Empirical Study of Session–based Workload and Reliability for Web Servers

Katerina Goševa-Popstojanova, Sunil Mazimdar, and Ajay Deep Singh
Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV 26506-6109
{katerina, mazimdar, ajays}@csee.wvu.edu

Abstract

The growing availability of Internet access has led to significant increase in the use of World Wide Web. If we are to design dependable Web–based systems that deal effectively with the increasing number of clients and highly variable workload, it is important to be able to describe the Web workload and errors accurately. In this paper we focus on the detailed empirical analysis of the session–based workload and reliability based on the data extracted from actual Web logs of ten Web servers. First, we address the data collection process and describe the methods for extraction of workload and error data from Web log files. Then, we introduce and analyze several intra–session and inter–session metrics that collectively describe Web workload in terms of user sessions. Furthermore, we analyze Web error characteristics and estimate the request–based and session–based reliability of Web servers. Finally, we identify the invariants of the Web workload and reliability that apply through all data sets considered. The results presented in this paper show that session–based workload and reliability are better indicators of the users perception of the Web quality than the request–based metrics and provide more useful measures for tuning and maintaining of the Web servers.

1 Introduction

The World Wide Web is arguably the biggest distributed system ever built. In fact, for many people, when they talk about Internet, they really mean the Web. The WWW is essentially a huge client-server system with millions of clients and servers distributed worldwide. So far the number of Web clients and servers has grown with exponential rate. Even more, common to many Web sites is the exponential rate of increase in number of requests, bytes transferred, number of files and number of bytes on the site. The most astounding fact is that the World Wide Web traffic, which in December 1992 was almost nonexistent, today is the one of the dominating components of the Internet traffic [2].

In the past few years, the popularity of e-commerce sites and large-scale Internet infrastructure services (e.g., AOL, Google, Yahoo, and Hotmail) has grown enormously. Despite the success of these services, their architectures - hardware, software, and operational - have developed in an ad hoc manner that few have surveyed or analyzed [13]. Users increasingly see large–scale Internet services as essential to the world's communication infrastructure and demand 24/7 availability and response time within seconds. However, they very often experience long delays on the Web that may be attributed to server or network failures, overloaded servers, or network congestion. The economic loss because of unavailability due to failures or poor performance is estimated to be in the range of billions dollars per year in United States alone [10].

With the tremendous growth and change in Web sites, users, and technology, expanding usage in different application domains, and high consequences of failures and poor performance, a comprehensive analysis and prediction of Web quality attributes is mandatory. In this paper we introduce and empirically analyze session–based workload and reliability and argue that they are better indicators of the users perception of the Web servers quality.

The rest of this paper is organized as follows. Related work and the contributions of this paper are discussed in section 2. The data extraction and analysis process are briefly described in section 3. Session–based workload is analyzed in section 4. Then, we present the analysis of request–based and session–based error behavior in sections 5 and 6, respectively. The invariants of the Web workload and reliability that are common to all ten data sets are summarized in section 7. Finally, the concluding remarks are given in section 8.

2 Related work and our contributions

The Web is based on client–server model and the communication is always in the form of request and response. The information about all requests and responses is stored in the Web server logs. Considerable amount of research work in the past has been focused on studying the characteristics of the Web traffic on per request basis. Thus, in [2] the access logs from six different Web servers were examined and emphasis was placed on finding request–based workload characteristics that are common to all data sets. Workload characterization was focused on the document type distribution, document size distribution, document referencing behavior, and geographic distribution of server requests. In [6] it was shown that the WWW transfers from the actual Web logs are consistent with self–similarity notion, characterized by burstiness and heavy tail distributions. Similar findings were reported in the recent study of the end–to–end response times required to download Web pages from a set of well–known Web sites [5].

A unique characteristic of Web workload is the concept of session which was first introduced as a unit of Web workload in [4] and further addressed in [3], [9], [10]. Session is described

Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE'04)
1071-9458/04 $ 20.00 IEEE

as a sequence of requests from the same user during a single visit to the Web site; sessions boundaries are delimited by a period of inactivity by a user. Some Web sites enforce a threshold and close inactive sessions to save resources allocated to these sessions. In [4] the authors introduced the concept of session–based admission control aimed at increasing the chances that longer sessions will be completed. In [3] authors studied how the threshold value affects the number of sessions and focused on other session characteristics such as distribution of number of requests per session, session length, and inter–session arrival times. The work presented in [9] went further in characterizing Web user sessions. First, starting from Web access logs, the session logs were generated. Then, a set of typical state transition graphs called Customer Behavior Model Graph (CBMG) were generated based on clustering analysis. As continuation of this work, in [10] priority–based resource management policies based on CBMG representation and simulated workload were proposed in order to increase the business–oriented metrics such as revenue/sec. The work presented in [11] studied the request, function, and session characteristics of two actual e–commerce sites.

As described above, a considerable amount of research work in the past was focused on the characteristics of Web workload. However, only few papers published so far have considered analysis of Web error behavior and modelling or measurement of Web server reliability. In [3] the authors analyzed the server response codes from the Web access logs of 1998 World Cup Web Site and reported the percentage of request with successful responses, partial content responses, not modified responses, and responses with errors. In [8] the information about errors was extracted from the Web error logs of the School of Engineering and Applied Science at the Southern Methodist University. The error analysis presented in the paper consisted of the summary on the number of different types of errors per day; the Web server reliability was computed as the number of non-erroneous requests over the total number of requests. In [1] a simple high level model of Web applications, which consists of a user model, a browser model, and a server model, was proposed. This paper lacks any numerical or empirical assessment of the reliability. Recently, an approach for deriving a transition tree was proposed and used for estimating Web reliability in [15]. This work, however, did not present any empirical results based on the actual Web logs. Instead, two simple hypothetical examples were provided as an illustration of the proposed approach.

The focus of our study is on characterization of the Web workload and reliability in terms of user sessions defined as a sequence of requests coming from the same user within a given threshold. Our analysis is based on data extracted from Web logs of seven public and three private Web servers. Although there are many existing commercial and open source tools for analysis of Web access logs, they either do not consider sessions at all or provide limited predefined reports on sessions. Furthermore, there are no tools available for analyzing the Web error logs. To overcome the limitations of the existing tools for analysis of the Web logs, we have developed a tool that extracts the detailed workload and error information from Web access and error logs and allows us to conduct systematic and flexible analysis. The contributions of this paper include:

- **Characterization of the Web workload in terms of sessions.** For this purpose we introduce and empirically analyze several intra–session and inter–session workload characteristics. Some of the session–based workload characteristics considered in this paper (i.e., session length and number of request per session) have been analyzed earlier in [3]. However, while the authors in [3] have used data from only one server, we present empirical analysis of session–based workload characteristics of ten Web servers and compare their values for public and private Web servers. It is important to emphasize that analysis of private Web sites, to the best of our knowledge, has not been presented in the literature earlier.

- **Characterization of the request–based and session–based reliability.** Unlike the previously published results on Web errors that were limited on the number of request with different erroneous status codes [3], [8], we present a detailed analysis of the Web server error characteristics based on data extracted from the Web error logs. The error analysis includes severity and frequency of error occurrence, as well as the analysis of the probability distribution functions of the error rates. In addition to request–based reliability considered earlier [8], we introduce and empirically analyze the session–based reliability. We argue that the distribution of the requests with erroneous responses within Web sessions is a better indicator of the Web server quality. Unlike some of the earlier papers focused on Web reliability that presented models that were not supported by real data [1], [15], in this paper we present empirical analysis of the request–based and session–based reliability based on actual logs from ten Web servers.

- **Identification of the invariants of the Web workload and reliability**. Finally, we identify the common characteristics of the Web workload and reliability that apply through the all data sets considered.

