International Journal of Advanced Research in IT and Engineering

ISSN: 2278-6244

DISCOVERING WORKLOAD CHARACTERIZATION OF E COMMERCE WEB SERVER: A SURVEY

Vishal Srivastava*

Dr. Mohd. Husain**

Abstract: *Workload characterization of E commerce web server is an important area to analyze E-commerce web server workload. Workload characterization have significant performance implications with regard to capacity planning and server performance. It can be classified into four different types i.e. Web sites studied at request level, function level, resource level and session level. The aim of this paper is to provide past, current evaluation and update in each of the performance implications.*

This paper also reports the comparisons and summary of various methods and technique of Workload characterization with applications, which gives the overview of development in research and some important research issues.

Keywords: *E-commerce, performance evaluation, workload characterization*

*AZAD IET, Lucknow, India.

**Jahagirabad Institute of Technology, Lucknow, India

1. INTRODUCTION

The use of the Internet has made automatic knowledge extraction from various web log files a necessity. This can be used to improve the effectiveness of the web sites by adapting the information structure of the sites to the user behavior. The data set used in this workload characterization study is composed of the access logs collected from the e commerce web server. Web log consists of a series of entries arranged in reverse chronological order, often updated on frequently with new information about particular topics. The performance of any type of system cannot be determined without knowing the workload, that is, the requests being processed. This paper Understanding these aspects of Web characterization efforts.

Characterization involves determining and describing the fundamental character of the workload as presented over time[1]. 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). Research that analyses client, proxy and server traffic will be examined followed by a review of attempts to characterize the entire World Wide Web. The literature is reviewed in historical order within each category. This is followed by a discussion of the regularities that have been identified across studies and categories.

Service providers can now clearly recognize user visiting patterns to their sites and pages, and hence can reorganize their site structure as per the interests exhibited by their users. Capturing user's navigation pattern in particular pages; site owners can easily get implicit ratings about their pages and use such information in page reorganization or relocation. Moreover, with the advancement in Automatic Web Navigation [36], with the help of web usage mining it is possible to discover user access patterns from the log files which are create on web servers which are helpful to workload characterization.

2.0 WORKLOAD CHARACTERIZATION

Workload characterization has a significant impact on performance evaluation. Understanding the nature of the workload and its intrinsic features can help to interpret

performance measurements and simulation results better. Research on improving Web performance must be based on a solid understanding of Web workloads.

Workload characterization can also be used to validate trace reduction and trace sampling techniques used in performance evaluation [1, 2]. Eventually, workload characterization should lead to a program behavior model, which can be used in conjunction with a processor model to do analytical performance modeling of computer systems. The performance of a computer system for a workload is determined by the system itself (hardware and software) and the characteristics of the application.

Throughout the study, emphasis is placed on finding workload characteristics that are common across all the data sets studied. These characteristics are deemed important, since they potentially represent universal truths for all Internet Web servers. Our research has identified such characteristics for Web server workloads. Web server workload characterization can be performed at many levels including:

Levels	Description
Request level	**Request arrival process, file type, file popularity, file size distribution, and other aggregated workload features are characterized.**
Function level	**The functions provided at a Web site are analyzed**
Resource level	**The usage of the system resources is analyzed.**
Session level	**The client sessions are identified and characterized.**

Most published studies on workload characterization for Web servers are at the request level [3]. At request level characteristics of Web workloads for traditional information Web servers. The main results are as follows:

File type: Requests for HTML, Dynamic file, image files, Com components, ActiveX Components etc. of the traffic, indicating that the Web sites rarely used dynamic pages.

File size: The mean transfer size is less than 21 kilobytes, indicating that most files are small.

File size distribution: The file size distribution is heavy-tailed. The existence of files with a very large size is related to the heavy-tailed property.

File popularity: This property can be informally interpreted to indicate that if the Web documents are ranked by popularity, the number of references to a document will be inversely proportional to its rank.

One-time referencing: Approximately one-third of the files and bytes in the logs are accessed only once.

