WPI-CS-TR-00-05 March 2000

Embedded Objects in Web Pages

by

Mikhail Mikhailov
Craig E. Wills

Computer Science
Technical Report
Series

WORCESTER POLYTECHNIC INSTITUTE

Computer Science Department
100 Institute Road, Worcester, Massachusetts 01609-2280

Abstract

An important characteristic of a Web page—the number of objects embedded in it—is difficult to obtain using publicly available sources of Web characterization data, such as proxy and server traces. It is important, however, to have an estimate for the number of embedded objects found in Web pages for modeling of realistic Web sites and workloads. We analyzed two data sets collected using active measurement technique and present our findings in this paper. We compare our results to those obtained by others, and show that the complexity of Web pages, in terms of the number of embedded objects, steadily increases.

Keywords: Web Characterization, Embedded Objects, Active Measurements, Statistical Analysis

1 Introduction

The investigation presented in this paper is an offspring from two previous studies [13, 9] and is based solely on the data collected as part of them. Our primary interest is to determine the number of objects embedded in popular Web pages—an important metric from the Web characterization point of view. Objects in this case are not only embedded images, but other resources, such as cascading style sheets and frames. This metric is required for modeling of realistic Web sites and workloads. For example, with the HTTP 1.0 protocol [3], a separate TCP connection is required to retrieve the base HTML object and every object embedded in it. The default persistent connection behavior of the HTTP 1.1 [6], however, allows multiple HTTP request/response pairs to be transmitted over a single TCP connection. Knowing a realistic distribution of the number of embedded objects is useful in modeling the use of persistent connections.

Embedded objects do not always translate directly into HTTP GET requests that can utilize a single TCP connection. Some embedded objects reside on the same server as the HTML page itself (we call it the *main* server), others may be on *auxiliary* server(s) [9]. For example, ad banner images often come from a Web site of the advertising company. Also, popular Web sites, in an attempt to improve end-user experience, offload some of their static content to a set of servers located geographically closer to the users. Akamai, for example, offers content providers to distribute their static images on Akamai servers, placed at Internet Service Providers (ISPs) around the world [1].

A secondary interest of our study is to determine if and how the number of embedded objects changes over time in a given set of popular Web pages. Less attention is devoted to this part of the project, partly because only a subset of the available data could be used for this analysis.

The rest of the paper is organized as follows. Section 2 discusses the data sets used in this study. The data analysis methodology is described in Section 3 and the results of the data analysis are presented in Section 4. An overview of the related research is given in Section 5. Section 6 outlines a number of discovered pitfalls. The paper concludes with the summary and future work in Section 7 and a list of references at the end.

2 Data Sets

The raw data sets used in this study were collected at different times, for different studies, by different people, using different software and methodology. In this section we discuss each data set in detail.

2.1 Data Set 1

This data set was collected in the Fall of 1999 for our previous study [13]. The collection was done as follows. First, we obtained a list of 50 most popular Web sites and a list of 50 most popular E-Commerce sites (business-to-consumer or home shopping) from the 100hot.com Web site [12]. 100hot.com ranks Web sites in various categories based on the number of hits they receive weekly: 100 most popular sites, 100 most popular shopping sites and so on.

We then wrote software to perform an unconditional HTTP GET for each URL in the two initial lists on a daily basis for over a week. Each retrieved HTML page was parsed, and a list of embedded images stored in a file. When our software detected the presence of frames, it fetched each frame along with the images embedded in it. Two resulting sets of raw data, called Top50 and ECom, comprise Data Set 1 (DS1).

2.2 Data Set 2

Gathered for another study [9], Data Set 2 (DS2) is more extensive than Data Set 1. The collection was performed as follows. A set of 711 popular Web sites was identified from a variety of sources, such as MediaMetrix [10], Netcraft [11], 100hot.com [12], Fortune 500 [7] and Global 500 [8]. For each site, the home page and all the embedded objects were retrieved. Retrievals were performed at three points in time: November 1999, December 1999 and February 2000. Unlike in the DS1 case, the [9] study focused on the end-to-end performance and the retrieval software simply counted the number of embedded objects discarding them upon retrieval. The software also noted how many embedded objects came from the *main* server and how many were fetched from *auxiliary* server(s). Lists of (count;server-name) pairs comprise DS2.