3 Data extraction and analysis process

Logs maintained at the Web server can reveal a huge amount of relevant information about the workload and errors experienced by the Web server. Access log files contain a separate entry for each request to the Web server. A sample entry from the access log of a Web site using Apache Web server that follows the Common Log Format (CLF) [16] is given in Figure 1. This entry tells us that at February 9, 2003, at one second after midnight, eastern time, the user from the IP address 217.72.93.150 asked the server for the file `/vti-inf.html`. The server was informed that the client supported HTTP/1.1 protocol. The server successfully responded to this request (indicated by the status code 200) and transferred 292 bytes to the client.

Although access logs record the status code of the response, separate error logs are used by the Web server to record any errors that it encounters in processing requests. Error logs are the first place to look when a problem occurs, since it often contains details of what went wrong and how to fix it. The format of the error log is relatively free–form and descriptive. However, there is certain information that is contained in most error log entries. For example, a typical entry from the error log

```
217.72.93.150 - - [09/Feb/2003:00:00:01 -0500]"GET /vti-inf.html HTTP/1.1"
200 292
```

Figure 1. A sample entry in access log

```
[Wed Apr 2 20:25:10 2003] [error] [client 12.44.188.239] File does not exist:
/projects/msrc/www/images/marble2d.gif
```

Figure 2. A sample entry in error log

of a Web site using Apache Web server is given in the Figure 2. The first item in the log entry is the date and time of the message, followed by the severity of the error being reported (`[error]`). The third item gives the IP address of the client (12.44.188.239). Beyond that is the error message itself and the file system path (as opposed to the Web path) of the requested document. In this example, the user was trying to access the file `/projects/msrc/www/images/marble2d.gif` which does not exist on the server.

The Web logs used in this paper were obtained from ten Web servers: three public and three private Web servers at the NASA Independent Verification and Validation Facility (NASA IV&V)[1], the Lane Department of Computer Science and Electrical Engineering (CSEE) Web server at West Virginia University, the Web server of a commercial Internet provider ClarkNet, the Web server at the NASA Kennedy Space Center (NASA-KSC), and the campus wide Web server at the University of Saskatchewan. The data sets of the NASA IV&V and CSEE Web servers consist of the access and error logs collected recently at the six NASA IV&V servers and two redundant Web servers at our department. For the other three data sets, which were downloaded from the Internet traffic archive [18], only the access logs are available.

Table 1 summarizes the raw data from the ten access logs. The access logs span different time durations, from two weeks to seven months. Furthermore, they provide information on servers with different workloads which vary by three orders of magnitude. Also, the servers are from different domains: two from educational institutions, seven from research institutions, and one from a commercial Web site.

While there are many existing tools for analyzing Web access logs [17], most of them provide limited analysis and generate predefined, fixed reports about the number of Kbytes transferred, access organized by domains, etc. These tools either do not consider sessions at all or provide limited information about sessions. Moreover, tools for analyzing Web error logs do not exist. To overcome the limitations of the existing tools, we developed our own tool for extracting relevant information from the Web access and error logs which provides bases for flexible and systematic analysis.

Figure 3 represents our data extraction and analysis system. Since the textual format of the log data is not suited for flexible, customized analysis, we include the log entries from the access and error logs as records in the corresponding relational databases. In order to extract sessions from Web access logs from an architecture that employs redundant Web servers (i.e., CSEE) we merge the logs from all Web servers using the time stamps of the incoming requests. This is necessary since, depending on the load balancing algorithm used, it is possible that consecutive requests from the user within the same session are directed to different Web servers. The error log files of the redundant Web servers are analyzed separately which allows us to observe any discrepancy in the error behavior of the servers.

The analysis of session–based workload is focused on several intra–session and inter–session characteristics that collectively describe the Web workload in term of sessions. The analysis of errors includes severity and frequency of occurrence of unique errors, identification of probability distribution functions of error rates, analysis of unique files with errors and their contribution to the total number of errors. In addition to detailed error analysis, we study the request–based and session–based reliability.

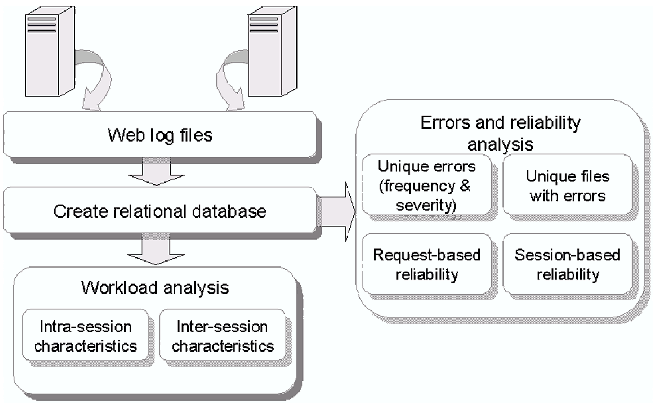

Figure 3. Data extraction and analysis process

4 Session-based workload

A unique characteristic of Web workload is the concept of session. A session is defined as a sequence of requests from the same user during a single visit to the Web site. A session begins when the user issues a request for a particular page on a Web site. Upon receipt of the response the user's Web browser parses the file and automatically generates requests for all embedded files. Therefore, a session may be present even in the case when the client requests only a single Web page. For example, accessing a home page involves requesting the HTML page, and then making further requests for all the images embedded in this document. The Web session continues as the user makes subsequent requests on the Web site. For example, placing an order through the e–commerce site involves requests related to selecting a product, providing shipping information, arranging payment, and receiving confirmation. So, for a customer trying to place an order or a retailer trying to make a sale, the real measure of a Web server success is its ability to process the entire sequence of requests needed to complete a transaction.

Next, we discuss in detail the process of session extraction.

[1]The Web logs of the NASA IV&V servers were sanitized, that is, IP addresses were replaced with unique identifiers.

Data set	Log duration	Start date	Requests	Average requests per day	Sessions	Average sessions per day	MB trans-ferred	Average MB per day
NASA-Pvt1	3 weeks	Mar. 16, 2004	3,862	184	115	5	62	2.9
NASA-Pvt2	3 weeks	Mar. 16, 2004	11,863	565	523	25	18	0.8
NASA-Pvt3	3 weeks	Mar.16, 2004	59,061	2,812	2,444	116	244	11.6
NASA-Pub1	3 weeks	Mar. 16, 2004	10,459	498	2,584	123	1,552	73.9
NASA-Pub2	3 weeks	Mar. 16, 2004	106,571	5,075	10,783	513	904	43
NASA-Pub3	3 weeks	Mar. 16, 2004	15,373	732	1,906	91	617	29.3
CSEE	6 weeks	Mar. 2, 2003	5,815,202	135,237	252,573	5,873	80,913	1,881
ClarkNet	2 weeks	Aug. 28, 1995	3,328,632	237,759	283,961	20,282	27,646	1,974
NASA-KSC	2 months	July 1, 1995	3,461,612	59,682	306,523	5,284	62,488	1,974
Saskatchewan	7 months	June 1, 1995	2,408,623	11,255	463,684	2,166	12,344	57

Table 1. Summary of the access log characteristics (raw data)

In this paper, for practical reasons, we define a session as a sequence of requests issued from the same IP address with the time between requests less than some threshold value. Two points of this definition require more detailed elaboration: (1) identifying the user by the IP address and (2) the value of the threshold that delimits the sessions.