Self-similarity: Arlitt and Williamson also discussed the self-similarity of World Wide Web traffic, which had been characterized earlier by Crovella and Bestavros [4].

Inter-reference time: File inter-reference times follow exponential distribution.

At the function level, the functions provided at a Web site are analyzed. Analysis at this level is seldom performed on workload of traditional Web servers for information providers, since there were not many functions available.

At the resource level, the usage of the system resources is analyzed. There are many resources in a system, such as CPUs, memory, disks, caches, I/O, and network bandwidth.

At the session level, the client sessions are identified and characterized. A session consists of a sequence of requests from the same client during a visit to the Web site. A session is complete if, after receiving a request from a client, the server does not receive any more requests from the same client within a threshold time (this is called the session timeout threshold).

These results are based on Web servers for information providers and are based mostly on HTTP logs collected in the early years of Web servers, the period from the early to the mid-1990s.

Some more recent studies on Web workloads, in particular, on workloads to E-commerce servers, provide more updated characterization request, function, resource, and session levels.

The analysis of resource usage in an E-commerce system is not an easy task. The workload data on application servers and back-end database servers is more difficult to obtain than that on Web servers. There are many resources in an E-commerce system.

SUMMARY OF WEB SERVER WORKLOAD CHARACTERISTICS

Fundamental to the goal of improving Web performance is a solid understanding of www workloads. Most studies present data from only one measurement site, making it difficult to

generalize results to other sites. Furthermore, some studies focus on characterizing Web clients and Web proxies, rather than Web servers.

Workload Characteristic	Description
1. Successful Requests	About 65-70% of requests to a Web server result in the successful transfer of a document.
2. Document Types	HTML and image documents together account for 70-85% of the documents transferred by Web servers.
3. Transfer Size	The median transfer size is small
4. Distinct Requests	Small fractions (about 1%) of server requests are for distinct documents.
5. One-time Referencing	A significant percentage of files (15-26%) and bytes (6-21%) accessed in the log are accessed only once in the log.
6. File Size Distribution	The file size distribution and transfer size distribution are heavy-tailed.
7. Concentration	The busiest 10% of files account for approximately 80-90% of requests and 80-90% of bytes transferred
8. Inter-Reference Times	The times between successive requests to the same file are exponentially distributed and independent.
9. Remote Requests	Remote sites account for 70% or more of the accesses to the server, and 80% or more of the bytes transferred.
10. Wide-Area Usage	Web servers are accessed by hosts on many networks, with 10% of the networks generating 75% or more of the usage.

Table 2: Summary of Web Server Workload Characteristics

These results are based on Web servers for information providers and are based mostly on HTTP logs. Some more recent studies on Web workloads, in particular, on workloads to E-commerce servers, provide more updated characterization request, function, resource, and session levels.

A session has many attributes and can be represented in many ways. To group sessions, one must select one or more session attributes to represent a session and use an algorithm for clustering. The selection of session representations and clustering algorithms has been mainly dependent on how the resulting session group would be applied to ensure a successful shopping experience for customers. In previous studies, session groups have been used either for performance analysis or for Web usage mining.

3.0 ROLE OF CLUSTERING OF SESSIONS IN PERFORMANCE ANALYSIS

In order to improve the Web performance and better serve the user needs, a solid understanding of user sessions is essential. A session is defined as a sequence of requests from the same user during a single visit to the Web site.

Summary of related Work

Author	Prior Studies	Publication Year
Felix Hernandez et al. *1	They observe that Web usage by both content providers and Web clients has significantly evolved. Continuous monitoring of Internet traffic to track its evolutionary patterns.	**2003**
Cherkasova et al. *2	The authors speculate that the differences arise from Web server side performance improvements, available Internet bandwidth, and a greater proportion of graphical content on Web pages.	**2001**
M. Arlitt. et al.	The author studied the effect of a wide range of session timeout values on numerous user session characteristics and showed how these characteristics can be utilized in improving Web server performance.	**2000**
Barford *et al.*	They conclude that document size distributions did not change over time, though the distribution of file popularity did. Their analysis was only for Web client workloads rather than Web server workloads.	**1999**
Martin F. Arlitt et al.	The workload characterization focuses on the document type distribution, the document size distribution, the document referencing behavior, and the geographic distribution of server requests.	**1997**

Arlitt [5] carried out a session level workload characterization on the 1998 Football World Cup site. The session length, duration, and other factors were discussed.

