

Summary of WWW Characterizations

James E. Pitkow
Chair, HTTP-NG WCG
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto CA 94306 USA
pitkow@parc.xerox.com

Abstract

To date there have been a number of efforts that attempt to characterize various aspects of the World Wide Web. This paper presents a summary of these efforts, highlighting regularities and invariants that have been discovered.

Keywords: Statistics, Metrics, Analysis, and Modeling

Introduction

In August 1997, the World Wide Web Consortium embarked upon the HTTP-NG (Next Generation) project [HTTP-NG 1997] to redesign the Web's underlying object model and protocols. In order to determine the efficiency of the new protocol, controlled experiments need to be conducted that compare the performance to HTTP 1.1. To ensure the validity of the tests, the scenarios must be as representative of actual Web traffic as possible. The current plan is to include model-based scenarios representing corporate, educational, and home usage. Given the amount of work required to develop a new protocol and create representative testing scenarios, two groups within HTTP-NG were formed: the Protocol Design Group (PDG) and the Web Characterization Group (WCG) [WCG 1997]. As part of the ramping up phase in building the test scenarios, the WCG collected and reviewed the existing body of research that characterizes WWW usage. This paper will summarize prior Web characterization efforts. Research that analyses client, proxy and server traffic will be examined followed by a review of attempts to characterize the entire World Wide Web. While the below summary of papers is meant to be complete, unintentional errors and omissions undoubtedly are present—apologies in advance. In addition, given the space constraints of this paper, the WCG recommendations for client, proxy, server, and Web characterizations and a discussion the representative scenarios generated for the testbed will be presented in future papers.

Summary of Characterizations

The decentralized nature of the Web makes measuring and gathering representative characterizations of Web usage difficult. As reviewed by [Pitkow 1997], various infrastructure, privacy, and protocol level issues can make even the simplest of metrics, like determining how

many people visit a Web page, difficult to obtain. This difficulty is expressed as a common thread in the literature that attempts to analyze various portions of the Web. Examination of these difficulties and recommending solutions are also part of the WCG mission.

Client

Characterizations of user behavior captured at the browser are the most informative yet are quite rare. The scarcity of browser level analysis is primarily due to the difficulty in instrumenting WWW browsers and/or ensuring that the data collection environment is not biased towards a particular user segment. Despite this, method like channeling requests through proxies, monitoring operating system events and privacy flaws in various scripting languages enable characterizations to be performed today. Still, without source code, none of the above techniques permits all user interface events to be captured, limiting the scope and accuracy of the results.

The first study to characterize WWW client behavior was performed by [Catledge and Pitkow 1995] during the summer of 1994. The study was conducted for a three-week period using a fully instrumented version of Xmosaic at the Georgia Institute of Technology. The primary contributions of the 107 person study were a) the verification of the overwhelming use of navigational features (over 80%), especially the "Back" button, b) the empirical basis for 25 minutes session timeouts, and c) a new classification method that categorizes user modes as searching, browsing, and serendipitous based up the length of navigation subsequence patterns.

Shortly thereafter during the fall of 1994, the Boston University Oceans Group conducted a much larger study—over 600 users during a six-month period—also using an instrumented version of Xmosaic that focused solely on navigation events. Many interesting heavy tail distributions were documented in [Cunha et al 1995] and later in [Crovella et al 1997] and [Cunha 1997] including: a) document size versus the number of documents (Pareto) b) document size versus the number of accesses (Power, with smaller files being most frequently accessed) and c) a reconfirmation of the applicability of Zipf's Law for document popularity versus rank noted initially by [Glassman 1994]. Additionally, over 75% of the requests were to off campus servers. A comparison of file sizes to UNIX file sizes and implications for caching algorithms were also presented. An application level specific caching analysis using the same data can be found in [Bestavros et al 1995].

The same data set was used to demonstrate and explain the self-similar nature of WWW traffic [Crovella and Bestavros 1995]. Their explanation uses the Pareto distribution of file sizes, Web transmission times (also Pareto), inter-request times, and the inverse relation between the number of times a document is requested and the size of the document. From this, [Crovella and Bestavros 1995] conclude that the self-similar nature of Web traffic is not likely to be machine induced.