First, we address the inaccuracies introduced by using the IP addresses as surrogate for users. As in all other research papers that considered sessions, we consider each unique IP address in the access log to be a distinct user. Clearly, this is not true in all cases [3], [14]. For example, if a proxy server exists between the user and the server, the IP address in the Web access log will be the address of the proxy, rather than the address of the originating machine. Furthermore, even when a unique IP address is assigned to a single machine, it may be a machine available for public access, such as for example machines in the university laboratories. In spite of these inaccuracies, we believe that using the IP address provides a reasonable approximation of the number of distinct users.

Next, we address the value of the threshold used to delimit the sessions. Some Web servers close sessions after a period of inactivity longer than a given threshold to save resources allocated to inactive sessions. If the Web site does not enforce a threshold, we need to estimate the threshold from the Web access logs. In the remainder of this section we examine the effects of various threshold values on the total number of user sessions. Figure 4 shows the effects that different threshold values (i.e., 1, 3, 5, 10, 20, 30, and 40 minutes) have on the total number of sessions seen in the ten data sets described in Table 1[2]. As the threshold value increases, the total number of sessions decreases rapidly. Once the threshold values larger than 30 minutes are used, there is a little further reduction in the total number of sessions, even with substantial increases in the threshold value. This result confirms the de facto standard of 30 minutes for the threshold value [11]. Therefore, for the analysis of the session characteristics presented in the rest of this paper we adopt a 30 minutes time interval as the threshold value.

Next, we study the session–based workload characteristics. For this purpose, we define several intra–session and inter–session metrics.

[2]Note that the ten data sets span significantly different time intervals. Therefore, the number of sessions presented in Figure 4 is not a measure of the traffic intensity.

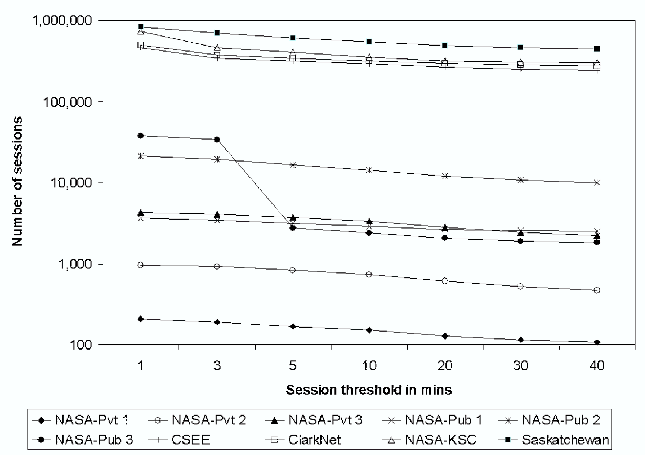

Figure 4. Effect of the session threshold on the number of sessions

4.1 Intra–session characteristics

In this section we analyze the session length, number of request per session, and number of bytes transferred per session. We refer to these characteristics collectively as intra–session characteristics.

4.1.1 Session length

Figure 5 presents the histogram of the session length in minutes. For all data sets, most of the sessions are very short and last less than a few minutes. It is obvious that as the session length increases, the percentage of session decreases rapidly. Despite the fact that the data sets are from different domains, the following common observation can be made from the results presented in Figure 5:

A significant percentage of sessions (e.g., 45–85%) last less than 1 minute. Most of the sessions (e.g., 73–92%) last less than 10 minutes.

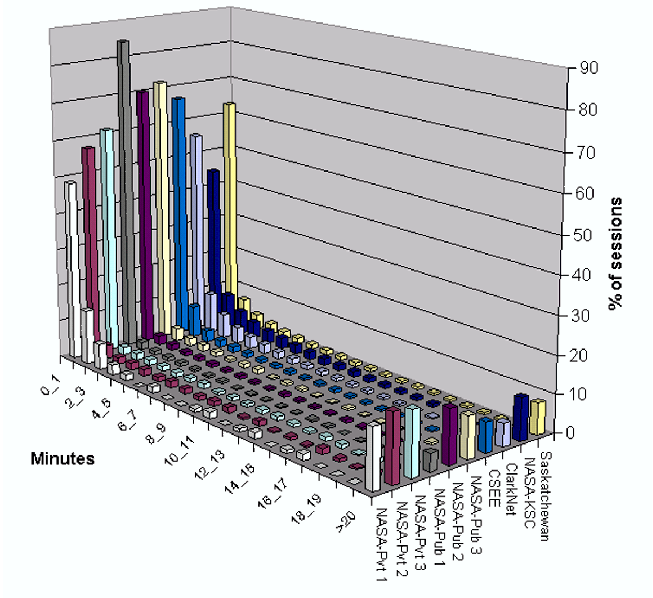

Figure 5. Histogram of the session length

4.1.2 Number of request per session

Next, we analyze the number of request issued by the user during a session. Several observations can be drawn from Figure 6. First, it is obvious that the distribution of the number of request per session differs significantly for the private and public Web servers. Thus, most public servers' sessions consist of only several request. Another intriguing observation is that most of the public servers have high percentage of sessions with only a single request. Also, as the number of requests per session increases, the percentage of sessions drops rapidly. This is not the case with the private Web sites; they have insignificant percentage of a single request sessions and do not show monotonically decreasing trend with the increase of the number of request. For example, NASA IV&V private Web servers have 30%, 31%, and 38% sessions with 10–11, 16–17, and 12–13 requests respectively. The summary of the number of requests per session workload characteristic includes:

> *The percentage of sessions with less than 10 requests is significantly smaller for the private (14–35%) than for the public (63–94%) Web servers. Unlike private servers that have few single request sessions (0.49–2.61%), public servers have a significant percentage of single request sessions (10–71%).*

4.1.3 Number of bytes transferred per session

Our last intra–session measure is the total number of bytes transferred per session. For this measure, we count the bytes transferred for both completed and partial transfers. As it can be seen from Figure 7, in seven of the servers the highest percentage of sessions transferred between 10KB and 100KB. Furthermore, 65–100% of sessions in all servers transferred between 1KB and 1MB of content. Another interesting observation is that private Web servers have few sessions with no data transfer (0–1.74%). The percentage of sessions that transferred no data is slightly higher for the public Web sites (3.12–12.62%).

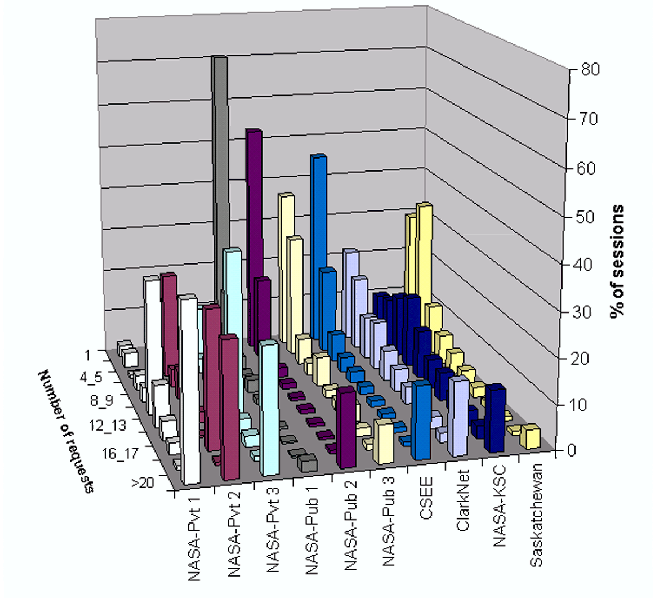

Figure 6. Histogram of the number of requests per session

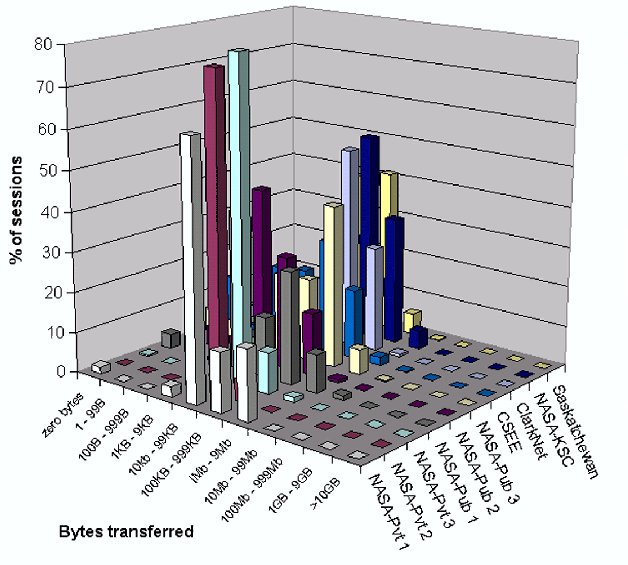

Figure 7. Histogram of bytes transferred per session

The following common workload characteristics hold for all data sets:

> *Most of the sessions (e.g., 65–100%) transferred between 1KB and 1MB data. Approximately 0–13% of sessions transferred no content data. Very few sessions (e.g., 0–1.2%) transferred more than 10MB of content data.*

4.2 Inter–session characteristics

In this section we analyze the inter–session characteristics that address the relationship between different sessions. For this pur-

pose we consider the number of sessions per user and number of sessions initiated per day and per hour.

4.2.1 Sessions per user

Figure 8 presents the sessions per user measure which considers the number of sessions originated by the unique users. Here we consider each session as one visit to the Web site. The vast majority of the public Web site's users (i.e., 69.48–93.09%) visited the site only once, that is, had only one session. This constitutes 24.58–71.40% of the total number of sessions. On the other side, the number of users of the private Web sites that have visited the site only once is significantly smaller (22.38–44.44%) which constitutes 1.31–13.91% of the total number of sessions. In each data set there were users that had revisited the Web site large number of times. The maximum number of sessions per user ranges from 27 in NASA-Pvt1 to 2897 in Saskatchewan. Note that the number of sessions per user depends on the intensity of the workload and tends to be larger for data sets that span longer time period.

The following common observations can be made about the number of sessions per user:

> *69–93% of the public Web sites' users and 22–44% of the private Web sites' users had only one session. This constitutes approximately 25–71%, that is, 1–14% of the total number of sessions. Very few users in each data set had extremely large number of sessions.*

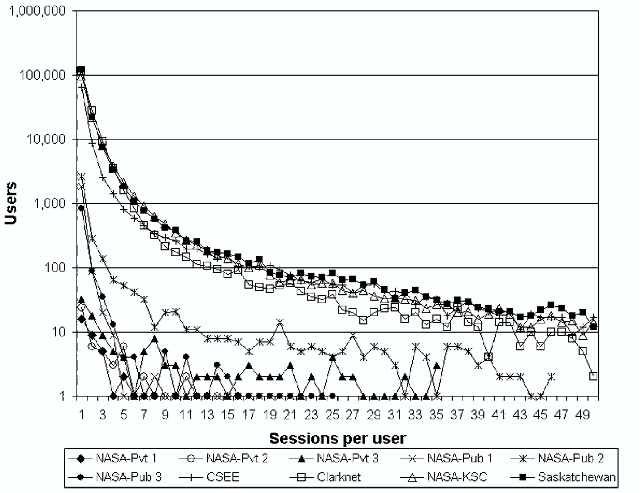

Figure 8. Number of sessions per user

4.2.2 Number of sessions initiated per day and per hour

We next study the number of session initiated per day. As it can be seen from Figure 9, which presents one week of data extracted from the access logs of each Web site, all ten data sets exhibit the expected periodical weekday/weekend behavior. Furthermore, the workload in sessions per day varies by three orders of magnitude. The Web server of the commercial Internet provider ClarkNet has the highest number of sessions initiated per day which is one order of magnitude higher than CSEE, NASA-KSC, and Saskatchewan, that is, $2-3$ orders of magnitude higher than the NASA IV&V Web sites.

At a time scale of one hour, we observe high variability in the number of sessions initiated per hour since the session threshold of 30 minutes is of the same order of magnitude as the time scale. We also observe the periodical day/night behavior.

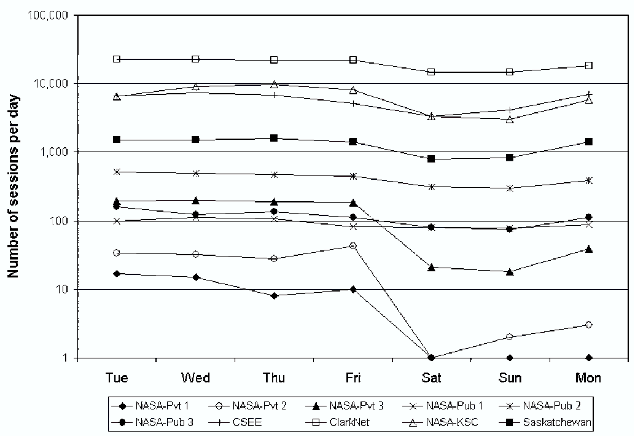

Figure 9. Sessions initiated per day

5 Request–based error characteristics

In this section we focus on request–based analysis of the Web errors. The first part of the analysis is based on the response codes for each request that are given in the Web server access logs. Some of the possible response codes to client requests include:

- `Successful (2xx)`
  - `OK (200)` - Successfully completed request.
  - `Partial content (206)` - The server has fulfilled the partial GET request for the resource.
- `Redirection (3xx)`
  - `Moved permanently (301)` - The requested resource has been assigned a new permanent URI.
  - `Found (302)` - The requested resource resides temporarily under a different URI.
  - `Not modified (304)` - The client, which already has a copy of the document in its cache, is told that the document has not been modified at the server.
- `Client error (4xx)` - Include errors such as `Unauthorized`, `Not found`, `Method not allowed`, `Conflict`, or `Unsupported media type`.
- `Server error (5xx)` - Indicate that the server is aware that it has erred or is incapable of performing the request. Examples include `Internal server error`, `Not implemented`, `Service unavailable` or `HTTP version not supported`.

Table 2 shows the breakdown of the server response codes in percentage of requests for the ten data sets. They reveal that the majority of request resulted in responses without errors

Status code	NASA Pvt1	NASA Pvt2	NASA Pvt3	NASA Pub1	NASA Pub2	NASA Pub3	CSEE	ClarkNet	NASA KSC	Saskat-chewan
200	69.2128	45.2921	43.5313	81.5948	74.9791	69.8823	26.5694	88.7764	89.5687	91.0692
206	0.5697	0.0000	0.1134	5.2299	0.8286	5.3861	1.2581	0.0000	0.0000	0.0000
301	0.0000	0.0000	0.0068	0.6310	0.1877	0.6505	1.2719	0.0000	0.0000	0.0000
302	0.0000	0.0000	0.0000	0.0000	0.2712	0.0000	22.3007	0.8737	2.1109	1.6904
304	16.7012	50.3751	54.9195	4.9240	17.8463	18.6171	45.7964	8.0736	7.7066	6.2955
4xx	13.3868	4.2907	1.4206	7.5724	5.8449	5.4316	2.8031	2.2037	0.6107	0.9216
5xx	0.1295	0.0421	0.0085	0.0478	0.0422	0.0325	0.0004	0.0616	0.0031	0.0233
$R_{request}$	0.8648	0.9567	0.9857	0.9237	0.9411	0.9454	0.9720	0.9774	0.9939	0.9906

Table 2. Breakdown of status codes (in percentages) and request–based reliability

(codes 2xx and 3xx). The most common status code in three of the Web sites (i.e., NASA-Pvt 2, NASA-Pvt 3, and CSEE) is `Not modified` with 45.80–54.92% which is significantly higher than in the case of the other Web sites. The higher percentage of `Not modified` responses can be attributed to the improved caching capability of the Web which is especially effective for certain usage patterns that include revisiting the same content and/or Web sites that contain large amount of static content (e.g., HTML files and images).

