The Content and Access Dynamics of a Busy Web Server*

Venkata N. Padmanabhan†
Microsoft Research
padmanab@microsoft.com

Lili Qiu‡
Cornell University
lqiu@cs.cornell.edu

ABSTRACT

We study the MSNBC Web site, one of the busiest in the Internet today. We analyze the dynamics of content creation and modification as well as client accesses. Our key findings are (a) files tend to change little upon modification, (b) a small set of files get modified repeatedly, (c) file popularity follows a Zipf-like distribution with an α much larger than reported in previous, proxy-based studies, and (d) there is significant temporal stability in file popularity but not much stability in the domains from which popular content is accessed. We discuss implications of these findings.

1. INTRODUCTION

An accurate understanding of the World Wide Web workload is essential for the development of sound algorithms to improve the functioning of the Web. To this end, we have analyzed the dynamics of the MSNBC Web site[1], a large and busy news site. Our study is distinguished in two ways: (1) unlike the vast majority of Web studies that focus on proxy workloads, we study the workload of an origin server, and (2) our definition of the workload encompasses both client access (i.e., *frontend* activity), and content creation and modification (i.e., *backend* activity). We discuss implications of our findings for the effectiveness of Web caching and prefetching, cache consistency algorithms, and optimizations such as delta-encoding. It is difficult to know how well the results derived from the MSNBC site generalize. However, we believe that the operation of this site is typical of large news sites such as CNN and ABCNews, hence our study is valuable despite its limitations.

2. EXPERIMENTAL SETTING

The MSNBC site comprises a cluster of over 40 server nodes, each running the Microsoft Internet Information Server (IIS). The load balancing algorithms spreads the client requests across the server nodes in such a manner that at different times a particular client's request may be served by any of the server nodes.

We used several logs in our analysis:

Server access log: For each client access, this records a timestamp, the client's IP address, the URL accessed, etc. Due to administrative reasons, accesses to images were not logged. Despite this, the site saw over 25 million accesses to its HTML content alone each day.

Content creation and modification logs: This log, derived from the Microsoft Content Replication System (CRS), indicates the time of creation and modifications of files.

Content Logs: For a small fraction of the content, we obtained a log of HTML content itself, i.e., successive versions of the files as they underwent modification.

Most of these logs were obtained during October 1999, with a few during other periods, notably during the U.S. Operation Desert Fox against Iraq in December 1998.

3. SERVER CONTENT DYNAMICS

The file creation and modification process at MSNBC exhibits marked diurnal and weekly cycles. During a one-week period in October 1999, there were approximately 6000 file creation events and 24000 modification events. The number of unique files modified was only about 10% of the latter implying that a subset of the files tends to get modified repeatedly.

For the subset of files that undergo modification, 90% of the inter-modification intervals lie between an hour and a day. CRS pushes new content to the servers once an hour with more frequent pushes happening only when there are "hot" news events. For a given file, the duration of a modification interval is a poor predictor of the duration of the next modification interval. The coefficient of correlation between the durations of successive intervals is only 0.23. On the other hand, using a greater amount of history results in a good predictor. The mean inter-modification duration computed over at least 10 intervals is well correlated with the duration of the next interval (coefficient of correlation of 0.79).

*A full-length version of this paper is available as [3]

†http://www.research.microsoft.com/~padmanab/

‡http://www.cs.cornell.edu/lqiu/

[1] http://www.msnbc.com/

The extent of change upon file modification tends to be very small. In over 70% of the cases, the change in file size is 1% or less. The change in the textual (HTML) content is also minimal as evident from the *cosine similarity metric* [2] between successive versions being close to 1.

4. CLIENT ACCESS DYNAMICS

We begin by examining the relationship between server content dynamics and client access dynamics. There is a strong correlation between the age of a file and its popularity. Most files receive a lot more accesses soon after their creation than afterwards. There are exceptions, e.g., files such as index pages that remain "hot" long after their creation.