3 Methodology

The methodology of this study is straightforward. We wrote a set of Perl scripts to process and analyze the raw data in both DS1 and DS2. For each data set, the software computes the MIN, MAX, MEAN and MEDIAN number of objects and plots a cumulative distribution function. For DS1, since we had over a week worth of retrievals for each object, we first computed the average number of embedded objects per page, and then used these averages in all our statistical calculations.

For DS2 set we performed our analysis for all objects first, as if they all came from the same main server, and then only for the objects which came from the main server.

4 Results

Cumulative distribution functions of the number of embedded objects per HTML page in DS1 are shown in Figure 1. Results indicate that only a small fraction of pages (under 10%), both in Top50 and ECom sets, have no embedded objects. This percentage might be overestimated, however, due to reasons discussed in Section 6. Results also show that roughly the same percentage of home pages for the most popular Web sites and most popular E-Commerce sites (a little under 40%) have about 10 embedded objects. About 50% of pages from both Top50 and ECom sets have between 10 and 40 embedded objects, but ECom pages in this category are slightly *heavier* than Top50 pages in terms of the number of embedded objects. The top 10% of the pages in Top50 and ECom sets, however, compare differently: Top50 pages embed more objects than ECom pages.

Results for Data Set 2 are shown in Figure 2. The first two curves, Nov-99 and Nov-99(*), show the cumulative distribution of the number of embedded objects as sampled in

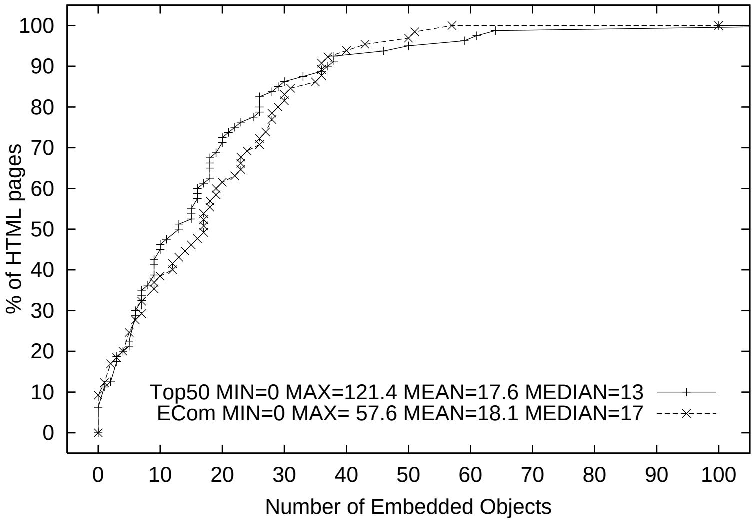

Figure 1: Cumulative Distribution Function for the Number of Embedded Objects in DS1

November 1999. The curve marked with (*) takes into account that some objects came from servers other than the main server and represents only embedded objects retrieved from the main server. The unmarked curve considers all embedded objects. The two curves are slightly apart, which shows that a number of sites offload some objects to other servers. The two curves are fairly close together, however, which is an indication that either only a small subset of sites studied explicitly use more than one server for their content or that sites offload only a small fraction of their embedded objects. The Feb-00 and Feb-00(*) curves are similar to the first two, except they represent data sampled in February 2000.

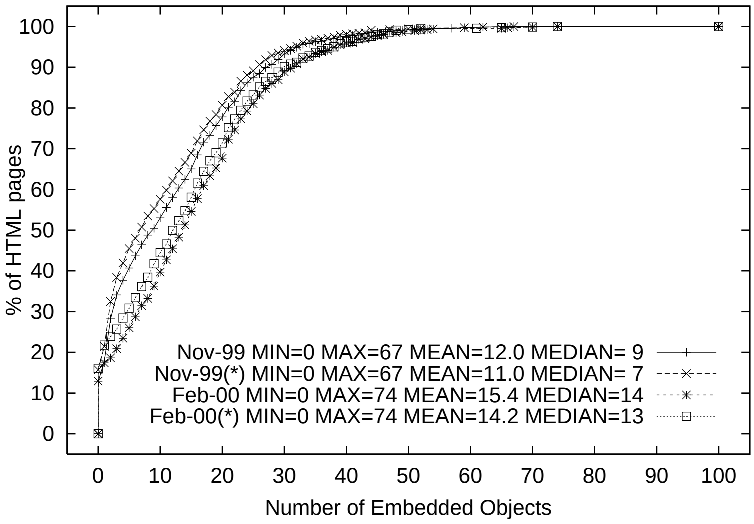

Figure 2: Cumulative Distribution Function for the Number of Embedded Objects in DS2

