

Web Server Traffic Characterization at the UFL College of Design, Construction, and Planning

Zornitza Genova Prodanoff

School of Computing
University of North Florida
4567 Saint Johns Bluff Road
Jacksonville, Florida 32224
zprodano@unf.edu

Karen Cano

School of Computing
University of North Florida
4567 Saint Johns Bluff Road
Jacksonville, Florida 32224
cank0005@unf.edu

Abstract - Our study evaluates user (Web browser) accesses to a Web server. We examine a Web server log that was collected for the span of several months at the College of Design, Construction, and Planning in the University of Florida in Gainesville, Florida. We have produced statistics that give us an insight into user access behavior. We were able to isolate cases that exhibit trends in request origination. We were also able to quantify the increase of on-campus originating requests (in-network traffic) during registration periods – about 3.6% absolute increase of total number of hits, which translates to about 2% relative increase, taking into account fluctuations in all other types of traffic. These results will be used to study a possible relationship between recent trends in student enrollment, originating from exchange programs and Web server traffic. They can also be used for server-side capacity planning.

Keywords

Web server, URL, HTTP 1.1, HTTP GET Request.

General Terms

Measurement, Performance, Experimentation, Theory.

1. Introduction

As defined by [9], a *Web server* houses files that are addressable by a single URL, such as HTML files, image files, applets, audio files, and video files. A Web server implements the server-side of the HTTP, which is the protocol between a Web server and a browser. A *Web site* is an HTML file, addressable by a single URL, which can consist of other files— such as other html files, images, applets, audio files, and video files— that are also addressable by a single URL [9]. The notion of a Web site is formally defined in [2] in an attempt to facilitate simple computation of related statistics. This formalization is summarized below:

```
<anything>(.edu or .com or .gov or .net)
<anything>(.co or .com).<country-digraph>
<anything>(.ac or .edu).<country-digraph>
<anything>(.army.mil, .af.mil or .navy.mil)
<anything-else>.mil
<anything>.<country-digraph>
```

and a series of ad-hoc rules to help with the .k12, <state-digraph>.us and <province-digraph>.ca sites.

A Web site can be comprised of many Web pages. The size of the average Web page is about 10 Kbytes - 60% of all Web pages are of this size [2]. Most Web pages are rich in content, for example, 50% of all pages contain at least one image reference [2]. The majority of Web pages are poorly connected to the rest of the Web. Most pages are linked to only by other pages at the same site. 80% of all sites contain no off-site links. In other words, “a small proportion of Web sites are carrying most of the load of hypertext navigation [2].” Nevertheless, roughly 75% of all sites contain at least one link to an outgoing URL (mostly on-site URLs) [2].

When a visitor requests a document from a Web site and the server delivers it to the user, a Web hit is said to have occurred. Figure 1 shows the flow of some hypothetical user (browser) request to a Web page, identified with a URL: <http://www.some.com/page>. When a user accesses a page, the browser (or Web client) sends a request to the Web server in the form of a URL. If the *resource* being requested is an HTML page, it may include other URLs embedded into its content, which are often images, audio files, video files, applets, as well as other text.

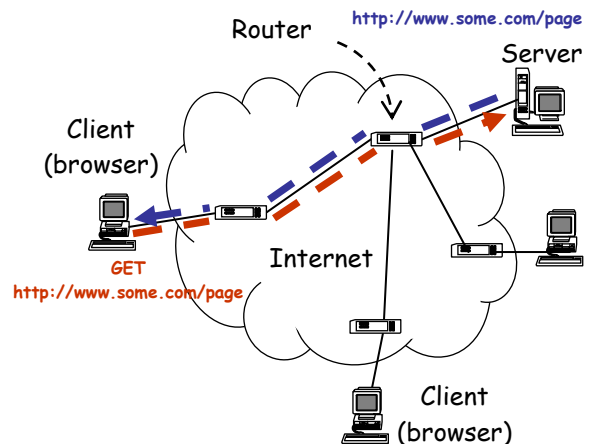


Figure 1. A Web client (browser) request to a Web server

When the client gets the requested HTML page back from the server, it reads the contents of the page and makes subsequent requests for the embedded URLs.