Relatively few requests resulted in error responses, that is, the total percentage of responses with client or server errors (4xx or 5xx) is reasonably low in all data sets. Another important observation is that the number of `Client errors` (4xx) is 1–4 magnitudes higher than the number of `Server errors` (5xx). Most of the errors that did occur were the result of the server not finding anything matching the request-URI (status code 404 `Not found`), which usually represent bad links and should be counted as Web failures.

Based on the error analysis, we can estimate the request–based reliability of the Web sites. Using the Nelson's model [12], one of the most widely used input domain models, the request–based reliability is defined as

$$R_{request} = 1 - \frac{f_r}{n_r} = \frac{n_r - f_r}{n_r} \quad (1)$$

where f_r is the number of requests that have resulted in responses with erroneous code (4xx and 5xx) and n_r is the total number of requests. The estimates of the request–based reliability for the ten Web sites are given in Table 2 and Figure 10. The request–based reliability estimated in [8] for the Web site of the School of Engineering and Applied Science at the Southern Methodist University is also very high and close to our values (0.9597).

This analysis leads to the following common error characteristic:

The request–based reliability is in the range of 0.8648 – 0.9939.

Note that these estimates of the request–based reliability are conservative due to the fact that some of the errors, such as for example `Unauthorized` may actually be the expected behavior of the server which is denying unauthorized access to restricted resources. Our current research is focused on analysis of different types of errors that will improve the accuracy of the estimates for the Web reliability.

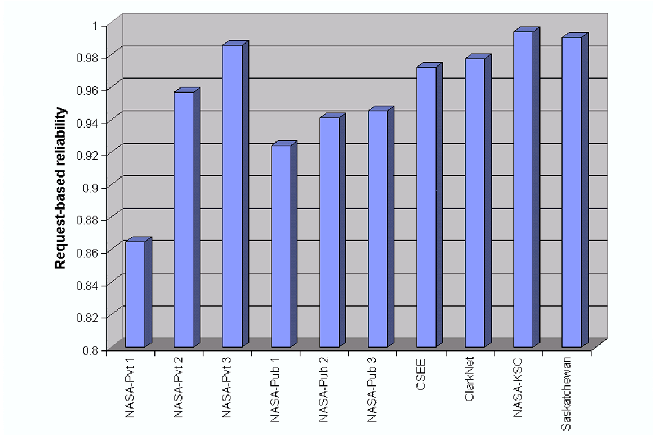

Figure 10. Request–based reliability

5.1 Detailed errors analysis

In this section we present a more detailed analysis of the errors based on the data extracted from the Web error logs of the six NASA IV&V Web servers and two redundant CSEE Web servers.[3] We focus on the analysis of the severity of the errors, unique errors, and unique files with errors. These topic were not addressed in any of the previously published papers that considered analysis of Web errors and reliability [1], [3], [8], [15].

5.1.1 Severity level of errors

Error logs provide very important information on the severity of the errors being reported which is extremely useful in making decisions on the priority and emergency of fixing errors. The following severity levels (in order of decreasing significance) are available in the Apache Web server error logs: `emerg`, `alert`, `crit`, `error`, `warn`, `notice`, `info`, and `debug`. The server can be configured to record error messages starting from a particular severity level and including all other levels of higher significance.

Table 3 presents the percentage of errors in each of the severity levels for the NASA IV&V and CSEE servers. Majority of the errors (up to 100%) in all servers have the `error` level of severity. CSEE servers, however, have a small percentage of errors with `crit` and `alert` severity level. Of course, fixing the errors with higher level of severity should have higher priority.

[3] For the other three data sets considered through the paper Web error logs were not available.

Severity level	NASA Pvt1	NASA Pvt2	NASA Pvt3	NASA Pub1	NASA Pub2	NASA Pub3	CSEE 1	CSEE 2
`alert`	0	0	0	0	0	0	0.018	0.005
`crit`	0	0	0	0	0	0	1.401	1.023
`error`	100	100	100	100	99.049	100	97.054	97.651
`warn`	0	0	0	0	0	0	0.842	0.832
`notice`	0	0	0	0	0.951	0	0.685	0.490

Table 3. Distribution of the error severity (in percentages)

The following observation is made for the severity level of errors:

> *Most errors recorded in Web error logs have severity level* `error` *(97.054–100%). Small percentage of errors have* `crit` *(1.023–1.401%) and* `alert` *(0.005–0.018%) severity levels.*

5.1.2 Unique errors

Due to the fact that multiple error messages may be associated with a single file, we define a unique error as a unique combination of an error message and a file in the error log entries. It follows that multiple unique errors may be associated with a single file. In order to identify the unique errors, the entries that consist of debugging output from the CGI scripts related to a single CGI error were excluded from the analysis. It is important to notice that in the case of the CSEE Web servers errors in a single CGI file have produced half a million debugging entries in the error logs. Fixing these errors, in addition to improving the Web server quality, improved the efficiency of the Web server and saved the resources wasted on logging.

Figure 11 presents the percentage of unique errors in the six NASA IV&V and two CSEE Web servers. NASA-Pvt1 and NASA-Pvt2 Web servers have the highest percentage of unique errors (72.40% and 70.05% respectively). CSEE Web servers have the lowest percentage of unique errors (18.42% and 18.20%). Of course, Web servers with the smaller percentage of unique errors will show faster and more cost effective improvement in their quality.

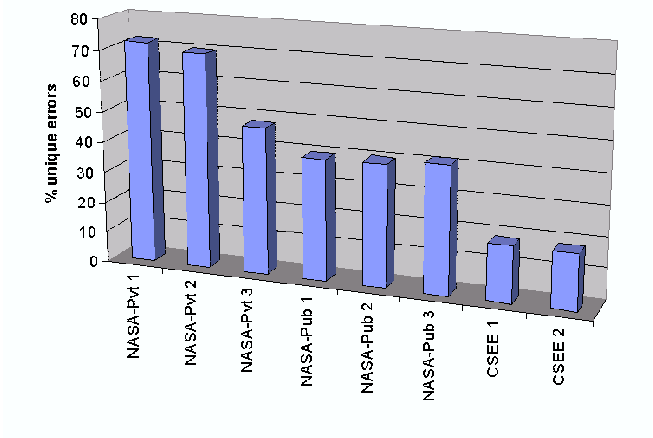

Figure 11. Percentage of unique errors

The frequency of occurrence of unique errors in the error log files of the NASA IV&V and CSEE Web servers are shown in Figures 12 and 13, respectively. It is obvious that the frequency of occurrence of many unique errors is low. For example, 213 – 1865 of unique errors in NASA IV&V Web servers have occurred only once in three weeks, that is, 8,032 of the CSEE 1 and 10,554 of the CSEE 2 unique errors have occurred only once in six weeks. However, some unique errors occur extremely large number of times. The highest number of occurrences of a unique error in NASA-Pvt1, NASA-Pvt2, NASA-Pvt3, NASA-Pub1, NASA-Pub2, NASA-Pub 3, CSEE 1, and CSEE 2 data sets are 37, 47, 106, 101, 304, 101, 3782, and 3970, respectively. Note that these numbers depend on the usage pattern (i.e., how often the users 'hit' a given erroneous resource), the traffic intensity, and the duration of the time considered.

