On the Fractal Nature of WWW and Its Application to Cache Modeling

Virgílio Almeida*
Adriana de Oliveira

Computer Science Department
Boston University
111 Cummington St,
Boston, MA 02215
virgilio@cs.bu.edu

Depto. de Ciencia da Computacão
Universidade Federal de Minas Gerais
Belo Horizonte, MG 30161
Brazil
dri@dcc.ufmg.br

Abstract

The World Wide Web (WWW or Web) is growing rapidly on the Internet. Web users want fast response time and easy access to a enormous variety of information across the world. Thus, performance is becoming a main issue in the Web. Fractals have been used to study fluctuating phenomena in many different disciplines, from the distribution of galaxies in astronomy to complex physiological control systems. The Web is also a complex, irregular, and random system. In this paper, we look at the document reference pattern at Internet Web servers and use fractal-based models to understand aspects (e.g. caching schemes) that affect the Web performance.

1 Introduction

The World Wide Web (WWW or Web) is growing rapidly on the Internet. Web users want fast response time and easy access to a enormous variety of information across the world. The Hypertext Transfer Protocol (HTTP) is the primary protocol used for sharing information [BLCL94]. A Web server is a system on a network that can process HTTP requests. Thus, Web server performance is becoming increasingly important [ASAW95, BrCl94, KwMR95]. To analyze the performance of WWW servers, one must construct a model of HTTP requests that arrive at a server. Many recent studies have used simulation models to analyze cache behavior for servers and proxies [PiRe94, ASAW95]. Simulation models of WWW requests have largely relied on traces of actual requests to servers. Such traces are indoubtedly more accurate than mathematical models, but they also have drawbacks. They are representative of a particular site and it is difficult to extrapolate their

*On sabbatical leave from Depto. de Ciencia da Computacão da UFMG, Brazil. Partially supported by CNPq-Brazil.

behavior to other servers. This paper proposes a mathematical model to represent the requests that arrive at a server. It is adequate for the purpose of understanding the WWW server behavior. The usefulness of a model relies upon its ability to answer questions of interest from the real phenomenon.

Fractals have been used to study fluctuating phenomena in many different disciplines, from the distribution of galaxies in astronomy too complex physiological control systems. The Web is also a complex, irregular, and random system. It is estimated that the Web now links over 40,000 local networks, 400,000 computers, and over 40,000,000 individuals. The Web contains over 1,000,000 documents. These numbers and the complex network of interactions among individuals and documents leaded us to conjecture that fractals can be used to enhance the understanding of the WWW behavior.

This paper discusses a method of modeling WWW requests. There are many aspects of WWW requests that one might wish to model, such as interarrival times, file sizes, type of files requested, etc. The discussion here will be limited to the document reference behavior of HTTP requests. The purpose of this paper is to use fractal-based models to understand various aspects that affect the Web performance. The remainder of this paper is organized as follows. In the next section, we present a brief introduction to fractal geometry and review the work on fractals in the context computer systems. Section 3 introduces the stack model used to analyze the behavior of an HTTP stream arriving at a server. Section 4 proposes a fractal-based analytical model for representing the behavior of a document cache in a WWW server. In section 5, we use logs from the WWW servers at Boston University and NCSA to validate the proposed model. A summary of our observations and comparisons with related work is provided in the conclusions.

2 The Fractal Nature of Computing and Networking

As defined by Mandelbrot [Mand83], fractal geometry is a mathematical theory conceived and developed to provide a model for irregularity and fragmentation in nature. Fractals have been used to describe spiky, irregular or variegated objects, such as coastlines, mountains, and crystals. Fractals are said to be self-similar structures. Each time a self-similar structure is magnified, another similar layer of finer detail is revealed. The statistical distributions of interest in fractal geometry are hyperbolic distributions of the form $P(U > u) = (u/u_0)^{-\theta}$, where U is a random variable, u_0 is a constant and θ is the fractal dimension. One key property of this distribution is the following: if it is plotted in a log-log format, the distribution appears as a straight line of slope θ. The hyperbolic distributions are closely associated with an important concept called statistical self-similarity, which involves sets with invariant probability distributions under scaling. Objects are said to be self-similar in a statistical sense, when parts of the whole fit the whole in distributions, rather than being exact copies. Hyperbolic distributions are shown to satisfy the requirements of self-similarity [CrBe95, VMHK83]. Fractals and self-similarity are intimately related.