With the current Version of HTTP (1.1) Web pages are downloaded by requesting a main URL as described above. The main HTML document and a set of image, plain text, and other embedded files are then sent automatically to the client machine. Thus, a page with three embedded images results in a total of four hits as measured by the server (one hit for the HTML page, and one hit each of the three images).

Since the composition of pages differs within a site as well as across sites, the comparison of hits reported by different sites is inadequate measure of Web server load. Hence, in this study we do not use hit related statistics. We look at number of page accesses instead.

An HTTP GET request is a message sent to a Web server, requesting the download of a specific Web resource (Web page). Figure 2 shows a sample HTTP GET request header. HTTP is application layer protocol. Hence, the HTTP header will be appended on top of TCP/IP headers.

In this study we evaluate browser accesses to a target Web server and make an attempt to characterize the Web traffic to that server. The input data used in our evaluation is a several month long Web server log, produced by an Apache 2.0 Web server. This log has the same format as depicted in Figure 3. Figure 3 shows a sample from a Web log, consisting of a sequence of 6 GET requests, where each request (character string) is abbreviated and shown on two consecutive lines.

```
GET/ index.htm HTTP/1.1
Accept: image/gif, image/x-xbitmap, image/jpeg,
image/pjpeg, application/vnd.ms-powerpoint,
application/vnd.ms-excel, application/msword,
*/ *
Accept-Language: en-us
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/4.0 (compatible; MSIE 5.5;
Windows NT 5.0)
Host: 131.247.3.42
Connection: Keep-Alive
```

Figure 2. HTTP GET request header

```
30633 73.139.209.145 TCP_MISS/200 74052 GET
http://gigex1.com/M0016 ...
14104 227.229.152.199 TCP_MISS/200 17048 GET
http://fourohfour.xoom ...
13549 73.139.209.145 TCP_CLIENT_REFRESH_MISS/200
368 GET http://www ...
704 148.97.138.187 TCP_REFRESH_MISS/504 1339 GET
http://www.rocksho ...
706 112.211.98.27 TCP_DENIED/403 1149 GET
http://store2.yimg.com/I/ ...
707 244.60.215.3 TCP_MISS/503 1265 GET
http://www.linkexchange.ru/c ...
```

Figure 3. Web server access log

This paper is organized as follows: Section 2 is an overview of existing work and an introduction of the metrics used to perform Web traffic characterization, Section 3 presents our evaluation model and results, Section 4 provides a summary and outlines future work, Section 5 lists acknowledgements, Section 6 lists references, Section 7 present the biographies of the authors.

2. Related Work

There are limitations in the current methods used to collect usage data on the Web. These limitations stem from the decentralized nature of the Web. A Web user is not uniquely identifiable across the system, inhibiting the collection of reliable statistical data. Also, the Web employs various levels of caching, and since users can select different cache management policies, no promises can be made about the measurement of page views. Caching at the proxy level poses a similar scenario. Since many users share the cache, only one page request will be issued to the Web server. If other users choose to view the same page at a later time, they will be provided with the cached copy of the page. For all these reasons, gathering reliable visitor and page data is a difficult task [10].

The standard solution for tackling these limitations involves using cookies to track visitor identity and cache-busting to determine the number of page views. Nevertheless, hit-metering and sampling are proposed improvements on the current practices. Hit metering involves removing cache-busting and implementing a new HTTP header that enables proxy-caches to report usage and referral data to the originating Web server. Sampling is a form of inferential statistics that formulates conclusions about all users based on a subset of users. The field is supported by mathematical equations that have proven to function properly and consistently [10].

Gathering reliable data only solves the first part of the problem. Next, the gathered data has to be analyzed using statistics. Temporal and path analyses are methods for analyzing the gathered data [10].

2.1 Temporal Analysis

Temporal analysis is the study of how frequently an event occurs. In Web analysis, the collection of these statistics assumes that visitors can be uniquely identified and that requests can be traced in the sequence they occur, correctly and completely [10]. These assumptions can be made since cookies provide a consistent approach for identifying visitors and several proposals are underway for gathering solid data concerning page views. Some metrics used in temporal analysis are *reading time*, *session length*, and *Inter-visit period*.