Oke [6] studied a Web server access log collected in June 1998 from a busy commercial Internet Web site. At the request level, some workload characteristics were very close to those for Web sites of information providers [6], although the Web site was an E-commerce site.

A workload characterization could be also used to create synthetic workload [7], in order to benchmark a site, a monitoring agent checks the resource usage metrics during the test to

search for system and network bottlenecks. In [8] two important models characterizing user sessions are introduced: Customer Behavior Model Graph (CBMG) and
Customer Visit Models (CVM).

In [9] K-means algorithms are used to obtain clusters of CBMG with similar patterns. A CVM is a more compact model than the CBMG. It represents sessions as a collection of session vectors, one per sessions.

The k-means clustering algorithm was applied to group sessions. In this algorithm,

i) a session is considered as a point in a virtual space;
ii) k points in the space are selected as estimated centroids of the k clusters; and
iii) the remaining points are grouped to the cluster with the nearest centroid.

Arlitt et al. [10] characterized E-commerce workload based on the level of demand on resources. The motivation was to study the scalability. Requests were classified into roughly three classes: cacheable, non-cacheable, and search.

4.0 CHALLENGES ON WEB.

The Web is based on the client-server model. Communication is always in the form of request-response pairs, and is always initiated by the client. The boom in the use of the World Wide Web WWW) and its exponential growth are well known facts nowadays. Just the amount of textual data available is estimated in the order of one terabyte. Some challenges faces are:

- Unstructured and heterogeneous
- Multimedia
- Size + rapid growth
- 1 new server every 2 hours
- Dynamic
- Networked/distributed

Web data is Web content (text, image, records, etc.), Web structure (hyperlinks, tags, etc.), Web usage (http logs, app server logs, etc.). Web servers use the log files to record an entry for every single access they get. As the complexity of the web site or application increases, simple statistics give no meaningful hints on how the web site is being used. Web mining refers to the application of such techniques to web data repositories, to enhance the

analytical capabilities of the known statistical tools. An early taxonomy of web mining has been proposed in [11]. Mining the web data is one of the most challenging tasks for the data mining and data management scholars because there are huge heterogeneous, less structured data available on the web and we can easily get overwhelmed with data [12].

4.1 Resource Discovery for Characterized Web Server Workload

Workload characterization is the basis for studies on server performance. Web servers can be configured to record (in an access log) information about all of the requests and responses processed by the server [13]. Each line from the access log contains information on a single request for a document. The log entry for a normal request is of the form:

Hostname- - -[dd/mmm/yyyy:hh:mm:ss tz) request status bytes

The access log provides most of the data needed for workload characterization studies of, web servers. However, they do not provide all of the information that is of interest. For example, the log entries tell only the number of bytes transferred for a document, not its actual size, there is no record of the elapsed time required for a document transfer; and there is no information on the complete set of files available on the server, other than those documents that are accessed in the logs. Furthermore; there is no record of whether a file access was human-initiated or software-initiated (e.g., by a Web crawler?), or what caching mechanisms, if any, are in place at the client and/or the server. These issues are outside the control of our study: our focus is solely on characterizing the workload seen by a typical Internet Web server in its de facto configuration.