Traditional performance models based on Markov characteristics have been extensively used to analyze computer systems [Klei76, MeAD94]. Poisson distributions have been successfully used to construct performance models of computer systems [MeAD94]. The exponential distribution has been used to represent service time and interarrival time. Its memoryless property simplifies the mathematical analysis even for complex system topolo-

gies. Most of the performance modeling studies of computers and networks have relied on the exponential assumption, because of its memoryless property [Klei76]. However, measurement data has been suggesting with increasing emphasis that new models are needed for describing distributed computing systems with more accuracy, as pointed out by Leland et al. [LTWW94].

There is little work exploring fractals and self-similarity in the context of computer systems and networks. A pioneering paper is that by Voldman et al. [VMHK83], where they noted that the interaction of software with memory hierarchies is naturally modeled by a stochastic process having a larger than usual amount of irregularity. The authors looked at memory traces of different software environments and found out that the distribution of the intermiss gaps is fractal. Thiebault [Thie89, ThWS92] noticed in several traces that the number of unique memory locations accessed by a program, as a function of the total number of memory accesses, converges toward a hyperbola, showing the fractal nature of programs. Thiebault also proposed and validated a simple fractal model of fully associative caches. In fact, the paper by Thiebault introduced fractal geometry to the field of computer performance evaluation and prediction. Recently, Peterson and Grossman [PeGr95] showed that the frequency distributions of measured parameters of disk I/O activity were observed to obey power laws indicating self-similarity over orders of magnitude of time. Based on a study on actual Ethernet traffic data, over several years, Leland et al. [LTWW94] demonstrated that Ethernet LAN traffic is statistically self-similar. They also show that none of the used traffic models is able to capture the fractal-like behavior of Ethernet LANs. The authors point out that analytical results show a clear distinction between predicted performance of certain queueing models with traditional Poisson streams and the same queuing models with self-similar inputs. Recent work by Crovella and Azer [CrBe95] shows evidences that the World Wide Web traffic may be self-similar. The authors also explain that the self-similarity in a wide area network traffic stems from factors such as the underlying distribution of of WWW document sizes and user think times. In this paper, we look at a different aspect of the WWW behavior. We focus our analysis on the document reference pattern at Internet Web servers.

3 Document Reference Model

To analyze the behavior of a WWW server, one must construct a model of HTTP requests arriving at the server. There are many aspects of an HTTP stream that one might wish to model, such as the distribution of HTTP interarrival times or the distribution of file sizes requested from the server. This paper analyzes the reference behavior of the documents requested from a WWW server. In order to develop good caching strategies, one need to know the document access patterns. There are various types of documents in a Web server, such as HTML, gif, postscript, and MPLG. As pointed out in [KwMR95], effective document caching is one of the key principles in a server architecture, allowing a WWW server to respond quickly to requests for frequently accessed documents. Our model is influenced by the Spirn's work on paging [Spir76]. The properties of program behavior in a paging environment, using a Least Recently Used (LRU) replacement policy, have been covered by Spirn. Those properties and a hyperbolic relation defined by Spirn are central to the document reference model proposed in this paper.

Let us consider a Web server with N different documents. Let r_t denote an HTTP request at time t. During an observation interval, a stream ($\rho = r_1, r_2, ..., r_t$) of HTTP requests arrive at the server. Let D_t be the number of the document referenced by an HTTP request at time t. Let us define an LRU stack s_t, which is an ordering of all N documents. Let $s_t = [D_1, D_2, ..., D_N]$, where $D_1, ..., D_N$ are the documents of the server. Thus, D_1 is the most recently referenced document, D_2 is the next most recently referenced document, and D_n is the least referenced document. Initially, all documents are simply grouped together in an arbitrary order at the right hand side of the stack. Whenever a reference is made to a document, the stack must be updated. Considering that $r_{t+1} = D_i$, then the stack becomes $s_{t+1} = [D_i, D_1, D_2, ..., D_{i-1}, D_{i+1}, ..., D_n]$. Suppose now that $s_t = [D_1, D_2, ..., D_N]$ and $r_{t+1} = D_i$. Then, we could say that request r_{t+1} is at distance i in stack s_t. Let d_t denote the stack distance of the document referenced at time t. We then have the following relation:

