Temporal Locality in Web Request Streams

*Sources, Characteristics, and Caching Implications**

An Extended Abstract

Shudong Jin and Azer Bestavros
Computer Science Department, Boston University
111 Cummington St, Boston, MA 02215
{jins,bestavros}@cs.bu.edu

1. INTRODUCTION

Web access patterns exhibit a number of unique properties that have been identified and characterized. The prevalence of some of these properties has motivated the development of many protocols (and optimizations thereof) that exploit such properties. One such property is the temporal locality of reference exhibited in Web request streams. Temporal locality in Web request streams emerges from two distinct phenomena, the *long-term popularity* [1, 2, 3] of Web documents and the *short-term temporal correlations* of references. Delineating between these two sources is important because they have different implications for caching and replication protocols. The highly skewed popularity of Web documents suggests the use of long-term frequency in caching and replication algorithms, while the temporal correlations of references suggests the use of short-term residency information. In this paper we show that temporal locality metrics proposed in the literature are unable to delineate between these two sources. In particular, the commonly-used inter-request time distribution [2, 3, 4] is predominantly determined by the power law governing the long-term popularity of documents in the request stream. Such an inherent relationship tends to disguise the existence of short-term temporal correlations. We propose a new method that enables accurate delineation between these two sources. We use that method to characterize temporal locality in a number of representative proxy cache traces. Our study shows that there are measurable differences between the degrees of temporal locality across these traces. We have demonstrated the significance of our findings by proposing and evaluating a novel Web cache replacement policy, which exploits both long-term popularity and short-term temporal correlations in an adaptive fashion. Trace-driven simulations show the superior performance of our algorithm.

*This work was partially supported by a NSF research grant ESS CCR-9706685. A full version of this paper is available as technical report BU-CS-99-014 from http://www.cs.bu.edu/techreports/

2. CHARACTERIZATION

We analyzed a one-week trace from DEC and three two-week traces from the NLANR sites. As suggested in [2, 3, 4], we used the distribution of document inter-request times to characterize the temporal locality of reference in these traces. The log-log scale plots for this distribution were nearly straight lines with slopes varying across the traces. The inter-request time distribution is reflective of the two sources of temporal locality: namely document popularity and temporal correlations of document requests. An important question is whether such a distribution can be used to assess accurately the strength of both of these sources.

To isolate the effects of skewed popularity, we applied a random permutation to the original traces. This permutation eliminates the effect of temporal correlations. Figure 1(a) shows the inter-request time distribution for both the original and scrambled traces (more results are available in our full paper [5]). The closeness of slopes suggests that the distribution of document inter-request times is predominantly determined by the popularity distribution.

The skewed popularity distribution of documents contributes to the heavy-tailed nature of inter-request time distributions. Highly popular documents are requested frequently, and thus tend to have shorter inter-request times; less popular documents, are requested infrequently, and thus tend to exhibit longer inter-request times. The relationship between popularity and inter-request distributions is established by the following (see [5] for a proof).

THEOREM 1. *If the distribution of document popularity in a request stream asymptotically follows a power law with parameter α, where $0.5 \ll \alpha \leq 1$, then the distribution of inter-request time in a random permutation of this request stream can be characterized asymptotically using a power law with parameter $(2 - 1/\alpha)$.*

To isolate and measure the effect of temporal correlations on temporal locality, we propose a new metric—namely, *the probability distribution of inter-request times for equally popular documents.* Since we characterize temporal locality for *equally* popular documents, the impact of the highly-skewed popularity of documents in our traces is masked. We illustrate the sensitivity of our metric to temporal correlations by plotting the distribution of inter-request times for equally popular documents for the original traces and for the scram-