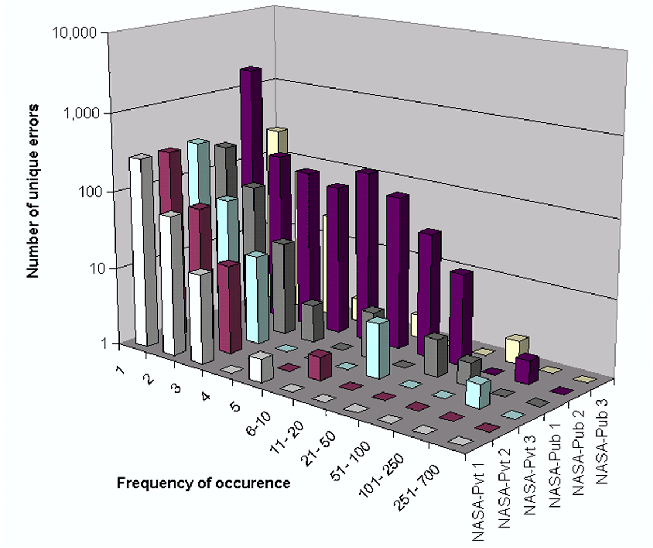

Figure 12. Frequency of occurrence of unique errors for NASA IV&V Web servers

The above observations clearly show that some unique errors occur large number of times with non-negligible probabilities. This fact has motivated us to formally assess whether the error rate, defined as a number of occurrences of a unique error per week, follows a heavy–tailed distribution. Given a particular data set, there are various methods of checking that a heavy tail model is appropriate. We choose the Hill estimator [7] which is statistically more rigorous method that does not depend on the existence of a finite mean. For the discussion that follows, let $X_1, X_2, \ldots, X_n$ denote observed values of the error rates and let

$$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)} \quad (2)$$

be the ordered statistics of the data set. The random variable X,

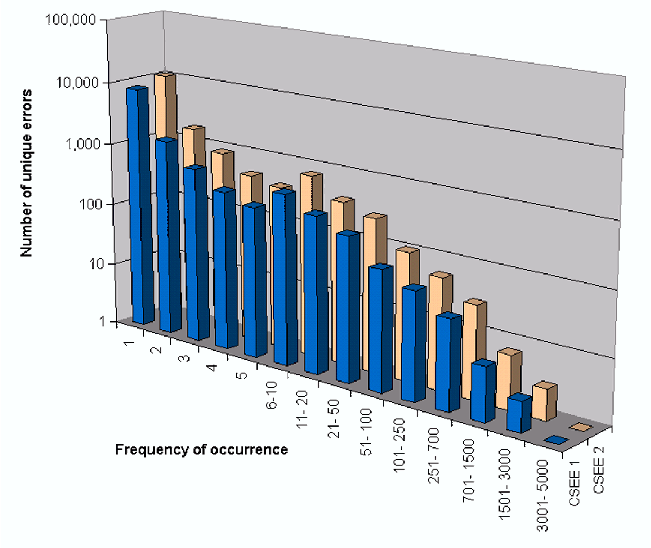

Figure 13. Frequency of occurrence of unique errors for CSEE Web servers

with cumulative distribution function $F(x)$ is said to be heavy–tailed if

$$1 - F(x) = x^{-\alpha} L(x) \tag{3}$$

where $L(x)$ is slowly varying as $x \to \infty$, i.e., $\lim_{x\to\infty} L(ax)/L(x) = 1$ for $a > 0$. A typical member of this distribution class is the Pareto distribution. There is an important qualitative property of the moments of heavy–tailed distributions. If X is heavy–tailed with parameter α then its first $m < \alpha$ moments $E[X^m]$ are finite and its all higher moments are infinite.

The rough idea behind the Hill estimator is to use only k upper-order statistics, that is, to sample from the part of the distribution which looks most Pareto-like. Therefore, we pick $k < n$ and compute the Hill estimator defined as

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log X_{(i)} - \log X_{(k+1)}. \tag{4}$$

The Hill statistic provides an estimate of the tail index α of a semiparametric Pareto type model given by (3). Thus, for each value of k we obtain an estimate of the tail index parameter $\alpha_{k,n} = 1/H_{k,n}$. In practice, the estimates of the tail index $\alpha_{k,n}$ are plotted as a function of k, for the range of k-values. A typical Hill plot varies considerably for small values of k (i.e., only a small fraction of the largest observations are considered), but becomes more stable as more and more data points in the tail of the distribution are included (often up to a cut-off value, to the left of which (3) no longer holds). If the plot stabilizes to a constant value one can infer the value of the tail index α. The absence of such straight line behavior is a strong indication that the data are not consistent with the heavy tail distribution (3).

Figure 14 depicts the Hill estimate plot corresponding to the CSEE 1 and CSEE 2 data sets for varying k restricted to the upper 2.5% tail. Both plots are gratifyingly stable after settling down and are in a tight neighborhood. Moreover, the Hill estimators of tail indexes are $\alpha_{\text{CSEE 1}} \approx 0.92$ and $\alpha_{\text{CSEE 2}} \approx 1$. This means that error rates of the CSEE 1 and CSEE 2 Web sites have infinite mean and variance.

Figure 15 presents the value of the Hill estimate for α for the NASA-Pub 2 Web site, again restricted to the upper 2.5% tail. The Hill plot shows that the estimator seems to settle to a relatively constant estimate in the interval between 1.4 and 1.7. It follows that the error rate of the NASA-Pub 2 Web site has a finite mean, but infinite variance.

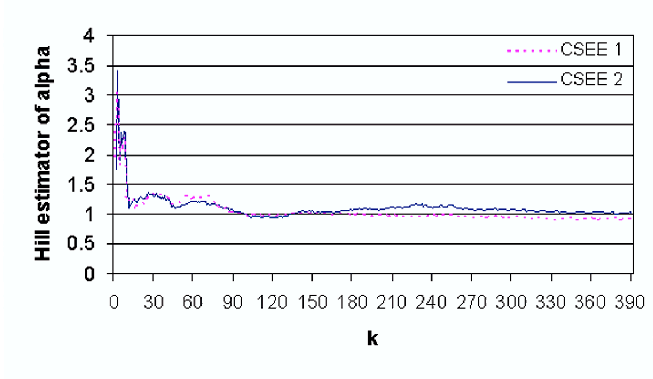

Figure 14. Hill plot for the CSEE 1 and CSEE 2 data sets

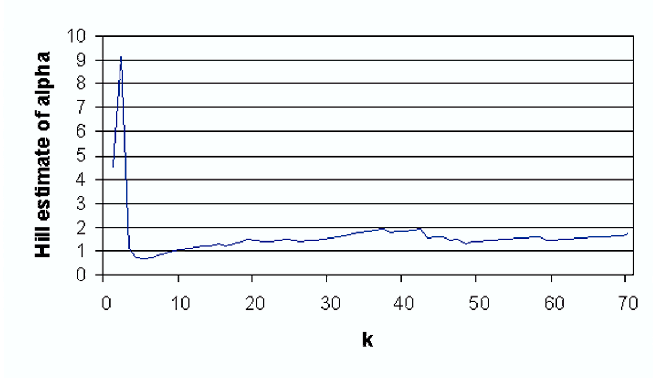

Figure 15. Hill plot for the NASA-Pub 2 data set

The Hill estimators of the other five NASA IV&V Web sites do not settle down as k increases, which is a typical behavior of the Hill plot when the data are inconsistent with assumption (3). Thus, the Hill estimator proves, in a statistically rigorous manner, the behavior that can be observed from Figures 12 and 13 - error rates of the NASA-Pub 2, CSEE 1, and CSEE 2 Web sites can plausibly be explained by a heavy tail model. This means that some unique errors will have extremely high rates of occurrence with non-negligible probability. For example, the highest error rates for of a unique error for NASA-Pub 2, CSEE 1, and CSEE 2 servers are 101, 630, and 662 errors per week, respectively.