Reading time is the time that a user spends reading a page. It is measured as the time between a request for a page and a subsequent request for a different page, by the same user. However, there is no guarantee that the user will spend that time actually reading the page. To make this statistic valid, a sample size of users is selected and studied. 30 minutes has become the standard time out period [10]. *Session length* is the total duration of a user's visit to a site. It is measured as the amount of time the

user spends requesting pages from the site [10]. In this study we do not use reading time or session length as a metric.

Inter-visit period is the time between a user's visits to a site. It is measured as the time between a user's visit to the site and the same user's subsequent visit to the site [10]. Our outlined future work will use inter-visit period as the basic metric to study trends in student enrollment during periods when high percentage of the traffic can be assumed to be caused by student accesses such as student class registration periods.

2.2 Path Analysis

Path analysis is the study of the sequence of page views by a user. The connectivity of pages in a Web site can be represented in a tree organization [10].

Average depth is the average number of pages visited by a user down a given path [10].

Average internal nodes accessed is the average number pages, characterized by facilitates navigation of the site, that were visited by a user [10].

Average leaf nodes accessed is the average number of pages, characterized by containing content, that were visited by a user. These metrics shed light on the amount of time spent navigating the site, as opposed to viewing actual content [10].

Entry point is the page where a user enters a site. This can be identified by analyzing differences between the incoming paths to the page and the number of requests made for the page. A small difference would indicate the users are relying heavily on the site topology [10].

Exit point is the last page of a site visited by a user. This can be identified by observing the last element in the path of requests [10].

Attrition is the measure of users who stop traversing a site via the links provided on a given page verses users who continue to traverse the site via its links [10].

Attrition curves are plots of attrition ratios for every page down a given path [10].

This study performs a temporal analysis, based on a real traffic trace as recorded by a Web server log.

3. Evaluation

Our testbed consisted of Dual 2.4 GHz Xeon processors with 1GB of memory and 3 U320 10K RPM SCSI drives in a RAID 5 configuration. For the purposes of this measurement, we used the Webalizer software, version 2.01.

The major headings in the Webalizer report are: hits, files, sites, visits, pages, and Kbytes.

Hits represent the total number of requests made to the server during the given time period [1].

Files represent the total number of requests that were answered and actually sent data back to the client. Not all hits will result in the server sending data. For example, *404-Not Found* requests and requests for pages that are already in the browsers cache are not recorded. By looking at the difference between hits and files, we can make an estimation of repeat visitors, since this difference translates into regular visitors accessing pages from cache [1].

Sites refer to the number of unique IP addresses that made requests to the server [1].

Visits record requests for a page made to the server for the first time. Since only page request will result in a visit, remotes sites that link to non- page URLs will not be counted [1].

Pages are requests made for files with extensions .htm, .html, and cgi [1].

Finally, a *Kbytes* (1024 bytes) is used to show the amount of data that was transferred between the server and the client [1].

The average size of Web pages is about 10 KB [2]. About 60% of all Web documents are of that size. The URL directory of the Web server we evaluate has a typical structure, where the Web page size ranges between 1KB and 40 KB, resulting in an average of 12 KB per Web page.

The input to our evaluation is a Web server log from the College of Design, Construction, and Planning at the University of Florida in Gainesville, Florida.

We first looked at geographical origination of requests. The college has student exchange cooperation agreements with the following countries: Finland, Australia, Sweden, Germany, Thailand, Costa Rica, and China. All the countries with existing agreements are listed in Table 1. This table also shows their corresponding placement in the top 30 list of most frequent request originations. The top 30 countries report generated by Webalizer lists the number of hits, files and Kbytes that characterize the request. The default list is limited to 30 and can be changed when configuring the Webalizer. A Boolean value is shown in each of the columns, indicating a placement in the top 30 list within a specific month.