$$\text{if } r_{t+1} = D_i \text{ then } d_{t+1} = i \tag{1}$$

Thus, to any HTTP request string $\rho = r_1, r_2, ..., r_t$ corresponds a distance string $\delta = d_1, d_2, ..., d_t$. We can then consider that the distance string and the request string are equivalent in terms of reference information. Since the distance string reflects the pattern in which WWW users request documents from a server rather than the actual identity of the documents, it will be used as our model of document reference pattern.

4 Fractal Evidences in the Document Reference Pattern

The data sets used in our analysis were the logs of accesses to WWW servers at the NCSA (National Center for Supercomputing Applications) and BU (Boston University) . The NCSA HTTP daemons produce a log entry for each request. To understand the access patterns and characteristics of the WWW servers, we analyzed the trace logs of BU and NCSA for selected weeks, from October to December of 1995. Table 1 summarizes the main characteristics of the log data we have used in this study. We analyzed these logs searching

Item	NCSA 1	NCSA 2	Boston
Access Log Duration	1 day	2 days	2 weeks
Access Log Size (MB)	5	8	9
Total Requests	43,384	79,440	80,518
Unique Requests	4,018	4,851	4,471
Total Bytes Transferred (MB)	556	1,158	887
Mean Transfer Size (bytes)	13,438	15,285	11,551

Table 1: Summary of Access Log Data

for fractal evidences, that would allow us to develop fractal-based models, that capture the generic mechanisms of the WWW and do not depend on details of specific servers.

4.1 Pictorial View of Self Similarity

Figure 1 exhibits a sequence of plots of the aggregate distance of referenced documents (i.e., the summation of the distances of the documents referenced during a time unit) for three different values of time units. Plot A was obtained with a time unit of 1000 seconds. Graph B was derived from graph A, by increasing the time resolution by a factor of 10 and by randomly choosing a subinterval of the data presented in the previous graph. The same procedure was used to produce plot C from B. The three plots suggest evidences of self-similarity in the document reference pattern. The graphs of Figure 1 seem to look the same for the three time units. All the three plots contain distance bursts, observable over three orders of magnitude. This scale-invariant feature of the server accesses shows signs of the fractal nature of WWW.

4.2 The R/S Analysis

The objective of the R/S (rescaled range) analysis of the document reference data set is to infer the degree of self-similarity H (Hurst parameter) for the pattern of accessing documents in a WWW server. This method is based on the fact that for a self-similar dataset , the R/S index grows according to a power law, with exponent H as a function of the number of points considered. Using regression, we calculated the slope of R/S plot, which yielded an estimate for the Hurst parameter of 0.78. According to Mandelbrot [Mand83], self-similar processes exhibit a Hurst parameter value in the range $(0.5, 1.0)$. For the purpose of comparison of the estimated H parameter with the asymptotic slopes, Figure 2 shows the R/S plot with two lines of slopes 0.5 and 1.0. The estimated value for H suggests some degree of self-similarity in the document reference pattern.

5 Cache Model

Our initial conjecture was that accesses to documents in a WWW server behaved as a random walk on a lattice. It is like the movement of a walker as it jumps from cell to cell [Thie89]. The walker is not constrained to only jump to a neighboring cell, but can indeed jump to a cell at any distance away. Requests would move from one document to another in a WWW server, resembling the walker movement. In the previous section, we showed evidences of the fractal nature of the WWW accesses. Let us now consider those evidences and assume that a fractal random distribution can be used to characterize the document distance distribution. Let us define the document distances as samples from the random variable U, whose distribution is as follows:

$$P(U > u) = (u/u_0)^{-\theta} \tag{2}$$

where u_o is a constant and θ is the fractal dimension. A random walk is considered transient when $\theta < 1$ and recurrent when $\theta > 1$. A transient walk jumps constantly to unvisited portions of the lattice. In our context, a transient behavior would represent those requests that jump to unvisited documents of the server. A recurrent walk tends to stay in a neighborhood for a long time. A recurrent HTTP stream would tend to visit the same documents several times before jumping to a new neighborhood. Gillis and Weiss [GiWe70, Thie89] show that the number of lattice-cells visited as the number of total cells visited grows asymptotically

