Internet Measurement: Understanding Web Server Workloads

Assignment 3, COMPSCI 314

Due: Oct 20, 2015 10 a.m.

1 Introduction

Data analytics is the process of examining raw data to find actionable information that can support decision making in organizations. Data analytics is a growing field, with demand for data scientists who can parse through large volumes of data to find meaningful patterns. These results can be utilized by organizations to improve their business processes.

Internet traffic measurement involves collecting network data that can be analyzed for several purposes such as traffic modeling, designing better network protocols, and traffic management. The growth in popularity of Web in the 1990s resulted in researchers trying to characterize Web traffic. These research works have utilized Web server logs to understand the workload characteristics of Web servers. The results of the research has led to improving performance of Web applications, designing better caching and load balancing techniques, and providing better user experience to clients, among other things [1, 3, 5, 6].

After graduating, you may go to industry and take up the job of an analyst at an enterprise. One of the duties of the analyst may entail understanding the traffic characteristics of the organization's Web servers, which is possible by analyzing the Web server logs. For example, analyzing the access patterns of users would allow the analyst to perform site updates during periods of reduced activity. The server logs can also be used for market research by a business organization [4]. Insights from the Web server analysis can be used to understand usage behaviour of the Web site, and help increase user visits to the site. For example, the business could perform appropriate search engine optimization to promote the site more effectively.

2 Web Server Access Logs

In this assignment, you will analyze a university Web server access log called `UofS_access_log` [2]. The server access log contains information about all requests made to the server and the corresponding server responses. The server log is in the fixed text-based *Common Log Format* and has the following syntax:

```
hostname - - [dd/mm/yyyy:hh:mm:ss time_zone] object response_code transfer_size
```

The `hostname` is the resolved name or IP address of the client making a request for an object stored on the Web server. The following fields (`- -`) are usually empty, but may contain user-identifier information such as username. The next field indicates the day and time the request was made along with the time zone. The URL requested is noted in the `object` field. The `response_code` field indicates the HTTP response code returned by the server. The `transfer_size` field records the bytes transferred by the server.

For example, the following is a line from the access log:

```
imhotep.usask.ca - - [15/Sep/1995:16:02:09 -0600] "GET /changes.html HTTP/1.0" 200 1254
```

This line represents a request made by host `imhotep.usask.ca` on September 15, 1995 at 4:02:09 p.m. The time zone is central time (GMT-0600). The host requested the HTML file called `changes.html` using HTTP version 1.0. This request was successfully completed by the server as shown by the status code 200. The server transferred 1254 bytes to the host `imhotep.usask.ca`.

3 Web Server Workload Analysis

You will analyze the log, and answer the following questions:

1. How many requests are made per day on average?
2. How many bytes were transferred during the entire log duration expressed in Megabytes (MB)?
3. What is the average number of bytes transferred per day expressed in MB per day?
4. Produce a breakdown of the server response codes expressed in percentage of the total number of requests. Group the status code as follows: *Successful, Not Modified, Found, Unsuccessful*. A successful response (status code: 200) means that the server received a request for a valid object (for which the client has the necessary access privilege), the object was found, and returned successfully to the client. A not modified response (status code: 304) means that the client already has a copy of the requested object in its cache, wants to verify if the object is up-to-date, and the client is informed that the object has not been modified at the server. A found response (status code: 302) results when the requested object is known to be stored in a different location than the URL requested by the client. The server responds with the new URL in this situation. A unsuccessful response (status code: 4XX and 5XX) happens when the requested object does not exist on the server, the client did not have access permission, or there was a server-side error.
5. How many requests are made by local clients and remote clients, respectively? Report your answer as a percentage of total requests. A local client is one containing `usask.ca` in the hostname or an IP address with `128.233.X.X`. All others are considered remote clients.
6. How many bytes are transferred by local clients and remote clients, respectively? Report your answer as a percentage of total bytes transferred.
7. Produce a breakdown of requests by file type category. The file categories are as follows: *Video, Sound, Dynamic, Formatted, HTML, Images, Others*. Report your answer as a percentage of total requests. The file categories by file extensions are described in Table 1.

Table 1: File categories

Category	File extension
HTML	html, htm, shtml, map
Images	gif, jpeg, jpg, xbm, bmp, rgb, xpm
Sound	au, snd, wav, mid, midi, lha, aif, aiff
Video	mov, movie, avi, qt, mpeg, mpg
Formatted	ps, eps, doc, dvi, txt
Dynamic	cgi, pl, cgi-bin
Others	Everything else

8. Using Table 1, produce a breakdown of bytes transferred by each file category. Report your answer as a percentage of total bytes transferred.
9. Using Table 1, calculate the average transfer sizes (in bytes) of each file category.
10. Identify all unique object requests in the log and sort them based on frequency. Next, identify all the objects that were requested only *once* in the log. What percentage of unique objects are accessed only once in the log? What percentage of bytes are accessed only once in the log?

For questions 5 through 10, your analysis should be based on successful requests only. Report your results to 2 decimal places. Some requests in the log may be malformed. It is safe to ignore these requests in your analysis. Please check that these requests account for a negligible fraction of the total requests.

4 Submission

Complete the attached `results.txt` file with your answers. Submit the file to Asisgnment Drop Box. A code template is being provided to help you get started with the assignment. For added challenge, you may write the parser and analysis scripts from scratch using a programming language of your choice. You are free to use online resources (e.g., online code, tools) as long as you provide appropriate attribution. *You do not need to submit the code.* You should keep the code, in case we wish to see it. You are encouraged to discuss the assignment with each other, however, the code and the produced results must be done individually.

Due to the tight marking deadline, no extensions will be given for this assignment. We will only accept assignments submitted on Assignment Drop Box. Late assignments (sent through email) will not be accepted. Please ensure that you have uploaded the correct file to Drop Box. Please make sure you have received a receipt from Drop Box after you have uploaded your answer. *Questions regarding the code template should be directed to the course tutor.*

5 Grading Scheme

Each question is worth 5 points. The maximum marks for the assignment is 50 points. This assignment counts for 5% of your total course grade. For each question, you will receive full points for the correct answer. You will receive 50% points for an answer, which is close to the correct answer. You will receive zero points for an answer that is far off from the correct answer.

References

[1] Martin Arlitt and Tai Jin, *A Workload Characterization Study of the 1998 World Cup Web Site*, IEEE Network **14** (2000), no. 3.

[2] Martin Arlitt and Carey Williamson, *Internet Web Servers: Workload Characterization and Performance Implications*, IEEE/ACM Trans. Netw. **5** (1997), no. 5, 631–645.

[3] Leeann Bent, Michael Rabinovich, Geoffrey M. Voelker, and Zhen Xiao, *Characterization of a Large Web Site Population with Implications for Content Delivery*, WWW **9** (2006), no. 4.

[4] D. Menascé, V. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, and W. Meira, *A Hierarchical and Multiscale Approach to Analyze E-business Workloads*, Perform. Eval. **54** (2003), no. 1.

[5] Venkata Padmanabhan and Lili Qiu, *The Content and Access Dynamics of a Busy Web Site: Findings and Implications*, Proc. ACM SIGCOMM, 2000.

[6] Weisong Shi, Y Wright, Eli Collins, and Vijay Karamcheti, *Workload Characterization of a Personalized Web Site and its Implications for Dynamic Content Caching*, Proc. WCW, 2002.