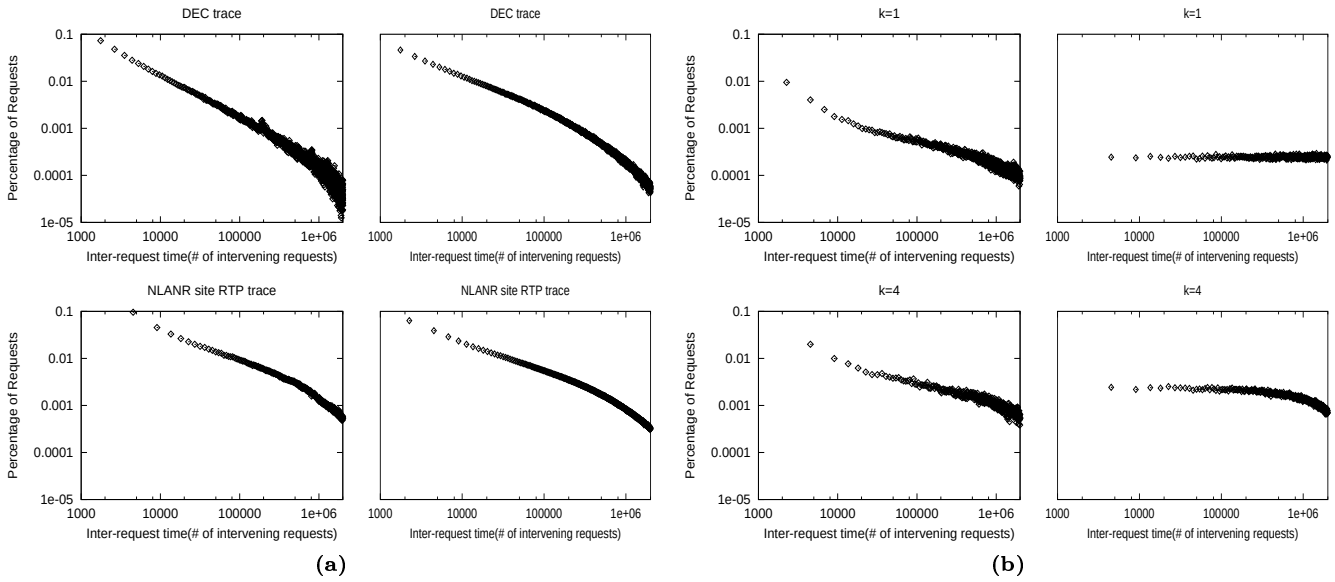

Figure 1: Inter-request time distributions: In both (a) and (b) the left plots are for the original traces, the right plots are for the scrambled traces. In (a), the closeness of the inter-request time distribution for the original and for the scrambled traces suggests that inter-request time distribution is predominantly determined by popularity distribution, and thus it is not able to identify temporal correlations. In (b), we show the inter-request time distribution of equally popular documents in the NLANR site RTP trace. The plots are drawn for the inter-request time after the k^{th} requests. The exhibited slope in the distribution for the original traces (a slope that disappears as a result of scrambling) suggests the existence of temporal correlations and provides a measure of its strength.

bled traces. Figure 1(b) shows that the slope of log-log scale plots—which is evident for the original traces—almost disappears for the scrambled traces.

The slope that characterizes the inter-request times for equally-popular documents is the result of only temporal correlations. Thus this slope can be used to gauge the level of such correlations. Specifically, the degree of short-term temporal correlations can be captured by a parameter β, the slope of log-log scale inter-request time distribution for equally popular documents. Using this method, we also observe that there is evidence of a weakening in temporal correlations from the DEC traces to the NLANR traces.

To summarize, (1) temporal locality is induced through two sources: long-term popularity of documents and short-term temporal correlations of reference; (2) there is an inherent relationship between document popularity and temporal locality, which tends to disguise the existence of temporal correlations; and (3) a novel model can characterize popularity and temporal correlations separately. Specifically, temporal locality can be characterized using a pair of parameters (α, β), where α is the parameter of the power law characterizing the popularity of documents in the request streams and β is the parameter of the power law approximately characterizing the inter-request time distribution of equally-popular documents.

3. CACHING IMPLICATIONS

Our characterization of temporal locality indicated that long-term popularity is the predominant contributor to temporal locality, but that temporal correlations exists. It advocates the use of a frequency-based, recency-aware cache replacement policy. In [5], we derived an algorithm—called GreedyDual*—that captures both long-term popularity and short-term temporal correlations in an adaptive fashion. We compared the performance (both hit ratio and byte hit ratio) of this algorithm with other algorithms such as LRU, GDS and LFU-DA. Our findings include: (1) Our proposed algorithm consistently achieves the best performance. The improvement is usually 10-30%. (2) The incorporation of long-term frequency is important for Web proxy caching algorithm, in line with our finding that document popularity is the main contributor to temporal locality. (3) Capturing temporal correlations is more crucial for smaller cache, in line with our finding that temporal correlations exists in short term. (4) Sensitivity analysis indicated that incorporation of β successfully captures the relative strength of long-term popularity and short-term temporal correlations.

REFERENCES

[1] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing reference locality in the WWW. In *Proceedings of PDIS'96*, December 1996.

[2] M. Arlitt and C. Williamson. Web server workload characteristics: The search for invariants. *SIGMETRICS'96*.

[3] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. *INFOCOM'99*, April 1999.

[4] P. Cao and S. Irani. Cost-aware WWW proxy caching algorithms. In *Proceedings of USITS'97*, December 1997.

[5] S. Jin and A. Bestavros. Temporal locality in web request streams: Sources, characteristics, and caching implication. Tech. Report BU-CS-99-014, CS Dept, Boston Univ, October 1999. (Available from `http://www.cs.bu.edu/techreports/`)