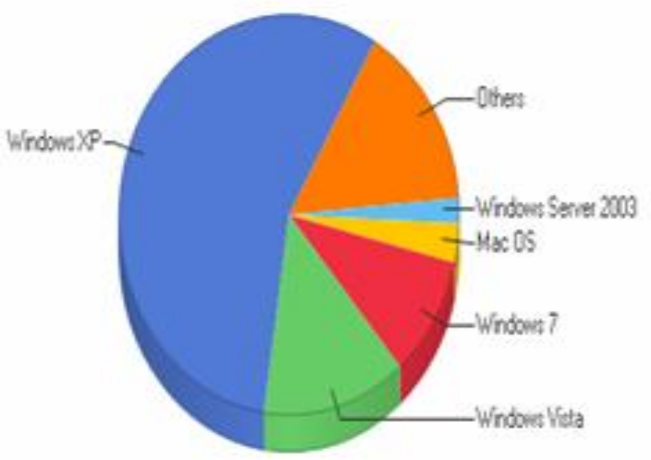

Figure 7. Top Platforms used in Law et al. News website

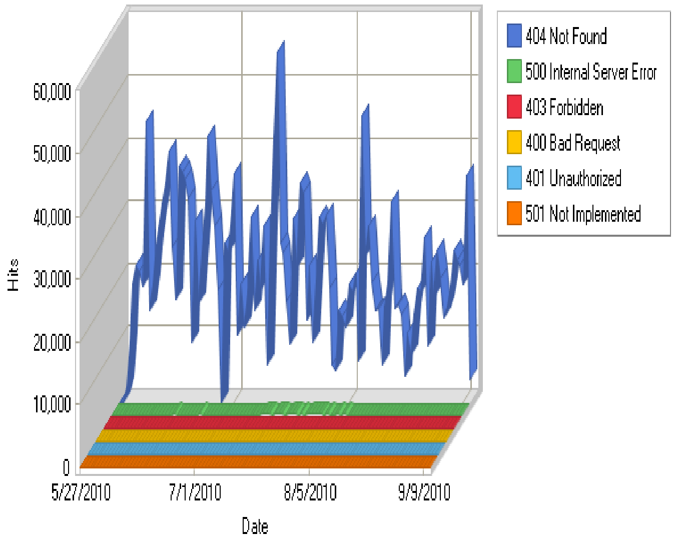

Figure 8. Error generated

Figure 3 plots the aggregate traffic by hour for each day of the week in the period studied. One interesting observation 4 is that the volume of traffic was quite equal on every day of week but on Monday some additional visits are made with the possible reasons some additional users getting information of legal news of whole week. Sunday is a very little slower days.

The Figure 4 depicts very small number of visits in late night to early morning because the target users belong to Asian sub continent. That's why as the sun rises the number of users grows.

Figure 5 shows the details of daily page access by visitors and Figure 7 shows the top operating platforms used by clients.

Figure 8 depicts the errors generated by server while response. As figure 5 shows the most common files, the popularity for the content at a site tends to follow a power law distribution; some studies suggest that content popularity specifically follows a Zipf distribution. In general a power law distribution is one of the form:

$$Y = x^a$$

Where a is a constant. In the case of the Zipf distribution the constant a is exactly -1. Put another way, Zipf's law predicts that the number of hits of a document (H) is related to the document's popularity rank (r) via the formula H = 1/r.

4. UNIQUE FILE SIZE DISTRIBUTION AND CHARACTERIZATION

The types of files at a Web site are closely related to the technology used to construct the Web site. Analysis regarding file category and characterization is helpful in understanding the composition of Web workload. In this section we analyze the distribution of sizes for all unique files requested from the Law et al. News Website. Our first analysis looks at the sizes of each of the unique files requested and successfully transferred at least once in the access log. For the purpose of this study we utilize the initial nonzero size recorded for each unique file. Thirty Eight Lacs fifty four thousand seven hundred eighty nine hits and six Lacs sixty one thousand six hundred twenty four files were requested (and successfully transferred) from the Law et al. news website during the measurement period. Bandwidth on Least Active Date is 81.35 KB and Bandwidth on Most Active Date is 400 MB. Most Active Hour of the Day 11:00 - 11:59 and least active hour of the day 04:00 - 04:59.