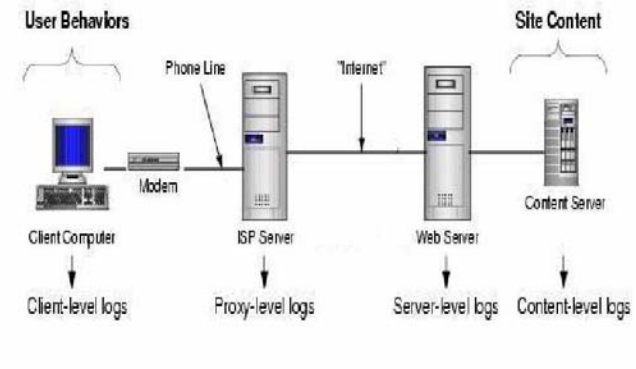

Figure 1: Various data sources

There are many kinds of data that can be used in Web Mining. Such data classifies into the following types

- **Content:** The real data in the Web pages, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited to, text and graphics.
- **Structure:** Data which describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure, where the (html) tag becomes the root of the tree. The principal kind of inter-page structure information is hyper-links connecting one page to another.
- **Usage:** Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses.
- **User Profile:** Data that provides demographic information about users of the Web site. This includes registration data and customer profile information.

The phenomenal growth in Web traffic has led to many performance problems, which in turn has resulted in much research activity on "improving" the World-Wide Web. The overall performance of the Web is determined by the performance of the components which make up the Web: the clients, the servers, the proxies, the networks, and the protocols used for communication.

Improving the performance of Web servers is vital to the goal of reducing response times. Web server activity, to help identify performance bottle- necks [14, 15] evaluates the performance impact of different Web server designs. Other researchers have studied the use of tile caching in reducing Web server loads.

Web proxies are useful for reducing response times and network traffic [16]. Various cache replacement policies for Web proxies are useful for workload characterization for Web servers, [17], [18], and [19].

5.0 ROLE OF WEB MINING FOR CLUSTERING SESSIONS

One of the main goals of workload characterization is to extract workload properties and use them to construct a workload model, while the purpose of Web usage mining is to discover Web usage patterns to better serve customers. However, both session level

workload characterization and Web usage mining follow similar procedures to process data and use the same clustering techniques to group sessions.

The results from Web usage mining can be applied in roughly two ways one is learning user profile in order to build adaptive (or personalized) servers and the other is learning user navigation patterns for system improvement or site reorganization or modification [20].

Web is a collection of inter-related files on one or more Web servers. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.
- Web graph, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

Web mining refers to the application of such techniques to web data repositories, to enhance the analytical capabilities of the known statistical tools. Web mining involves three tasks: An early taxonomy of web mining has been proposed:

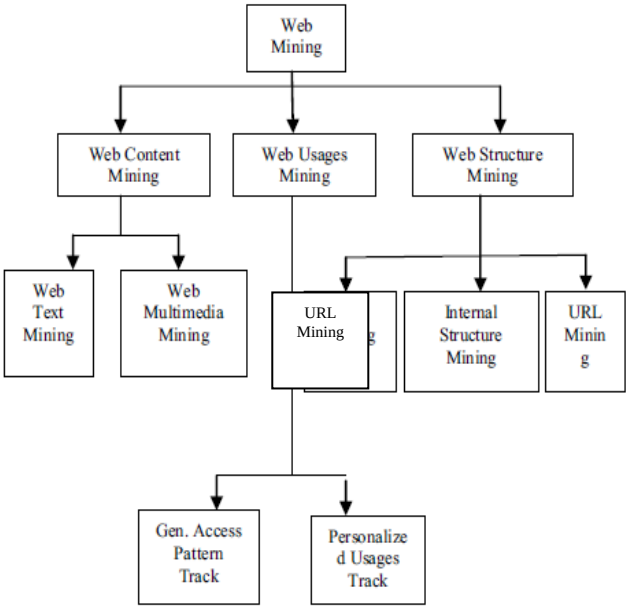

Figure-2: Taxonomy of Web Mining

(1) Structure Mining, use of the hyperlink structure of the Web as an (additional) information source

(2) Content mining: application of data mining techniques to unstructured or semi-structured data, usually HTML-documents.