During the tail end if 1995, [Tauscher 1996] conducted a study of 28 university students that used a fully instrumented version of Xmosaic for a six-week period. An extensive analysis was performed on the user interface events and compared to [Catledge and Pitkow 1995]. Additional metrics and results were introduced including: a) URL vocabulary growth rate (the rate new URLs are visited as a function of all requests), b) reoccurrence rate (the rate previously visited URLs are revisited), c) temporal and spatial locality (the number of items revisited across time) and d) an analysis of longest repeating subsequences that went into significantly more depth than

[Catledge and Pitkow 1995]. The results were used to measure the effectiveness of the algorithms used to manage bookmarks/favorites. A concise summary of the findings can be found in [Tauscher and Greenberg 1996].

Proxy and Gateways

Another way to characterize user behavior is by analyzing proxy traffic [Luotonen and Altis 1994]. Most of the proxy research focuses on improving caching algorithms and therefore does not typically specify general parameters. This makes comparing results and determining which distributions to use in HTTP-NG testbed scenarios difficult. It is hoped that future proxy trace research will include a section that describes the distributions and characterizations of various parameters.

The first characterization of WWW proxy traffic was performed by [Glassman 1994] using the Digital Equipment Company (DEC) proxy-cache. The proxy was established in January of 1994 and served nearly 300,000 pages to 600 different users, averaging 4,000 requests per day from 100 users. Several interesting observations were drawn. First, document popularity followed a Zipf distribution. Second, one third of the requested pages were not in the cache (cache misses), another third were in the cache (cache hit), and the final third were in the cache but no longer valid (cache invalid). Finally, the rate pages change was found to be uniform. As a comparison, higher hit rates were reported by [Smith 1994] for the HENSA proxy in the United Kingdom and lower hit rates by [O'Callaghan 1995] for a proxy in Australia during September 1994 through February 1995 and for the national cache in Japan [Nabeshima 1997].

Also using a corporate Intranet, [Sedayao 1994] characterized the distribution of mime types, mime type traffic, domains accessed, and temporal nature of WWW traffic generated by roughly 800 people at Intel during spring 1994. [Sedayao 1994] found that images are the most requested items and account for the most traffic. As was true in the early client studies, most of the domains accessed via the Intel network were educational. Another DEC proxy trace was collected by [Mogul 1996].

With the intent of determining the lifetime of documents, [Worrell 1994] conducted an experiment that gathered URLs from a variety of Harvest Caches, publicly readable browser history files, the top Harvest Web pages, and by a modified version of Xmosaic. Several interesting findings resulted including: a) the majority of items have small lifetimes (under 50 days), b) the time between modifications was found to vary significantly, with variance increasing as a function of time since last modification, and c) HTML files are more likely to be modified than images (mean lifetime for HTML files was 75 days versus 107 days for images).

One of the more prolific research groups studying proxies is the Network Resource Group at Virginia Tech. Using data from three student populations at Virginia Tech during the spring of 1995, [Abrams et al 1995] report a maximum hit rate between the range of 30-50%. This range compare well with the corporate hit rate published by [Glassman 1994] a year. Although distributions are not provided, the paper contains the location of the collected data files from which the distributions could be derived. The traces were expanded in [Williams et al 1996] to include proxies at various points in the Virginia Tech network. Interestingly, they show that the number of requests to a server also shows a Zipf distribution. Traces from the same sources albeit a later time (fall of 1996) were used in [Wooster and Abrams 1997] (see [Wooster 1996]

for in depth analysis) to compare caching algorithms as well. [Williams et al 1996] and [Wooster and Abrams 1997] contain a media type breakdown for each data set.

[Gwertzman and Seltzer 1996] used data from a proxy at Microsoft, a Boston University Web server, and several Harvard College Web servers during 1996 to determine mime type popularity and average life span. Confirming earlier work by [Worrell 1994], the average life span of an HTML document was 50 days, with media being 1,460 days. They also found that while images compose over 65% of the total number of requests they are the least likely to change. As far as modeling proxy traffic, both [Bolot and Hoschka 1996] and [Abdulla et al 1997a] have successfully used time series analysis to model and forecast Web traffic at proxy caches. Finally, [Scheuermann et al 1997] reconfirmed the distribution of file size by popularity using a two-day trace from students at Northwestern University during November 1996.