A more interesting comparison is between Nov-99 and Feb-00 curves or Nov-99(*) and Feb-00(*) curves. Pairs of the named curves are far apart from each other. This indicates that popular sites have enhanced their home pages with more embedded objects, most likely images. For all sites in DS2, over the period from November 1999 through February 2000, the average number of embedded objects increased by 3 and the median increased by 5-6. The increase is significant enough to warrant an additional investigation. The part of DS2 collected in December 1999 is not shown on the graph because the two curves representing it fall in between the four curves already shown, closer to the two right-most curves for February 2000.

Comparison of Figures 1 and 2 reveals that 2%-9% more pages in DS2 than in DS1 have no embedded objects. About 10% of pages from DS1 have over 40 objects, compared to about 5% in the DS2.

Table 1 shows a list of popular sites from DS1 with a number of embedded objects from small (under 10) to large (over 60) to give an idea which popular sites contain how many embedded objects. Since the set of objects on a page changes the numbers in the table are averages across multiple retrievals.

Table 1: Number of Embedded Objects for the Selected Sites

Site	Ave. Number of Embedded Objects
www.excite.com	7.4
www.microsoft.com	9.4
www.aol.com	13.0
www.amazon.com	18.3
www.onsale.com	19.0
www.ebay.com	26.9
www.etoys.com	36.4
www.cnn.com	38.2
www.fortunecity.com	46.1
www.usatoday.com	61.4

5 Related Research

Few published papers quantify the number of objects embedded in Web pages. Part of the reason so little data is available on this subject is because in order to determine the number of embedded objects in a given page, the contents of the page must be examined. Most of research on Web characterization has used proxy or Web server logs as a source of data. Since such logs do not contain contents of the retrieved objects, the required information is not available.

One early study on Web characterization was presented by Tim Bray [4] in May of 1996. Data collection was done in November 1995, over 4 years ago. The data set was larger than ours—over 1.5 million pages, and did not specifically focus on popular pages. Also, Bray's paper explicitly talks about embedded images, while our study included other embedded objects. In Bray's work, the median number of objects embedded on a page was 1, while in our study the median is between 7 and 17, depending on a data set. Bray's results showed that over 50% of all examined pages had at least 1 object. For both of our data sets, DS1 and DS2, that number is about 90%—a significant increase. A little over 45% of the pages had no images at all. In our case, only about 10% of the pages fall into that category. In our data sets, however, the largest number of embedded objects encountered was only 120+, while [4] reports a small percentage of pages with 256+ images. The reason our study did not find any pages with that many objects is because our sample size was small. One interesting piece of data from Bray's work is the percentage of pages with exactly 1 image—15%. Our results report a negligibly small percentage of such pages.

Another related work is by Paul Barford and Mark Crovella [2]. These researchers focused on modeling and reproducing realistic Web workloads and built a Scalable URL Reference Generator (SURGE). They went beyond counting a number of objects contained in a typical Web page and developed a distributional model. Barford and Crovella used traces of actual Web accesses recorded by a set of specifically instrumented Web browsers from November 1994 through February 1995 at Boston University [5]. As authors point out in [2], "... number of [embedded references] is difficult to extract from client trace data since there is typically no record in the data that indicates which documents are embedded". Barford and Crovella inferred the number of embedded objects from the traces by identifying sequences of file transfers initiated by a given user with the times between the transfers falling within a predefined threshold of one second. Using statistical methods on the extracted data, researchers determined that the Pareto distribution provided the best fit. Analogous to Bray's findings, Barford and Crovella report that over 45% of pages contained no embedded objects, and a small percentage of pages contained over 100 objects.

6 Pitfalls

A few issues with the data collection phase of the study affected our results.

- **Re-Direction**. Many Web sites, due to various reasons, use the HTTP re-direction mechanism by sending an HTTP 302 response code [6] and a new location for the requested content in the HTTP response header. When a browser receives such a code, it immediately retrieves the new URL. While our retrieval software properly understands re-direction, and fetches the content from its new location, it does not notify the analysis software that re-direction occurred. Even though re-direction is not an embedded object in itself, it does require one more GET request, and should be counted.