(3) Usage mining: analysis of user interactions with a Web server (e.g., click-stream analysis)

5.1 Web Usage Mining

Web usage mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Most of this information is usually generated automatically by Web servers and collected in server access logs. Other sources of user information include *referrer logs* which contains information about the referring pages for each page reference, and user registration or survey data.

Web usage mining techniques improve the effectiveness of the web sites by adapting the information structure of the sites to the user behavior. The ease and speed with which business transactions can be carried out over the Web has been a key driving force in the rapid growth of e-commerce. Specifically, e-commerce activity that involves the end user is undergoing a significant revolution. The ability to track user browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before. Service providers can now clearly recognize user visiting patterns to their sites and pages, and hence can reorganize their site structure as per the interests exhibited by their users.

Most data used for mining [20] is collected from Web servers, clients, proxy servers, or server databases, all of them produce noisy data. Because Web mining is sensitive to noise, data cleaning methods are necessary. Categorize data preprocessing into subtasks and noted that the final outcome of preprocessing should be data that allows identification of a particular user's browsing pattern in the form of page views, sessions, and click streams. Click streams are of particular interest because they allow reconstruction of user navigational patterns.

The web usage mining generally includes the following several steps: data collection, data pretreatment, and knowledge discovery and pattern analysis.

a) Data collection*:*

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

b) Data preprocessing:

Some databases are insufficient, inconsistent and including noise. The data pretreatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pretreatment work, mainly include data cleaning, user identification, session identification and path completion.

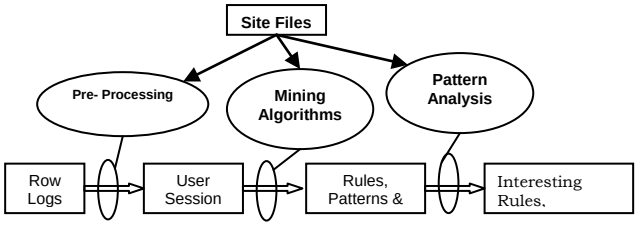

Figure 4: Preprocessing of Web Usage Data

i) Data Cleaning:

The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed:

1. The records of graphics, videos and the format information the records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record.