[Gribble and Brewer 1997] performed an interesting analysis of dial-in users with connections less than 28.8 kb/sec at the University of California at Berkeley for 45 days starting October 1996. They confirm that while self-similar network behavior occurs on minor scales (microsecond and seconds) as noted by [Crovella and Bestavros 1995][Crovella and Bestavros 1996], large time scales (hours, days, weeks) show strong periodic behavior as noted by [Bolot and Hoschka 1996][Abdulla et al 1997a]. Additionally, locality of reference, and hence cache performance, is demonstrated to be a function of the number of users the system supports.

Finally, [Abdulla et al 1997b] completed an extensive analysis of proxy logs from ten sources. These sources include: the two 1996 DEC proxy traces, the undergraduate and graduate 1994-1995 client BU traces, the three 1996 departmental proxy Virginia Tech proxy traces, a fall 1995 trace from the Korean national proxy, a few minute trace from AOL during December 1996, and a 1996 proxy trace from Auburn high school. Using [Arlitt and Williamson 1996] as a guide, [Abdulla et al 1997b] identified nine invariants. The regularities found in common include a) the median file size is approximately 2KB, b) the mean file size is less than 27 KB, c) 90-98% of the accessed files are either HTML, image, or CGI-map files, d) images account for the most traffic, e) less than 5% of the servers are accessed only once, f) less than 12% of the accesses are to unique servers, g) 25% of the servers are responsible for 80-95% of all accesses, f) 90% of the bytes are from 25% of the of servers, and g) 88-99% of the requested files were retrieved successfully.

Server

While nearly every Webmaster performs some form of Web site traffic analysis, the follow summary focuses only on scientific attempts to characterize site usage. As reviewed in [Crovella et al 1997], the Web, especially on the server side, is filled with many heavy tailed distributions and regularities, e.g., file size, transmission time, document popularity. Future research will hopefully refine our existing understanding of Web site usage as well as incorporate more diverse log files.

Within the educational domain, several studies were performed during the 1994. [Braun and Claffy 1994] analyzed the cluster of servers at NCSA during August 2-3 1994. Although the focus of [Braun and Claffy 1994] was on the geographical nature of WWW requests with respect to distributed caching, the paper contains several time and size dependent distributions from NCSA's server, one of the busiest WWW servers at the time (received over 500

requests/minute). At the same conference, [Pitkow and Recker 1994] published the results of applying a model of human memory to the January through March 1994 traffic from the Georgia Institute of Technology WWW server. Their findings suggested that recency (the time since last access) is a better predictor than frequency of future access using a granularity of one day. The paper also documented a cache-hit rate of nearly 80% using the same temporal granularity.

One of the most comprehensive analysis of a WWW server during 1994 was performed by [Mogul 1995] using the November 9, 1994 Californian Congressional election server setup by DEC. The site received over 1500 requests/minute during peak periods. [Mogul 1995] modified the server to record connection duration, number of disk I/O, and CPU usage in addition to normally logged fields. Numerous important findings were made including: a) no correlation existed between file size and connection time for files under 30K, b) the majority of traffic was consumed by small images, c) the cumulative distribution of requestors as a function of the number of requests is log-log, and d) the inter-arrival time of requests did not appear to follow a pure Poisson process. This latter point lends support for the self-similar nature of Web server traffic as put forth in [Lelend et al 1993] for Ethernet traffic and [Crovella and Bestavros 1995] for client generated HTTP traffic.

Another early attempt to model server behavior was done by [Burchard 1995]. File size, external link references, and session lengths are shown to generally fit lognormal distributions while the time between requests from the same host were modeled as exponential distributions.

In an extensive analytical study, [Arlitt and Williamson 1996] identified ten invariants using six different data sets that spanned the course of one week to one year. The data sets included the departmental servers at the University of Waterloo and the University of Calgary, a campus server at the University of Saskatchewan, the Kennedy Space Center (NASA) server, a commercial ISP server in Baltimore, Maryland, and NCSA's server. The first invariant discovered was that 88% of all requests were transfers of the content, with another 8% being requests where the client was checking to see if had been modified. Second, images and HTML files account for 90-100% of the files transferred, with images typically being requested far more than HTML files. This is consistent with the client side observations of [Sedayao 1994] and [Cunha et al 1995]. Third, the mean file transfer size was below 21K for all requested objects, confirming the observations by [Braun and Claffy 1994]. Only 3% of the requests were for separate (distinct) files and one third of the files and bytes accessed are accessed only once.

