Towards Understanding Modern Web Traffic

Sunghwan Ihm
Department of Computer Science
Princeton University
sihm@cs.princeton.edu

Vivek S. Pai
Department of Computer Science
Princeton University
vivek@cs.princeton.edu

ABSTRACT

As the nature of Web traffic evolves over time, we must update our understanding of underlying nature of today's Web, which is necessary to improve response time, understand caching effectiveness, and to design intermediary systems, such as firewalls, security analyzers, and reporting or management systems. In this paper, we analyze five years (2006-2010) of real Web traffic from a globally-distributed proxy system, which captures the browsing behavior of over 70,000 daily users from 187 countries. Using this data set, we examine major changes in Web traffic characteristics during this period, and also investigate the redundancy of this traffic, using both traditional object-level caching as well as content-based approaches.

Categories and Subject Descriptors

C.2.m [**Computer-Communication Networks**]: Miscellaneous

General Terms

Measurement, Design, Performance

Keywords

Web Traffic Analysis, Web Caching

1. INTRODUCTION

The World Wide Web is one of the most important Internet applications, and its traffic volume is increasing and evolving due to the popularity of social networking, file hosting, and video streaming sites [3]. Understanding these changes is important to overall system design. For example, analyzing end-user browsing behavior leads to a Web traffic model, which in turn can be used to generate a synthetic workload for benchmarking or simulation. In addition, analyzing redundancy in the Web traffic and the effectiveness of caching could shape the design of Web servers, proxies, and browsers to improve response times.

While there has been much research in the past decade to better understand the nature of Web traffic, unfortunately, we still have little understanding of today's Web. It is challenging because understanding changes requires large-scale

data spanning a multi-year period. Also, while content-based caching [2] is known to be very effective and becomes popular, understanding its effectiveness on Web traffic requires full content data rather than just access logs.

In this paper, we analyze five years (2006-2010) of real Web traffic from the CoDeeN content distribution network [6], a globally distributed proxy system which captures the browsing behavior of over 70,000 users per day from 187 countries. Using this data, we examine major changes in Web traffic characteristics over a five-year period, such as content type distributions and popular sites. In addition, we capture the full content of traffic, and study the redundancy and impact of caching, using both traditional object-based caching as well as content-based caching approaches.

2. DATA SET

CoDeeN is a semi-open [1] globally distributed proxy which has been running since 2003, and serves over 30 million requests per day from more than 500 PlanetLab [5] nodes. For this study, we consider a five-year period of access log data from 2006 to 2010, as well as full content data from 2010. Due to the large volume of requests, we sample one month (April) of data per year, and focus on the traffic of users from four countries from different continents – the United States (US) in North America, Brazil (BR) in South America, China (CN) in Asia, and France (FR) in Europe. Overall, our analysis on four countries covers 48-137 million requests, 689-1903 GB traffic, and 70-152 thousand users per month.

3. PRELIMINARY RESULTS

Content Types Figure 1 presents the content type distribution changes in the United States, France, and Brazil from 2006 to 2010, connected by arrows. The X axis is the percentage of requests, and the Y axis is the percentage of bytes, both in log-scale. We omit China's result that also exhibits similar changes.

First, we observe a sharp increase of javascript, css, and xml, primarily due to the popular use of ajax [1]. We also find a sharp increase of flash video (flv) traffic, taking about 25% of total traffic in the United States and Brazil in 2010. At the same time, non-flv video traffic sees a decrease, demonstrating the shift of the media delivery medium to the popular flash video. Still, the image traffic including all of its subtypes consumes the most bandwidth.

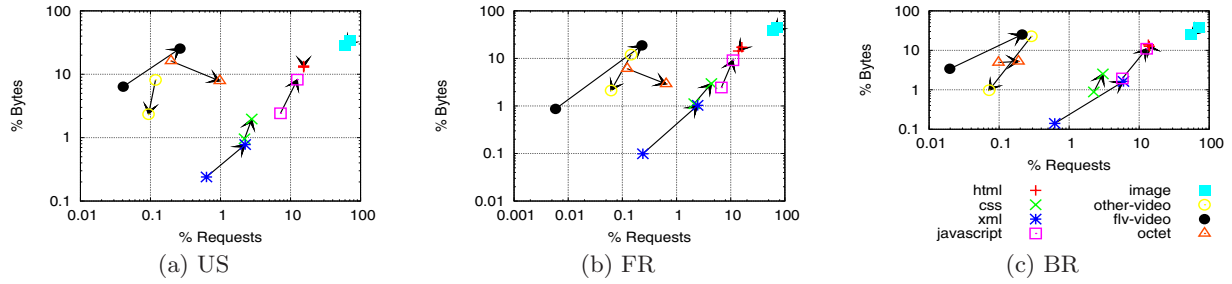
[1]It only allows GET requests for security reasons.

(a) US   (b) FR   (c) BR

Figure 1: Content type distribution changes from 2006 to 2010: Flash-video/ajax traffic is increasing.