2. The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed.

It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

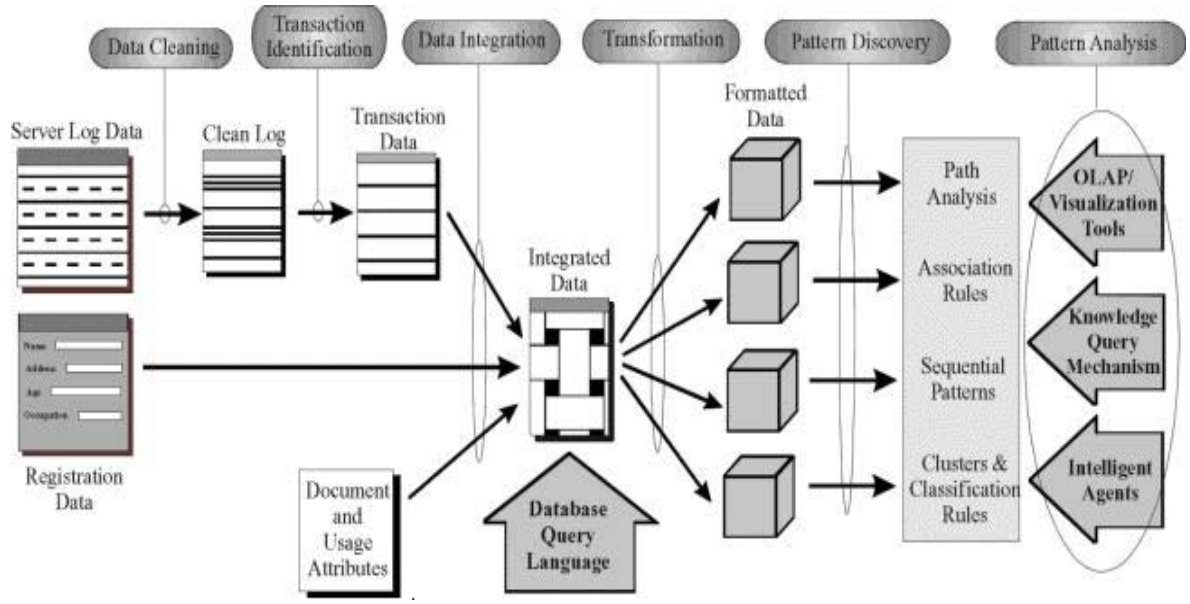

Figure 3: General Architecture for Web Usage Mining

ii) User and Session Identification:

The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The rules adopted to distinguish user sessions can be described as follows:

a. The different IP addresses distinguish different users;

b. If the IP addresses are same, the different browsers and operation systems indicate different users.

c. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account.

d. Session identification is done by using the time stamp details of the web pages. The total time used by each user of each web page. Session is the time duration spent in the web page.

iii) Path Analysis: Graph models are most commonly used for Path Analysis. A graph represents some relation defined on Web pages and each tree of the graph represents a web site. Each node in the tree represents a web page (html document), and edges between

trees represent the links between web sites, while the edges between nodes inside a same tree represent links between documents at a web site.

c] Knowledge Discovery: Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At resent the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

d] Pattern Discovery: Techniques adopted for this phase, strictly depend on the aim of the analysis. Methods available draw upon several fields such as: statistics, data mining, machine learning, and pattern recognition.

- Statistical analysis is in general applied to discover information such as most accessed pages or average length of a navigation path through a web site.
- Clustering is used to group items having similar characteristics. In this case clustering may be used to group users exhibiting similar navigation behavior (usage clusters) or groups of pages having a related content (page clusters).
- Classification techniques are often used to associate navigation behaviors to groups of users (or profiles)
- Sequential pattern discovery techniques to identify set of items followed by further items in a time ordered sequence.
- Dependency modeling is also used with the goal of developing a model to represent significant dependencies among various variables in the web (for instance, modeling the stages a user undergoes during a visit to an on-line store).

The results from Web usage mining can be applied in roughly two ways

- One is learning user profile in order to build adaptive (or personalized) servers and
- The other is learning user navigation patterns for system improvement or site reorganization or modification.

Session Group Identification and Characterization

Understanding the patterns and characteristics of incoming web request streams to a Web server is necessary for analyzing server resources and for evaluate server performance.

The session-level characterization results indicate that session groups can be identified for a Web site. The characterization of the session groups can be used to improve server performance, implement Web site personalization, and improve resource management.

It is observed that the session clustering results obtained by session representations Pages Requested, Navigation Pattern, and Resource Usage are similar.

The information on Web pages requested is available in any HTTP log. Clustering by Pages Requested works well in identifying session groups in all cases.

Author	Method	Application	Publication Year
Jaideep Srivastava, R. Cooley[29]	Statistical Analysis Association Rules	Personalization Site Modification etc	2000
Jianhan Zhu et al[31]	Clustering algorithm called Citation Cluster	Construct a conceptual hierarchy of the Web site	2002
Borges and M. Levene[32]	Dynamic clustering-based method	Representing a collection of user web navigation sessions	2004
TAN Xiaoqiu, YAO Min et al[33]	Improved WAP tree	Sequential pattern mining	2006
Yu-Hui Tao , Tzung-Pei Hong et al[34]	Taxonomy of browsing data	Decision support	2007
Mehdi Hosseini et al[16]	Web based recommender systems	predict user's intention and their navigation behaviors	2008
Mehrdad Jalali et al [36]	Longest common subsequences algorithm	Predict user near future movement.	2009
M. Jalali, et al[37]	WebPUM	Predict user near future Movement	2010

6. CONCLUSION

Several Studies have been published regarding the workload of information provider sites. However, very few studies are available for E-commerce sites. This paper used a chronological approach for workload characterization of ecommerce sites. The

characterization was done at the session, e commerce function and request levels. This paper show how we analyzed web server workload by web usage mining technique. The characterization of the session groups can be used to improve server performance, implement Web site personalization, and improve resource management. This paper focus on development of techniques which would improve automatic Web navigation and page pre-fetching on the behalf of the users and establishment of more robust procedures for reconstructing user behavior patterns while visiting Web sites and pages.