4.1 File Referencing Behavior

In this section we analyze the Law et al. News server workload for the presence of two important file referencing characteristics: temporal locality and concentration of references.

4.1.1 Temporal locality

Temporal locality means that a recently referenced file is likely to be referenced again in the near future[2]. To measure the temporal locality we utilize the standard Least Recently Used (LRU) stack-depth analysis. This investigation works in the following manner. When a file is primarily referenced it is placed to the top of the LRU stack (position 1). All files currently in the stack are pushed down by one position. When a file is referenced again its current depth (i.e., position) in the stack is recorded, and the file is shifted back to the zenith of the stack. The other files in the stack are pushed down as needed.

4.1.2 Concentration of References

The second file referencing characteristic on which we focus is concentration of references. Many studies, including [2], have found that a non uniform referencing pattern exists for files on the Web. This means that a small number of files on a Web site are extremely popular and receive most of the requests arriving at the site, while many unique files on a Web site are very unpopular.

4.2 File Type Characterization

Web workload characterization Characteristics such as observed file types, file size distribution, the popularity of Web objects, and request arrival processes are analyzed and compared to the results of previous studies. In addition, the request mix and requests through SSL are analyzed, as these also are workload characteristics associated with user behaviors and server resource usage. Previous studies [11, 12] showed that Image and HTML files accounted for 90% to 100% of requests. That also seems true in this study. There are two conflicting factors affecting the percentage of requests for images, particularly for Ecommerce-oriented Web sites. On one hand, E-commerce-oriented sites tend to use more small images for quick visual display. The average size for image files in Law et al. News -Log is only slightly larger than 1 kilo-bytes. The average transfer size for image files in Law et al News Log is also small. When the transfer size for images files decreases, more image files tend to be published on a Web page, increasing the percentage of requests for images about 45% of the requests in Law et al. News Log are for images. We have applied a logarithmic transformation to the file sizes to enable us to identify patterns across the wide range of values [8]. For a log2 transformation, bini includes values in the range $2i \le x < 2i +1 -1$. Similarly, for a loglo transformation, bini includes values in the range $10i \le x <$ l0i+1 - 1. In other Web workload characterizations the file size distribution has been found to be lognormal [3][4]. That is, after applying a logarithmic transformation to the data, the data appears to be normally distributed. We compare the unique file size distribution (the empirical data) to a synthetic lognormal distribution with parameters $\mu = 12.14$ and $\sigma = 1.73$.

Requests for dynamic pages are the most important part of the Web workload in the logs. Dynamic pages are necessary to make a database-driven Web site work since they are used to query databases and return the results to Web users. Most of the important E-commerce activities on a Web site, such as searching, selecting, and ordering goods/services, are related to dynamic pages. Law et al. news web Log shows that 47.42% dynamic pages are used. Changes in the technologies for building a dynamic page are also observed. CGI and Perl were the most popular technologies for implementing a dynamic Web page in the late 1990s. But in the logs used in this study ASP (Active Server Pages) are the most popular file types. Each log used in this study has a dominant dynamic file type. About 47.42 % of dynamic file in Law et al. News Log are .asp files.

Earlier we noted that much of the traffic to the Law et al. News web site came in large bursts that occurred while working days. In this section we analyze the busiest 59-min period from the overall one day workload. This period occurred from 11:00 until 11:59. on Aug 09, 2010.