(a) Ads network traffic by requests   (b) Video site traffic by bytes   (c) Single top site % IPs
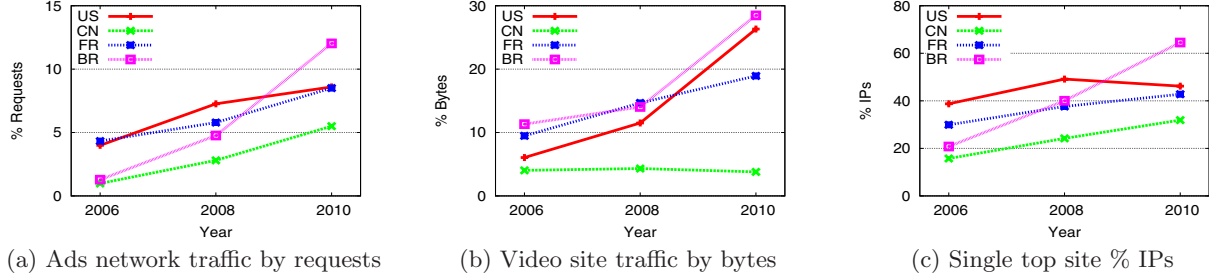
Figure 2: Top sites: Ads/video site traffic is increasing. A single top site tracks up to 65% of the entire users.
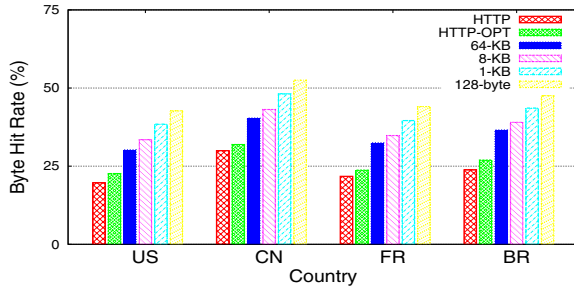

Figure 3: Ideal byte hit rate with infinite cache storage: Content-based caching with 128-byte chunks achieves almost 2x larger byte hit-rate than HTTP caching.

Top Sites We examine the share of 1) video site traffic (*e.g.*, `youtube.com`), and 2) advertising network/analytics traffic (*e.g.*, `doubleclick.com`, `google-analytics.com`) in Figure 2. We consider only the top 50 sites as we gain diminishing returns from further investigation, so the result conservatively estimates the actual share.

First, in Figure 2 (a), we observe that advertising network traffic takes 1-12% of the total requests, and it consistently increases over time as the market grows [4]. In addition, we find the volume of video site traffic is consistently increasing as shown in Figure 2 (b), taking up to 28% in Brazil in 2010. In China's case, however, the image-hosting site traffic takes the largest volume, and the share of video site traffic is lower than the share of other countries. Finally, we see the single top site reaches a growing fraction of all users over time in Figure 2 (c). All of the single top sites during a five-year period are either a search engine (`google.com` or `baidu.com`), or analytics (`google-analytics.com`). Especially in 2010, the percentage reaches up to 65% in Brazil, which might concern user privacy.

Redundancy and Caching We calculate the ideal bandwidth savings achievable with infinite cache storage using the traditional object-level HTTP caching and content-based caching. For the object-level caching, we assume two objects

are identical (cache hit) if they are cacheable and their URLs and content lengths match. We also consider a slightly optimistic behavior of object-based caching by discarding query strings from URLs in order to accommodate the case where two URLs with different metadata actually belong to the same object. For content-based caching, we vary the average chunk size from 128 bytes, 1 KB, 8 KB, to 64 KB, and exclude the metadata overhead.

In Figure 3, we observe that content-based caching outperforms the object-based caching with any chunk size. The cache hit rate of object-level caching ranges from 27-39% (not shown in the figure), but the actual byte hit rate is only 20-30%. The hit rate of the optimistic version (`HTTP-OPT`) is only slightly larger. On the other hand, the lowest byte hit rate of the content-based caching is 30-40% with 64-KB chunks, and the highest byte hit rate is 43-53% with 128-byte chunks, 1.7-2.2x larger than the object-level caching's.

4. ACKNOWLEDGMENT

References

[1] D. Crane, E. Pascarello, and D. James. *Ajax in Action*. Manning Publications Co., Greenwich, CT, USA, 2005.

[2] S. Ihm, K. Park, and V. S. Pai. Wide-area Network Acceleration for the Developing World. In *Proc. USENIX Annual Technical Conference*, Boston, MA, June 2010.

[3] ipoque. Internet Study 2008/2009. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009.

[4] JPMorgan Chase & Company. The Rise of Ad Networks. http://www.mediamath.com/docs/JPMorgan.pdf.

[5] PlanetLab. http://www.planet-lab.org/, 2008.

[6] L. Wang, K. Park, R. Pang, V. S. Pai, and L. Peterson. Reliability and security in the CoDeeN content distribution network. In *Proc. USENIX Annual Technical Conference*, Boston, MA, June 2004.