6. REFERENCES

[1] V. S. Iyengar, L. H. Trevillyan, P. Bose,Representative Traces for Processor Models with Infinite Cache, Proceedings of HPCA-2, 1996

[2] Lizy Kurian John, Purnima Vasudevan and Jyotsna Sabarinathan " Workload Characterization: Motivation, Goals and Methodology" IEEE, Nov 1998.

[3] J. Pitkow. Summary of WWW Characterizations. World Wide Web, 1999.

[4] M. Crovella and A. Bestavros. Self-similarity in World Wide Web Traffic: Evidence and Possible Causes. IEEE/ACM Transactions on Networking (TON), 5(6):835{846, 1997.

[5] M. Arlitt. Characterizing Web User Sessions. ACM SIGMETRICS Performance Evaluation Review, 28(2):50{63, 2000.

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

[7] P. Barford and M. Crovella, Generating Representative Web Workloads for Network and *Server Performance Evaluation*. In Proc. of ACM SIGMETRICS, 1998.

[8] D.A. Menascé, V. A.F. Almeida, R. Fonseca, M. A. Mendes Scaling for E-business: Technologies, Models and Performance and Capacity Planning, Prentice Hall, NJ, May, 2000.

[9] D.A. Menascé, V. A.F. Almeida, R. Fonseca, M.A. Mendes *A Methodology for Workload Characterization of Ecommerce Sites*, in Proc. Of ACM Conf. on E-Commerce, Denver, CO, Nov. 1999.

[10] M. Arlitt, D. Krishnamurthy, and J. Rolia. Characterizing the Scalability of a Large Web-based Shopping System. ACM Transactions on Internet Technology (TOIT), 1(1):44{69, 2001.

[11] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In International Conference on Tools with Artificial Intelligence,pages 558–567, Newport Beach, 1997. IEEE.

[12] Qingyu Zhang and Richard s. Segall," Web mining: a survey of current research, Techniques, and software", in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) 683–720

[13] National Center for Supercomputing Applications, "NCSA-httpd," 1994,

[14] J. Almeida. V. Ahneida, and D. Yates, "Measuring the behavior of a World-Wide Web server," in Proc. 7th Conf. High Performance Networking (HPN), White Plains, NY, Apr. 1997,pp. 57-72

[15] N. Yeager and R. McGrath, Web Server Technologies: The Advanced Guide for World Wide web information Providers, San Francisco. CA:Morgan Kaufmann, 1996.

[16] E. Markatos, 'Main memory caching of Web documents." in Electron.Proc. 5th World Wide Web Conf. Paris. France, May 6-10. 1996.

[17] M. Abrams, C. Standridge, G. Abdulla. S. Williams. and E. Fox,"Caching proxies: Limitations and potentials," in Electron. Proc. 4th World Wide Web Conf'95: T Web Revoiurion, Boston, MA, Dec.11-14. 1995.

[18] J. Bolot and P. Hoschka. "Performance engineering of the World-Wide Web: Application to dimensioning and cache design: in Electron. Proc.4th World Wide Web Conf'Paris, France, May 6-10, 1996.

[19] S. Williams, M. Abrams, C. Standtidge, G. Abdulla, and E, Fox,"Removal policies in network caches for World-Wide Web documents," in Proc. ACM SIGCOMM'96, Stanford, CA, Aug. 1996, pp. 293-305,

[20] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Web mining: Information and pattern discovery on the world wide web. In International Conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, 1997. IEEE.