The summary of the unique errors characteristics include:

The percentage of unique errors is in the range 18.20–72.40%. Many unique errors occur only once. However, some unique errors occur large number of times. The error rates of three Web sites are heavy tailed, which means that some unique errors will have extremely high rates of occurrence with non-negligible probability.

5.1.3 Unique files with errors

We also study the unique files with errors, regardless of the number and type of error messages associated with each file. In all data sets, the number of unique files with errors is slightly smaller than the number of unique errors because some files have more than one error message associated with them. Web administrators of the NASA IV&V and CSEE Web servers were given a list of unique files with errors ordered accordingly to the frequency of occurrence. Obviously, fixing the unique files with errors with the higher frequency of occurrence is the most cost effective way to improve the Web site quality. As an illustration, in Figure 16 we present the percentage of total number of errors that are due to the three files with the highest frequency of occurrence in each data set[4]. Thus, a significant percentage of the total number of errors (9.42% – 34.66%) is due to only three files, which means that fixing the errors in these files will improve the reliability of the Web sites significantly. For example, fixing the errors related to only 3 files in NASA-Pvt3, NASA-Pub1, NASA-Pub3, CSEE 1, and CSEE 2 Web servers has eliminated approximately the same or larger percentage of errors as if 294, 213, 234, 8,032, and 10,554 unique files with errors that occur only once were fixed. Although not as impressive, fixing the errors related to only 3 files in the other three Web servers (i.e., NASA-Pvt1, NASA-Pvt2, and NASA-Pub2) have eliminated 9.42–12.87% of the total number of errors. Of course, in addition to the frequency of occurrence of unique files with errors, the severity level of the errors have to be considered when deciding on the priority and emergency of fixing errors.

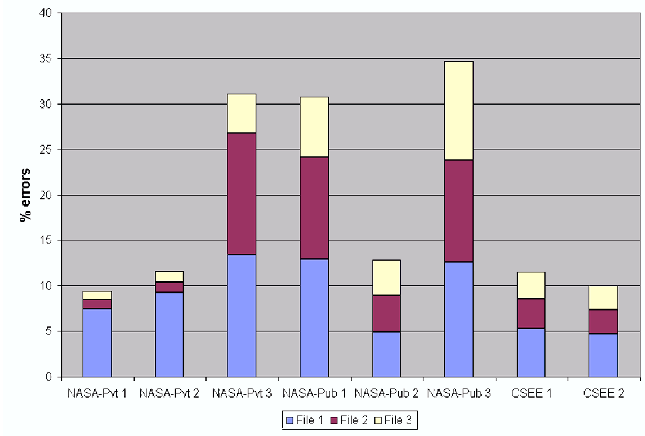

Figure 16. Percentage of errors due to the top three files with errors

The following important conclusion with respect to the unique files with errors is made:

> *9.42–34.66% of the total number of errors are due only to the three files with errors that occur most frequently in each data set. It follows that fixing only three files in each Web server results in a significant increase of the Web reliability.*

[4]The error logs contain the full file path name for each file. To protect confidentiality and privacy of the Web servers, the real path names are not revealed.

6 Session–based error characteristics

In this section we introduce a new Web reliability measure - session–based reliability - that has not been considered previously in the literature. For this purpose, we first study the distribution of errors within sessions. It is obvious from Figure 17, which presents the histogram of the number of errors per session, that the majority of sessions have responses to all requests without an error. In particular, 75 – 98% of all sessions do not have erroneous status codes (4xx and 5xx). Furthermore, the number of session with higher number of errors decreases rapidly as the number of errors per session increases.

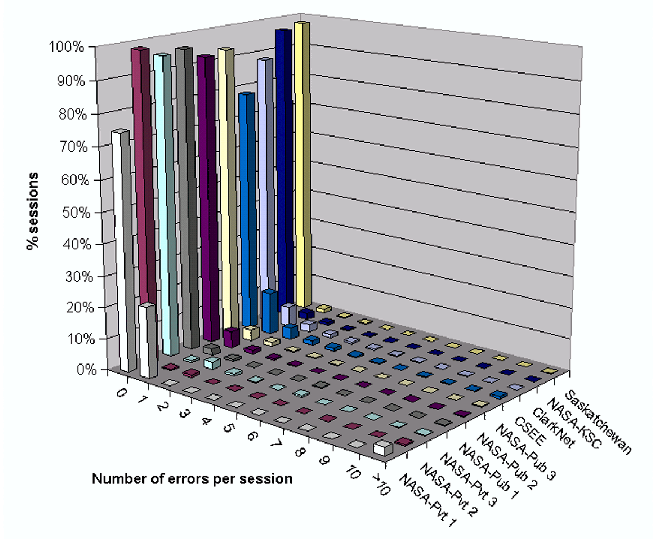

Figure 17. Histogram of the number of errors per session

Based on the analysis of the errors per session, we can estimate the session–based reliability as

$$R_{session} = 1 - \frac{f_s}{n_s} = \frac{n_s - f_s}{n_s} \quad (5)$$

where f_s is the number of sessions that have at least one request with erroneous code (4xx and 5xx) and n_s is the total number of sessions. Session–based reliability can be interpreted as the probability that a user of a Web server will not experience error in any of requests that constitute the user session.

The values for the session–based reliability are given in Table 4. An important observation is that for the most Web sites the session–based reliability is lower than the request–based reliability. This is due to the fact that if there is at least one request with an erroneous response code we consider that the whole session has failed. In particular, sites that have large number of sessions with a few errors (i.e., NASA-Pvt1, CSEE, and ClarkNet) exhibit significantly smaller session–based reliability than request–based reliability. The smaller session–reliability reflects the fact that many users will experience at least one error within their sessions. On the other side, two Web sites (i.e., NASA-Pvt2 and NASA-Pub1) have higher session–based reliability than request–based reliability. This phenomenon is due to the fact that there are few sessions that contain significant number of errors per session. Thus, although there might be a relatively high number of

	NASA Pvt1	NASA Pvt2	NASA Pvt3	NASA Pub1	NASA Pub2	NASA Pub3	CSEE	ClarkNet	NASA KSC	Saskat-chewan
$R_{session}$	0.7478	0.9828	0.9529	0.9613	0.9251	0.9360	0.7814	0.8806	0.9650	0.9782

Table 4. Session–based reliability

erroneous responses, if they are distributed in a small number of sessions, the session–based reliability will be high reflecting the fact that only a few users will be affected.

Based on the analysis presented in this section, we conclude that:

> *For the data sets considered in this paper, the session–based reliability is in the range of 0.7478 – 0.9828. For the most Web sites, the session–based reliability is lower than the request–based reliability.*

We believe that the session–based reliability estimates are very important for measuring the ability of Web servers to process the entire sequence of requests without an error and they are better indicators of the users perception of the quality of the Web servers than the request–based reliability. In our future research we will consider the impact of different types of errors and further refine the session–based reliability. For example, it would be useful to distinguish between insignificant errors such as an image that does not appear in a given Web page and critical errors that might prevent a user to place an online order.

7 Invariants of Web workload and error behavior

In addition to introducing and studying new metrics that collectively describe the session–based workload and reliability, our goal is to identify those characteristics that are common across all data sets studied. This aspect of our work can be seen as a continuation of the study published in [2] which has identified ten request–based invariants that involve characteristics such as successful requests, distinct request, remote requests, concentration of references, and file size distribution. In this paper we have identified session–based invariants that involve Web workload and reliability metrics. These characteristics are summarized in Table 5, including the reference of the section in the paper that describes them in more detail.