5. WORK LOAD CHARACTERIZATION

On the basis of study workload characteristics and their performance implications are as follows:

- CSS and JavaScript files have become widely used. The percentages of requests for these types of files range from 3% to 16% in the log.
- Almost all important functions of the Web site were implemented using dynamic files. Due to the large numbers of requests for dynamic files, the characterization of dynamic files should be separated from that for static files.
- In spite of the effectiveness of client or proxy caches, a large percentage (60% to 90%) of incoming requests are still for images. The performance implication is that the server cache is still very necessary, in addition to effective client and proxy caches.
- Embedded images in the same Web page tend to be requested together. If the Web page is a popular one, such as the home page for a site, the amount of workload and overhead generated by requesting these images is high. The batched-referenced images in a popular Web page should be cached as a bundle so that a client needs only one request to receive all embedded images. Thus, the server can send all these images in one reply. In order to do this, the caching mechanism must have knowledge regarding the objects that need to be handled together.
- In general, the popularity of dynamic Web objects follows a Zipf-like distribution, indicating that caching would be beneficial for system performance.
- The incoming request stream is bursty. Knowing the composition of the incoming request stream is helpful for organizing the server resource. For example, at the peak of the requests for dynamic pages, the back-end database would receive the highest service demand. More resources should be assigned to it so that it does not become a bottleneck.
- The request mix is relatively stable when the time scale to measure it gets large enough. This stability indicates that customers are looking for similar services throughout the day. The request mix can be taken into account when allocating server resources to optimize performance [6]. For example, jobs can be scheduled in such a way that each request type gets its share of resource based on request mix. The stable request mix is also useful in forecasting workload. For example, assuming sales will increase by 50%, the volume of a specific request type can be predicted based on request mix.
- For E-commerce servers, most pages related to revenue generation are requested through SSL. To optimize server resource management with respect to revenue-generating page requests, priority should be granted to requests using SSL.
- It is observed that some images were processed through SSL to save the cost of moving back and forth between SSL and non-SSL use. This practice should be careful evaluated since the cost of SSL processing is high. The performance effect has not been quantified because the detailed data is not available.
- However if the recommender systems[13] is used by the E-commerce website which utilize the times of yore experiences and preferences of the target customers as a basis to offer personalized recommendations for them as well as resolve the information overloading hitch, workload expands exponentially. And without the recommender systems the potential of E-commerce can not be exploited.

6. CONCLUSION

The Web Server under study is a busy Internet site receiving a quarter thousands hits per day and experiencing increasing traffic on a week-to-week basis. During the workload characterization phase we have not examined system resource utilization in depth; but based upon the data that have mostly appeared in the server system was under extreme pressure for the duration of this study. We conclude this by examining the weekly traffic graphs, based on General Statistics of visits, Daily Spider Activity, Visits by Day of Week, Visits by Hour of Day, Browsers Used, Top Platforms used, Error generated, Daily URL Access, Daily URL Access. We conclude this by examining the daily and weekly traffic graphs: systems under severe request pressure tend to have a pronounced “flat top”, a period during which the maximum number of hits or bytes per time period remains consistent. A “flat top” can indicate either that the request or response traffic exceeds the available network bandwidth or that the mean arrival rate at the server is greater than the mean service rate. In the latter case, a server system or network bottleneck is limiting the service rate. This simple analysis does not give any insights on how to improve the performance for the end user, nor does it indicate the point at which a resource bottleneck will limit system performance in the future.

7. REFERENCES

[1] M. Arlitt T. Jin, ``Workload Characterization of the 1998 World Cup Web Site'', IEEE Network, Vol. 14, No. 3, pp. 30-37, May/June 2000.

[2] M. Arlitt C. Wikamson, "Internet Web Servers: Workload Characterization and Performance Implications," IEEE/ACM Trans. Net., vol. 5, no. 5, Oct. 1997, pp. 631-45.

[3] P. Barford et al., "Changes in Web Client Access Patterns," World Wide Web J., Special Issue on Characterization and Performance Evaluation, 1999.

[4] M. Arlitt, R. Friedrich, and T. Jin, "Workload Characterization of a Web Proxy in a Cable Modem Environment," ACM SlGMETRlCS Perf. Eva/. Rev.,vol. 27, no. 2, Aug. 1999, pp. 25-36.

[5] K. Thompson, G. Miller, R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," IEEE Nehvork, vol. 1 1, Nov./Dec. 1997, pp. 1 C-23.