Table 1 shows the specific months, from November 2003 to February 2004, during which the countries with existing agreements made the top 30 list. The top 30 list presents the top 30 countries from where the maximum requests were generated. It is interesting to note that out of the 7 countries with agreements only 1 made the top 30 list for each month and 2 did not make the list at all.

Table 1 . Months in which the 7 countries with exchange cooperation agreements made the “top 30 list”

	11/2003	12/2003	01/2004	02/2004
Germany	Yes	Yes	Yes	Yes
Australia	No	Yes	Yes	Yes
Costa Rica	Yes	No	No	Yes
Sweden	No	Yes	Yes	No
Thailand	Yes	No	No	No
Cuba	No	No	No	No
Finland	No	No	No	No

The results in Table 1 indicate that Germany shows the most interest in the exchange cooperation agreement, making the top 30 list in each of the 4 months under study. Australia shows high interest, making the top 30 list in 3 out of the 4 months. Costa Rica, Sweden, and China all show the same amount of interest distributed uniquely for each country in 2 out of the 4 months. Thailand shows little interest, making the top 30 list only during 1 out of the 4 months.

No requests originating from Cuba or Finland were found. Hence, those two countries never made the top 30 list. The results indicate that the cooperation with Cuba and Finland needs to be evaluated to find possible reasons for the low level of traffic origination from these countries. The administration may look into a possible course of action with Germany for guidance, since Germany has made the top 30 list every month for 4 months.

In addition, many other countries that are not part of the exchange cooperation agreement consistently made the top 30 list. This may be an indication that the administration should explore establishing a connection with these countries.

In February, Canada, United Kingdom, Italy, Netherlands, Mexico, Singapore, India, Greece, Brazil, Israel, France, Belgium, Japan, Turkey, Nicaragua, South Africa, Ireland, and Portugal made the top 30 list.

In January, Canada, United Kingdom, Netherlands, Italy, Dominican Republic, Brazil, Japan, Hong Kong, Greece, Turkey, France, Portugal, Mexico, India, Colombia, Taiwan, Belgium, Lithuania, and Israel made the top 30 list.

In December, Canada, United Kingdom, Greece, Italy, Belgium, Brazil, Turkey, France, India, Austria, Taiwan, Netherlands, Mexico, Japan, Poland, Norway, Singapore, and Hong Kong made the top 30 list.

In November, Canada, United Kingdom, Italy, Netherlands, Japan, Brazil, Poland, India, Turkey, France, Belgium, Spain, Greece, South Africa, Israel, Hong Kong, Singapore, and Portugal made the top 30 list.

Canada, the United Kingdom, Italy, Netherlands, India, Greece, Brazil, France, Belgium, Japan, and Turkey show very high interest, making the top 30 list for four consecutive months. Furthermore, Canada, the United Kingdom, Italy and the Netherlands are consistently in the top 4, except during the month of December, showing reduction in hits from the Netherlands.

Mexico, Singapore, Portugal, and Hong Kong made the top 30 list in 3 out of the 4 months under study. Israel, South Africa, and Poland made the top 30 list in 2 out of the 4 months. Nicaragua, Ireland, Dominican Republic, Colombia, Taiwan, Lithuania, Norway, and Spain made this list during 1 month.

Only 8 countries out of the top 30 were inconsistent during the 4 months. In other words, at least 16 countries (excluding those in the exchange agreement) that are not part of the exchange agreement show consistent interest in the college. An avenue for incorporating them into the exchange program should be explored. In particular, Canada, the United Kingdom, Italy, and the Netherlands, show a very high and steady interest in the college.

We next made an attempt to quantify student registration related increase in traffic. Table 2 shows the percentage of accesses originating from US educational, US commercial institutions, as well as internal (local campus) requests. Note that each column does not sum up to 100%, since international and unresolved by DNS traffic is not shown.

These results indicate that there was an increase in local traffic, possibly caused by pre-registration research of course related publications. During the month of January there was about 3.6% increase of local traffic (going from 470,007 total hits to 487,006 total hits), which corresponds to about 2% relative increase in local traffic as shown in Table 2.