The fifth invariant presented was the Pareto distribution of file sizes (previous noted on the client side by [Crovella and Bestavros 1995]). The file inter-reference times were found to be independent and exponentially distributed a finding that also confirms [Burchard 1995]. Another invariant is that 10% of the documents account for 90% of all request and bytes transferred. Temporal locality, as measured by least recently used stack distance, was found in all log files as well, further clarifying the importance of recency introduced by [Pitkow and Recker 1994]. The ninth invariant discovered was that remotes sites account for over 70% of the requested files and over 60% of the requested bytes (similar to the client side findings of [Cunha et al 1995]). The final invariant uncovered was that 10% of the thousands of domains accessing the servers account for over 75% of the usage (similar to the proxy trace findings of [Abdulla et al 1997b]).

Shortly thereafter, [Almeida et al 1996] performed analysis on access logs from NCSA, SDSC, EPA, and BU during the fall of 1996. Zipf's Law was shown to apply to documents popularity

for documents served by Web sites as well as for sequences of requests from clients ([Glassman 1994] had shown this for documents requested by Web clients). Despite this, the Zipf-based model was not able to synthetically generate representative workloads. [Almeida et al 1996] argue that this is due to the failure of the Zipf-based model to capture spatial and temporal locality of reference, a property that was documented on the client side by [Tauscher 1996]. Their study provided further support for the limited utility for LAN based caching due to the lack of shared common interests as originally predicted by [Glassman 1994] and reiterated by [Abrams et al 1995].

Survival analysis of files within the Gvu Center at Georgia Institute of Technology during a 226-day period starting in 1996 was performed by [Pitkow and Pirolli 1997]. The origin of accesses was used to stratify the data, where files were classified as either being primarily accessed from users within Georgia Tech, outside of Georgia Tech, or accessed equally from both communities. Internally driven files were more likely to be deleted, than mutually and externally driven files. Confirming [Worrell 1994] and [Gwertzman and Seltzer 1996], HTML files were found to have shorter lifecycles than non-HTML files. Finally, they found the more frequently a file has been modified, the greater the chances of file being eventually deleted.

In his undergraduate thesis, [Manley 1997] examined the access log files from ten different sites. The sites were composed of the following servers: Harvard Engineering, Harvard Arts and Sciences, Rice University Engineering, an adult content, a professional organization, a government agency, a free software company, a Web site designer, and a traditional business. All logs were from 1996 or early 1997 and spanned in most cases over a year. Unlike previous analyses, [Manley 1997] focused on the changes in access patterns and Web site's content to categorize sites. The primary characteristics that categorize a site are the growth functions of the following metrics: a) the number of Web users, b) whether the site experiences a redesign, c) the number of documents on the site, d) the number of documents visited per user, e) the number of search engine hits, and f) whether the site charges for access. Finally, the results challenge the myth that CGI is becoming increasingly more important and that the main cause of latency is heavily loaded servers. A more concise summary of the central factors effecting the growth of Web sites can be found in [Manley and Seltzer 1997].

WWW

One of the earliest uses of citation analysis on the Web, if not the earliest published account, was performed by [Mauldin 1994] using the Lycos search engine. Of the 335,000 documents retrieved at the time, the Genome Database was the most cited, followed by a now extinct educational server, an RFC FTP site, CERN, and NCSA. For this analysis, citations from within host site were not excluded, hence the probable cause for the Genome site being the most cited.

[Woodruff et al 1996] used the Inktomi collection of 2.6 million documents as of November 1995 to characterize the Web. Many interesting descriptive statistics were provided although no attempts were made to model the distributions. The statistics presented include: a) the mean size for HTML documents was 4.4 KB with a median size of 2 KB and a maximum size of 1.6 MB, b) the average number of HTML tags per page was 71 with an average of 11 unique tags, c) the most cited sites were Xerox PARC, Yahoo, and The Cool Site of the Day (though as with [Mauldin 1994], self referencing citations were not excluded).