Previous research [4] has shown that up to 40% of accesses are to files that have not been accessed before by clients in the same domain. In a caching context, these lead to *first-time* (i.e., compulsory) misses. We have found that most first-time accesses are to old files that were created at least a day ago. As such, these files are unpopular and accesses to them are hard to predict.

As reported in the literature [1], file popularity tends to follow a Zipf-like distribution, C/i^{α}. However, unlike previous, proxy-based studies which found α to be well under 1.0, our server-based analysis yield a much higher α, typically in the range 1.4–1.6. The high degree of concentration of client accesses — in some instances, 2% of the files account for 90% of the accesses — suggests that techniques such as reverse caching and replication would go a long way in alleviating server load.

Turning to the stability of Web page popularity, we find that popularity remains stable over time-scales of around a week. Of the 100 most popular files on a given day, over 60 tend to overlap with the popular set for others days within a week. Also, the greater the popularity of a document on a given day, the more likely it is to remain popular on the following days.

The stability of the interest group of a Web page, i.e., the set of clients/domains interested in the page, tends to be low. Well under a 100 files have over a 50% overlap in their interest groups on successive days. A possible explanation is that domain-level proxy caching reduces the likelihood of multiple requests for a file emanating from a domain.

Finally, we evaluate spatial locality in client behavior by comparing the degree of overlap in requests of clients belonging to the same domain to that of clients grouped together at random. We find that domain membership is generally significant. However, a "hot" event of global interest, such as Operation Desert Fox, can become dominant enough to diminish the significance of domain membership.

5. SUMMARY AND CONCLUSIONS

The main findings of our study are:

1. Most (HTML) file modifications tend to be minor in terms of the change both in the file size and in the visible textual content.
2. Past modification behavior of a file, if averaged over a sufficient number of samples, tends to be a reasonably good predictor of future modification behavior.
3. The Zipf-like distribution of file popularity has an α of 1.4–1.6, much larger than what has been reported in proxy-based studies. The large α implies, in some instances, that just the top 2% of documents could account for 90% of the accesses.
4. The popularity of files tends to be stable over a timescale of a week. However, the set of domains from which the popular documents are accessed tends to change significantly from day to day.
5. Organizational (i.e., domain) membership of clients tends to have a significant (positive) impact on the degree of local sharing, unless there is a globally popular event that cuts across organizational boundaries.
6. Document popularity tends to drop off with age.
7. The majority of first-time accesses from a domain are to unpopular documents at least a day old.

These findings have several implications for the design of new algorithms for the Web:

1. Past modification behavior of files is a good basis for designing cache consistency algorithms.
2. Delta-encoding would be useful given the minor change in file content upon modification.
3. The lack of stability in client interest groups would make it difficult for the server to push documents discriminatively to clients.
4. The high-degree of concentration of Web accesses among popular documents implies that techniques such as reverse caching and replication would be very effective in reducing server load.
5. First-time misses, which occur frequently, are hard to cut down because most such accesses are to old and unpopular files and hence difficult to anticipate.

6. ACKNOWLEDGMENTS

Special thanks to Jason Bender and Ian Marriott for making the trace data available to us and for their patience with our many questions. Thanks also to Damon Cole, Susan Dumais, Niccole Golden, Chris Haslam, and Eric Horvitz.

7. REFERENCES

[1] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. "Web Caching and Zipf-like Distributions: Evidence and Implications". In *Proc. of INFOCOM'99*, March 1999.

[2] D. A. Grossman, O. Frieder. "Information Retrieval — Algorithms and Heuristics". *Kluwer International Series in Engineering and Computer Science*, September 1998.

[3] V. N. Padmanabhan and L. Qiu. "The Content and Access Dynamics of a Busy Web Server: Findings and Implications". *Microsoft Research Technical Report MSR-TR-2000-13*, February 2000.

[4] A. Vahdat, M. Dahlin, T. Anderson, and A. Aggarwal, Active Names: Flexible Location and Transport of Wide-Area Resources. In *Proc. USITS'99*, October 1999.