8 Conclusion

A solid understanding of the Web workload and error behavior is fundamental to improving Web quality attributes such as reliability, performance, and security. In this paper we have presented a detailed empirical analysis of the session–based workload and reliability based on the data extracted from actual Web logs of ten Web sites.

First, we have systematically analyzed the session–based workload in terms of several intra–session and inter–session characteristics, including the comparison between public and private Web sites workload that has not been addressed previously in the literature. The private Web sites considered in this paper have significantly higher number of sessions initiated by the same users and sessions with significantly higher number of requests than the public Web sites. These results clearly show that the operational (i.e. user) profiles are substantially different for the private and public Web servers. While the users of private (internal corporate) servers usually revisit the server regularly and tend to perform long sessions related to their work, significant percentage of the users of public Web servers are single time users and many sessions are extremely short-lived.

Then, we have analyzed Web error characteristics and estimated the request–based and session–based reliability of Web servers. The presented analysis of the severity and frequency of occurrence of errors is extremely useful in deciding on the priority for fixing errors. We have shown, in a statistically rigorous manner, that the error rates of some Web sites can be explained by a heavy tail model which means that some unique errors will have extremely high rates of occurrence with non-negligible probability. The analysis of the unique files with errors has proved that fixing the errors associated with only a few files is the most cost effective way to improve the Web server quality, leading to a significant reduction of the total number of errors.

Next, we have introduced and empirically evaluated a new measure of Web reliability - session–based reliability - and argued that it is a better indicator of the users perception of the Web quality than the request–based reliability.

The last contribution of this paper consists of the identification of the invariant workload and reliability characteristics that hold across data sets studied.

Our future research is focused on the study of the phenomena noticed in this paper, such as for example sessions with only one request, extremely long session, extremely long sequence of sessions originated from the same user, and detailed analysis of different types of errors.

Acknowledgements

This work is funded in part by grant from the NASA Office of Safety and Mission Assurance (OSMA) Software Assurance Research Program (SARP) managed through the NASA IV&V Facility, Fairmont, West Virginia and by grant from the West Virginia University Research Corporation, Program to Stimulate Competitive Research (PSCoR). The authors would like to thank David Krovich of the Lane Department of Computer Science and Electrical Engineering, West Virginia University and Brian Kesecker of the NASA IV&V Facility for making the Web logs available. They also thank Ken McGill for giving them permission to use the NASA IV&V data.

References

[1] V. S. Alagar and O. Ormandjieva, "Reliability Assessment of Web Applications", *Proc. 26th Annual International Computer Software and Applications Conference (COMPSAC'02)*, 2002.

[2] M. Arlitt and C. Williamson, "Internet Web Servers: Workload Characterization and Performance Implications", *IEEE/ACM Transactions on Networking*, Vol.5, No.5, October 1997, pp. 631-645.

Intra–session characteristics		
Session length	A significant percentage of sessions (e.g., 45–85%) last less than 1 minute. Most of the sessions (e.g., 73–92%) last less than 10 minutes.	Section 4.1.1
Request per session	The percentage of sessions with less than 10 requests is significantly smaller for the private (14–35%) than for the public (63–94%) Web servers.	Section 4.1.2
Bytes per session	Most of the sessions (e.g., 65–100%) transferred between 1KB and 1MB data. Approximately 0–13% of sessions transferred no content data. Very few sessions (e.g., 0–1.2%) transferred more than 10MB of content data.	Section 4.1.3
Inter–session characteristics		
Sessions per user	69–93% of the public Web site's users and 22–44% of the private Web site's users had only one session. Very few users in each data set had extremely large number of sessions.	Section 4.2.1
Sessions initiated	The number of sessions initiated per hour is more variable than the number of sessions initiated per day.	Section 4.2.2
Error characteristics		
Severity level of errors	Most errors have severity level `error` (97.054–100%). Small percentage of errors have `crit` (1.023–1.401%) and `alert` (0.005–0.018%) severity levels.	Section 5.1.1
Unique errors	Many unique errors occur only once. However, some unique errors occur large number of times. The error rates of three Web sites are heavy–tailed.	Section 5.1.2
Unique files with errors	9–35% of the total number of errors are due only to the three files with errors that occur most frequently in each data set.	Section 5.1.3
Request–based reliability	Request–based reliability is in the range of 0.8646 - 0.9939.	Section 5
Session–based reliability	For most Web sites session–based reliability is lower than the request–based reliability. For the data sets considered in this paper, the session–based reliability is in the range of 0.7478 - 0.9828.	Section 6

Table 5. Summary of the workload and error characteristics common across Web sites

[3] M. Arlitt and T. Jin, "Workload Characterization of the 1998 World Cup Web Site", *Hewlett-Packard Technical Report*, HPL-1999-35(R.1), Sep. 1999.

[4] L. Cherkasova and P. Phaal, "Session Based Admission Control: a Mechanism for Improving the Performance of an Overloaded Web Servers", *HP Labs Technical Reports*, HPL-98-119, 1998.

[5] P. Cremonesi and G. Serazzi, "End-to-End Performance of Web Services", *Performance 2002*, M.C.Clzarossa, S.Tucci (Eds.), LNCS 2459, Springer-Verlag, 2002, pp. 158-178.

[6] M. E. Crovella and A. Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, Vol.5, No.6, Dec.1997, pp. 835-846.

[7] B. M. Hill, "A Simple General Approach to Inference about the Tail of a Distribution", *Annals of Statistics*, Vol.3, No.5, 1975, pp. 1163-1174.

[8] C. Kallepalli and J. Tian, "Measuring and Modeling Usage and Reliability for Statistical Web Testing", *IEEE Transaction on Software Engineering*, Vol.27, No.11, Nov. 2001, pp. 1023-1036.

[9] D. Menasce, V. Almeida, R. Fonseca and M. Mendes, "A Methodology for Workload Characterization of E-commerce Sites", *Proc. ACM Conference on Electronic Commerce (EC-99)*, Denver, CO, Nov. 1999, pp. 119-128.

[10] D. A. Menasce, V. A. F. Almeida, R. Foneca, and M. A. Mendes, "Business–oriented Resource Management Policies for E-commerce Servers", *Performance Evaluation*, Vol.42, No.2-3, 2000, pp. 223-239.

[11] D. Menasce, V. Almeida, and R. Ried, "In Search of Invariants for E-Business Workloads", *Proc. 2nd ACM Conference on Electronic Commerce (EC'00)*, Minneapolis, MI, Oct. 2000, pp. 56-65.

[12] E. Nelson, "Estimating Software Reliability from Test Data" *Microelectronics and Reliability*, Vol.17, No.1, 1978, pp. 67-73.

[13] D. Oppenheimer and D. Patterson, "Architecture and Dependability of Large-Scale Internet Services", *IEEE Internet Computing*, September-October 2002, pp. 41-49.

[14] M. Rosenstein, "What is Actually Taking Place in Web Sites: E-Commerce Lessons from Web Server Logs", *Proc. 2nd ACM Conference on Electronic Commerce (EC'00)*, Minneapolis, MI, Oct. 2000, pp. 38-43.

[15] W. Wang and M. Tang, "User–Oriented Reliability Modeling for a Web System", *Proc. 14th International Symposium on Software Reliability Engineering (ISSRE 2003)*, Denver, CO, Nov. 2003, pp. 293-304.

[16] Log Files - Apache HTTP Server, http://httpd.apache.org/docs-2.0/logs.html

[17] A Listing of Access Log Analyzers, http://www.uu.se/Software/Analyzers/Access-Analyzers.html

[18] Internet Traffic Archive, http://ita.ee.lbl.gov/html/traces.html