Invariant	Sources	Metric
Requested file popularity	[Glassman 1994] [Cunha et al 1995] [Almeida et al 1996]	Zipf Distribution
File sizes (requested and from entire Web)	[Cunha et al 1995][Bray 1996][Woodruff et al 1996] [Arlitt and Williamson 1996]	Heavy tailed (Pareto) with average HTML size of 4-6 KB and median of 2 KB, images have an average size of 14 KB
Traffic properties	[Sedayao 1994][Cunha et al 1995][Arlitt and Williamson 1996]	Small images account for the majority of the traffic and document size is inversely related to request frequency
Self-similarity of HTTP traffic	[Crovella and Bestavros 1995] [Gribble and Brewser 1997]	Bursty, self similar traffic between the micro second and minute time range
Periodic nature of HTTP traffic	[Bolot and Hoschka 1996][Abdulla et al 1997a] [Gribble and Brewer 1997]	Periodic traffic patterns able to be model by time series analysis at the hour to weekly time range
Site popularity	[Arlitt and Williamson 1996] [Abdulla et al 1997b]	Roughly 25% of the servers account for over 85% of the traffic
Life span of documents	[Worrell 1994][Gwertzman and Seltzer 1996]	Around 50 days, with HTML files being modified and deleted more frequently then images and other media.
Occurrence rate of broken links while surfing	[WCG 1997-Xerox PARC, Virginia Tech]	Between 5-8% of all requested files
Occurrence rate of redirects	[WCG 1997-Xerox PARC, Virginia Tech]	Between 13-19% of all requested files
Number of page requests per site	[Huberman et al 1997][Catledge and Pitkow 1995][Cunha et al 1995]	Heavy tailed (Inverse Gaussian) distribution with typical mean of 3, standard deviation of 9, and mode of 1 page request per site
Reading time per page	[Catledge and Pitkow 1995][Cunha et al 1995]	Heavy tailed distribution with an average 30 seconds, median of 7 seconds, and standard deviation of 100 seconds
Session Time outs	[Catledge and Pitkow 1995][Cunha et al 1995]	25 minutes, with mean time of 9 minutes

Appearing at the same conference was a paper by [Bray 1996], which used the Open Text collection of 1.5 million documents as of November 1995. [Bray 1996] reports several similar findings, notably: a) the mean file size is between 6.5 KB, with a median size of 2 KB, b) slightly over 50% of all HTML pages contain at least one embedded image, with 15% of the pages containing exactly one image, c) 75% of all pages contains at least on hyperlink, with 15% of the pages containing no hyperlinks, d) over 80% of the sites are linked to by between 1 and 10

other sites, e) 80% of the sites contain no hyperlinks to other sites (which means a few sites are doing most of the navigation), f) NCSA, MIT, Yahoo, CMU, and Netscape respectively were the most linked to sites by other sites, and g) 44% of the files did not have a mime type extension, 36% were HTML, and 3% were GIF, which reverses the findings of media type requests and bytes transferred noted at the client and server level.

Conclusion

The above table summaries several invariants discovered across the client, proxy, server, and Web studies reviewed in this paper. Additionally, a few late breaking findings of the WCG not presented in the paper are also included. One of the graver shortcomings of the characterizations to date is the focus on US education settings and the subsequent scarcity of client characterizations from commercial Internet Service Providers, corporate environments, and international users. These sectors represent significant portions of today's WWW traffic and may have very different usage patterns. Another shortcoming of the research is that many of the studies reviewed about are out dated. While one could argue that Web usage behavior has not changed significantly in the past year or more, empirical evidence is necessary to construct representative scenarios of usage. Still, despite limitations of client monitoring and reliable usage measurement issues, the amount of high quality research being done that characterizes the Web continues to increase and will hopefully yield more characterizations.