4. Summary and Future Work

This paper has described an analysis of Web usage data based on Web usage statistics produced by Webalizer 2.01. We have evaluated the statistics on the Web site of the College of Design, Construction, and Planning at the University of Florida. We found trends in per country origination of accesses. We have also quantified Web site traffic increase during student registration periods – about 2% for the spring semester of 2004.

Table 2 . Percentage of accesses

	11/2003	12/2003	01/2004	02/2004
Educational	36.94%	29.66%	29.46%	32.48%
In Network	27.15%	27.90%	30.53%	28.54%
Commercial	16.14%	19.16%	17.06%	14.69%

This study has raised numerous issues for future work. For instance, Webalizer does not provide information about the states or cities in the United States that requests originate from. This data can be obtained by means of alternative techniques. We plan to work on an approach that goes beyond the scope of the Webalizer software. An interesting alternative is the use of third-party data collection software that provides robust databases for retrieving state and city information. Webalizer only queries a limited number of addresses in its resolution capabilities DNS database for the two-character country code based on ISO 3166. Using Webalizer in conjunction with alternative software, we can manipulate an IP address in such a way that a record for state and city can be found.

Another issue that we plan to address is that of access sessions. Much information can be gathered through the analysis of user sessions and navigation behavior. Navigation patterns can be extracted to improve the services provided by the Web site. The main goal is to make the access of online information more efficient. We can then follow this up by grouping similar users into communities based on their navigational behavior and use this information to build an adaptive Web site, which modifies its behavior according to the community into which it categorizes the user. This may be interesting for the Web site under study, since faculty, attending and non-attending students, and other colleges, are all possible communities that will have differing interests and navigational preferences.

In conclusion, we believe that discovering new techniques for Web usage analysis offers many promises for improving the efficiency of Web sites. This is an important field to focus on since the size of the Web and the societal dependency on on-line information becomes greater every day.

5. Acknowledgements

The authors would like to acknowledge the help of Julie Frey, the Director of Information and Publication Services and Todd Kisida, the Director of Information Technology for the College of Design, Construction, and Planning, at the University of Florida in Gainesville.

6. References

- [1] Barrett, Bradford L. "Home of the Webalizer." 2003. <http://www.mrunix.net> (8 March 2004).
- [2] Tim Bray, "Measuring the Web", *Proceedings of the Fifth International World Wide Web Conference*, Paris, France, May 1996.
- [3] Robert Buff, Arthur Goldberg, and Ilya Pevzner, "Rapid, Trace-Driven Simulation of the Performance of Web Caching Proxies", *Proceedings of WISP'98*, March 1998.
- [4] Ramón Cáceres, Balachander Krishnamurthy, and Jennifer Rexford, "HTTP 1.0 Logs Considered Harmful", *Proceedings of World Wide Web Consortium Workshop on Web Characterization*, Cambridge, MA, November 1998.

- [5] Lara D. Catledge and James E. Pitkow, "Characterizing Browsing Strategies in the World Wide Web", *Computer Networks and ISDN Systems*, vol. 26, iss. 6, pp. 1065-1073, 1995.
- [6] Ken-ichi Chinen and Suguru Yamaguchi, "An interactive prefetching proxy server for improvement of WWW latency", *Proceedings of the Seventh Annual Conference of the Internet Society (INET'97)*, Kuala Lumpur, June 1997.
- [7] Carlos R. Cunha, Azer Bestavros, and Mark E. Crovella "Characteristics of WWW Client-based Traces", *Technical Report TR-95-010*, Computer Science Department, Boston University, July 1 1995.
- [8] Fred Douglass, Anja Feldmann, Balachander Krishnamurthy, and Jeffrey C. Mogul, "Rate of Change and other Metrics: a Live Study of the World Wide Web", *Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS '97)*, December 1997.
- [9] James F. Kurose and Keith W. Ross, *Computer Networking, A Top-Down Approach Featuring the Internet, 3rd Edition*, Addison Wesley, 2006.
- [10] James Pitkow, "In Search of Reliable Usage Data on the WWW", *Proceedings of the Sixth International Conference*, Palo Alto, CA, 1997.