- **JavaScript**. Our HTML parser does not parse JavaScript code embedded between `<script>` and `</script>` tags within HTML. Some sites exploit JavaScript's ability to add HTML markup and text to the page on the fly. We have observed cases, for example, where JavaScript code was used to build frames and our parser failed to identify the presence of frame references and thus underestimated the number of objects embedded in a fully rendered page.

- **Refresh**. It is possible to include a meta header into HTML code, which will be interpreted by the browser. One such header is REFRESH. It specifies a timeout value

and a new URL to fetch. For example,
`<meta http-equiv="REFRESH" content="0; url=splash/">`.
Some sites use an empty HTML page as their home page and deploy this header to force the retrieval of another (real) page. Even though the real page might contain frames and images the initial page has no content. Our retrieval software does not understand the REFRESH header at this point, and thus underestimates the number of embedded objects.

7 Summary and Future Work

In this paper we have presented our study of the number of embedded objects in popular Web pages. We made an attempt to account not only for images but also for other embedded objects, such as cascading style sheets and frames. Comparison of our findings and the results obtained by others [4, 2] four years ago indicates a dramatic increase in the complexity of Web pages in terms of the number of embedded objects. Our results indicate that currently few popular pages embed no objects, and that the number of embedded objects in popular pages increases even over short time periods. We also discovered that home pages for the on-line shopping sites contain more embedded objects than other sites in our study, which is probably due to the fact that E-Commerce sites display more product related information, such as photographs. We believe the trend is towards the increase in complexity of Web pages, at least popular ones, and in the future we should expect Web pages to embed more objects.

We encountered a couple of pitfalls undermining the accuracy of our results. Diversity in the browser technologies used by content providers complicates the active data collection process. A more powerful retrieval agent, mimicking functionality of popular browsers, is required. This is left for future work.

Few research projects focused on the *number* of embedded objects, while many studied the proportion of *bytes* contributed by embedded objects to the overall size of a fully rendered page. We neglected the latter metric in our study, but plan to investigate it in the future.

References

[1] Akamai. http://www.akamai.com.

[2] Paul Barford and Mark Crovella. Generating Representative Web Workloads for Network and Server Peformance Evaluation. In *Proceedings of the SIGMETRICS Conference*, pages 151–161, June 1998.
http://www.cs.wpi.edu/~sigmet98/barford.ps.

[3] Tim Berners-Lee, Roy T. Fielding, and Henrik Frystyk Nielsen. Hypertext Transfer Protocol—HTTP/1.0. RFC 1945, May 1996.
http://www.ietf.org/rfc/rfc1945.txt.

[4] Tim Bray. Measuring the Web. In *Fifth International World Wide Web Conference*, Paris, France, May 1996.
http://www5conf.inria.fr/fich_html/papers/P9/Overview.html.

[5] Carlos R. Cunha, Azer Bestavros, and Mark E. Crovella. Characteristics of WWW client-based traces. Technical Report BU-CS-95-010, Boston University, Department of Computer Science, July 1995.
ftp://cs-ftp.bu.edu/techreports/95-010-www-client-traces.ps.Z.

[6] Roy T. Fielding, James Gettys, Jeffrey C. Mogul, Henrik Frystyk Nielsen, Larry Masinter, Paul J. Leach, and Tim Berners-Lee. Hypertext Transfer Protocol—HTTP/1.1. RFC 2616, June 1999.
http://www.ietf.org/rfc/rfc2616.txt.

[7] 1999 Fortune 500 companies. Fortune volume 139 number 8, April 1999.

[8] 1998 Global 500 companies. Fortune Magazine, 1998.

[9] Balachander Krishnamurthy and Craig E. Wills. Analyzing factors that influence end-to-end web performance. In *Ninth International World Wide Web Conference*, Amsterdam, Netherlands, May 2000.
http://www.cs.wpi.edu/~cew/papers/www9/e2e.html.

[10] Media Metrix. http://www.mediametrix.com.

[11] The Netcraft Web Server Survey. http://netcraft.co.uk/survey/.

[12] 100hot.com. http://www.100hot.com.

[13] Craig E. Wills and Mikhail Mikhailov. Studying the impact of more complete server information on web caching. Submitted, February 2000.