References

- [Abdulla et al 1997a] Abdulla, G., Nayfeh, A., and Fox, E. (1997). Modeling proxy traffic using Fourier analysis. Unpublished document.
- [Abdulla et al 1997b] Abdulla, G., Fox, E., and Abrams, M. (1997). [Shared user behavior on the World Wide Web](#). WebNet97, Toronto, October 1997.
- [Abrams et al 1995] Abrams, M., Standridge, C. R., Abdulla, G., Williams, S., and Fox, E. (1995). [Caching proxies: limitations and potentials](#). The World Wide Web Journal 1(1).
- [Almeida et al 1996] Almeida, V., Bestavros, A., Crovella, M., and de Oliveira, A., (1996). [Characterizing reference locality in the WWW](#). Proceedings of PDIS'96: The IEEE Conference on Parallel and Distributed Information Systems, Miami Beach, Florida, December 1996.
- [Arlitt and Williamson 1996] Arlitt, M. and Williamson, C. (1996). Web server workload characterization: the search for invariants. ACM SIGMETRICS Conference, Philadelphia, Pennsylvania, May 1996.
- [Braun and Claffy 1994] Braun, H. and Claffy, K. (1994). [Web traffic characterization: an assessment of the impact of caching documents from NCSA's Web server](#). In Proceedings of the Second International WWW Conference, Chicago, Illinois, October 1994.
- [Bestavros et al. 1995] Bestavros, A., Carter, R., and Crovella M. (1995). [Application-level document caching in the Internet](#). In Proceedings of the Second International Workshop on Services in Distributed and Networked Environments (SDNE '95). Whistler, Canada, June 1995.
- [Bolot and Hoschka 1996] Bolot, J., and Hoschka, P., (1996). [Performance engineering of the World Wide Web: Application to dimensioning and cache design](#). In Proceedings of the Fifth International WWW Conference, Paris, France, May 6-10, 1996.
- [Burchard 1995] Burchard, P. (1995). [Statistical properties of the WWW](#).

- [Bray 1996] Bray, T. (1996). [Measuring the Web](#). The World Wide Web Journal 1(3).
- [Catledge and Pitkow 1995] Catledge, L. D. and J. E. Pitkow (1995). [Characterizing browsing strategies in the World-Wide Web](#). Computer Networks and ISDN Systems 26(6): 1065-1073.
- [Crovella and Bestavros 1995] Crovella, M., and Bestavros, A. (1995). [Explaining World Wide Web traffic self-similarity](#). Tech. Rep. BUCS-TR-95-015, Boston University, CS Dept, Boston, MA 02215, August 1995.
- [Crovella and Bestavros 1996] Crovella, M. and Bestavros, A. (1996). [Self-similarity in World Wide Web traffic: Evidence and possible causes](#). In Proceedings of SIGMETRICS'96: The ACM International Conference on Measurement and Modeling of Computer Systems. Philadelphia, Pennsylvania, May 1996.
- [Crovella et al 1997] Crovella, M., Taquu, M., and Bestavros, A. (to appear 1997). Heavy-Tailed probability distributions in the World Wide Web. In Applications of Heavy-Tailed Probability Distributions (Adler, Feldman, and Taquu ed.), Birkhauser, Boston.
- [Cunha et al. 1995] Cunha, C., R., Bestavros, A., and Crovella M. (1995). [Characteristics of WWW client-based traces](#). Boston, MA, Computer Science Dept., Boston University.
- [Cunha 1997] Cunha, C. (1997). [Trace analysis and its application to performance enhancements of distributed information systems](#). Doctoral thesis, Department of Computer Science, Boston University, Boston, Massachusetts.
- [Glassman 1994] Glassman, S. (1994). [A caching relay for the World Wide Web](#). Computer Networks and ISDN Systems 27(2).
- [Gribble and Brewer 1997] Gribble, S. D, and Brewer, E. A. (1997). [System design issues for Internet middleware services: Deductions from a large client trace](#). Submitted to the 1997 Usenix Symposium on Internet Technologies and Systems, Monterey, California December 1997.
- [Gwertzman and Seltzer 1996] Gwertzman, J., and Seltzer, M. (1996). [World Wide Web cache consistency](#). In Proceedings of the 1996 Usenix Technical Conference, Boston, MA, Harvard College.
- [HTTP-NG 1997] (1997) [World Wide Web Consortium Project Page](#).
- [Huberman et al 1997] Huberman, B, Pirolli, P., Pitkow, J., and Lukose, R. (1997) Strong regularities in WWW surfing. To appear in Science.
- [Leland et al 1993] Leland, W., Taquu, M., Willinget, W., and Wilson, D. (1993). On the self-similar nature of Ethernet traffic, In Proceedings of ACM SIGCOMM '93. San Francisco, California, October 1993.
- [Luotonen and Altis1994] Luotonen, A. and Altis, K. (1994). [World-Wide Web proxies](#). Computer Networks and ISDN Systems 27(2).
- [Manley 1997] Manley, S. An analysis of issues facing World Wide Web servers. Bachelor of Arts. Department of Computer Science, Harvard College, Cambridge, Massachusetts, April 1997.
- [Manley and Seltzer 1997] Manley, S., and Seltzer., M., (1997). [Web facts and fantasy](#). In Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems, Monterey, CA, December 1997.
- [Mauldin 1994] Mauldin, M., and Leavitt, J. (1994) [Web agent related research at the Center for Machine Translation](#). Meeting of the ACM Special Interest Group on Networked Information Discovery and Retrieval, McLean, VA, Aug. 1994.
- [Mogul 1995] Mogul, J. (1995). [Network behavior of a busy Web server and its clients](#). CA, Digital Western Research Laboratory, October 1995.
- [Mogul 1996] Mogul, J. (1996). [Digital's Web proxy traces](#). Online reference.