[6] Bryan Veal , Annie Foong, “Performance scalability of a multi-core web server”, Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems, December 03-04, 2007, Orlando, Florida, USA

[7] Mohd. Husain, Raj gaurang Tiwari, Vishal Srivastava, “Convalescing PageRank Score Using Optimal Spam Farm Structure” in proceedings of International Conference on 'Challenges and Applications of Mathematics in Science and Technology (CAMIST 2010), held at National Institute of Technology(NIT), Rourkela, Orissa, India on Jan 11-13, 2010. ISBN No. 023-032-875-X, pp-553-561.

[8] V. Paxson, "Empirically-Derived Analytic Models of Wide-Area TCP Connections," IEEE/ACM Trans. Net., vol. 2, no. 4, Aug. 1994, pp. 31 6-36.

[9] M. Crovella, M. Taqqu, "Estimating the Heavy Tail index from Scaling Properties," Methodology and Computing in Applied Probability, vol. 1, Nov. 1, 1999.

[10] H. Yu, L. Breslau, S. Shenker, "A Scalable Web Cache Consistency Architecture," Proc. ACM SIGCOMM '99, Cambridge, MA, Sept. 1999. [l 11 J. Mogul et al., "Potential Benefits of Delta Encoding and Data Compression far HTTP," Pm. ACM SlGCOMM '97, Cannes, France, Sept. 1997, pp. 181-94.

[11] M. Arlitt, C. Williamson, “Web Server Workload Characterization: The Search for Invariants”, In Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pages 126{137, Philadelphia, PA, USA, May 1996.

[12] A. Oke. Workload Characterization for Resource Management at World Wide Web Servers, Msc. Thesis, University of Saskatchewan, Saskatoon, SK, Canada, April 2001.

[13] Raj Gaurang Tiwari, Mohd. Husain, Anil Agrawal, Vishal Srivastava, “Multidimensional Recommendation: A Step towards the Next Generation of Recommender Systems”, in proceedings of 2nd International Symposium on Emerging Trends and Technologies In Libraries And Information Services (ETTLIS 2010), held at Jaypee University of Information Technology, Solan, HP, on June 3-5,2010, ISBN 81-9079991-6, pp-237-242

[14] Extended Log File Format, http://www.w3.org/TR/WD-logfile.html

Author Biographies

Mr. Vishal Srivastava is pursuing Ph. D. in Computer Science from Dravidian University. He received his Masters degree in Computer Applications from VBS Purvanchal University Jaunpur in 2002. Currently he is working as Lecturer at AZAD Institute of Engineering and Technology, Lucknow, India. His research interest includes E commerce. He authored more than 10 International and national journal and conference papers

Mr. Raj Gaurang Tiwari is pursuing Ph. D. in Computer Science from Dravidian University. He received his Masters degree in Computer Applications from Dr. B. R. Ambedkar University, Agra in 2002 and Masters degree in Computer Sc. and Engg. From Gautam Buddh Technical University, Lucknow in 2010. Currently he is working as Assistant Professor at AZAD Institute of Engineering and Technology, Lucknow, India. His research interests are Knowledge-Based Engineering and Web Engineering. He authored more than 35 International and national journal and conference papers.

Dr. R. A. Khan is working as Reader and Head in Dept. of Information T echnology, School for Information Sc. & Technology, Babasaheb Bhimrao Ambedkar University, (A Central Govt. University), Lucknow. He received his Masters degree in Computer Applications from Punjab Technical University, Jalandhar in 2000 and & Ph.D Degree from Jamia Millia Islamia University in 2004. He has more than 9 years teaching and research experience in the field of Software Engineering. He has published more than 72 International and National publications

Prof (Dr.) Mohd. Husain is working as Director at AZAD Institute of Engineering and Technology, Lucknow, India. He got his Masters Degree from UP Technical University & Ph.D Degree from Integral University in 2007. He has more than 12 years teaching experience and 10 years research experience in the field of Data mining. He has published more than 65 International and National publications