- [Nabeshima 1997] Nabeshima, M. (1997). [The Japan cache project: an experiment on domain cache](#). In Proceedings of the Sixth International WWW Conference, Santa Clara, CA.
- [O'Callaghan 1995] O'Callaghan, D. (1995). [A central caching proxy server for WWW users at the University of Melbourne](#). Proceedings of AusWeb95, the First Australian WWW Conference, University of Melbourne, Australia.
- [Pitkow and Recker 1994] Pitkow, J., and Recker, M. (1994). [A simple yet robust caching algorithm based on document access patterns](#). In Proceedings of the Second International WWW Conference. Chicago, Illinois, October 1994.
- [Pitkow and Pirolli 1997] Pitkow, J., and Pirolli, P. (1997). [Life, death and lawfulness on the electronic frontier](#). In Proceedings of Human Factors in Computing Systems (CHI 97). Atlanta, Georgia, April 1997.
- [Pitkow 1997] Pitkow, J. (1997). [In search of reliable usage data](#). In Proceedings of the Sixth International WWW Conference, Santa Clara, CA, April 1997.
- [Scheuermann et al 1997] Scheuermann, P., Shim, J., and Vingralek, R. (1997). [A case for delay-conscious caching of Web documents](#). In Proceedings of the Sixth International WWW Conference, Santa Clara, CA, April 1997.
- [Sedayao 1994] Sedayao, J. (1994). "Mosaic will kill me network!" [Studying network traffic patterns of Mosaic use](#). In Proceedings of the Second International WWW Conference. Chicago, Illinois, October 1994.
- [Smith 1994] Smith, N. (1994). [What can archives offer the World-Wide Web](#). In Proceedings of the First International WWW Conference, Geneva, Switzerland, May 1994.
- [Tauscher 1996] Tauscher, L. (1996). [Evaluating history mechanisms: an empirical study of reuse patterns in World Wide Web navigation](#). Master's thesis, Department of Computer Science. University of Calgary, Alberta, Canada.
- [Tauscher and Greenberg 1996] Tauscher, L. and Greenberg, S. (1996). [How people revisit Web pages: Empirical findings and implications for the design of history systems](#). International Journal of Human Computer Studies 47(1).
- [WCG 1997] Web Characterization Group: Boston University Ocean's Group, Harvard College's Vino Group, Microsoft, Netscape, Virginia Tech's Network Resource Group, and Xerox PARC. (1997). Work in progress.
- [Williams et al 1996] Williams, S., Abrams, M., Standridge, C., Abdulla, G., and Fox, E. (1996). [Removal policies in network caches for World-Wide Web documents](#). In Proceedings of ACM SIGCOMM 96, Stanford, CA, Aug. 1996.
- [Woodruff et al 1996] Woodruff, A., Aoki, P., Brewer, E., Gauthier, P., and Rowe, L. (1996). [An investigation of documents from the World Wide Web](#). The World Wide Web Journal 1(3).
- [Wooster 1996] Wooster, R. (1996). [Optimizing response time, rather than hit rates, of WWW proxy caches](#). Master's thesis, Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.
- [Wooster and Abrams 1997] Wooster, R., and Abrams, M. (1997). [Proxy caching that estimates page load delays](#). In Proceedings of the Sixth International WWW Conference, Santa Clara, CA.
- [Worrell 1994] Worrell, K. (1994) [Invalidation in large scale network object caches](#). Master's thesis, Computer Science, University of Colorado, Boulder, Colorado.