toward a power function, given by:

$$\text{Number of unique cells} = u(n) = An^{1/(\theta-1)} \tag{3}$$

where A is a constant and θ is the locality parameter. Using equation (3) and considering a cache of size C, Thiebault [Thie89] shows that the probability of hit index of a selected document is greater than C is given by:

$$P[hit\ index > C] = \frac{A^{\theta}}{\theta} C^{1-\theta} \tag{4}$$

where hit index of a document is defined by document position in the LRU document stack. In fact, the probability defined in equation (4) is the probability of a miss in the cache of size C. Let us consider a cache capable of holding C documents. Using the log data from the servers at NCSA and BU, we simulated the HTTP request stream and plotted the graph (log-log) of the number of unique documents as a function of the total number of requests. Figure 3 shows three different curves. The real data curve was plotted with data provided by a trace-driven simulation, based on the logs of NCSA and BU. The asymptote is described by equation i(3). The 45-degree line coincides with real data line while the number of unique documents is less than or equal to the number of requests. From the curves of Figure 3, we extracted the coefficients A and θ, that define the asymptotic model for the cache behavior.

We also used trace-driven simulation to validate the cache model defined by equations (3) and (4). The comparison of the predicted and simulated miss rate (i.e., the probability of experiencing a miss when accessing a document) for different values of cache sizes is shown by the graphs of Figure 4. The quality of the prediction is reasonably good, especially when one considers that only two parameters (A and θ) were used to describe the HTTP request pattern in a cached environment.

6 Conluding Remarks

The use of fractal geometry to understand the behavior of complex systems such as the Web has several advantages. For instance, simple analytical models can be used to investigate new caching schemes (e.g., separate caches for text, graphics, and videos or geographical caches) that would improve performance of the Web. The work by Crovella and Azer [CrBe95, CuBC95] found evidences of self-similarity in the Web traffic. In this paper, we took a different approach. We defined a distance-based model for document references and found evidences of the fractal nature in the document reference pattern, described by logs of Web servers at NCSA and BU. We also proposed and validated a fractal-based analytical model for a cache of documents.

Acknowledgements

The authors thank Carlos Cunha of Boston University and Robert McGrath of the NCSA (National Center for Supercomputing Applications), for providing access to the trace logs of the Computer Science Department and NCSA WWW servers, respectively.

References

[ASAW95] Abrams, M., Standridge, C. R., Abdulla, G., Williams, S. and Fox, E. A., "Caching Proxies: Limitations and Potentials", Technical Report TR-95-12, Department of Computer Science, Virginia Polytechnic Institute and State University, July 1995.

[BLCL94] Berners-Lee, T., Cailliau, R., Luotoneu, A., Nielsen, H. F. and Secret, A., "The World Wide Web", *Communications of the ACM*, pp 76-82, Vol. 37, No. 8, August 1994.

[BCCC95] Bestavros, A., Carter, R., Crovella, M., Cunha, Carlos, Heddaya, A. and Mirdad, S., "Application level document caching in the internet", *IEEE SDNÉ96: The Second International Workshop on Services in Distributed and Networked Environments*, Whistler, British Columbia, June 1995.

[BrCl94] Braun, H.-W. and Claffy, K., "Web Traffic Characterization: an Assessment of the Impact of Caching Documents from NCSA's Web Server", *Proccedings of the Second International WWW Conference*, Chicago, Illinois, October 1994, pp. 1007-1027.

[CrBe95] Crovella, M. E. and Bestavros, A., "Explaining World Wide Web Traffic Self-Similarity", Technical Report TR-95-015 , Computer Science Department, Boston University, Boston, October 1995.

[CuBC95] Cunha, C., Bestavros, A. and Crovella, M., "Characteristics of WWW Client-based Traces", Technical Report TR-95-010 Computer Science Department, Boston University, Boston, April 1995.

[GiWe70] Gillis, J. E. and Weiss, G. H., "Expected Number of Distinct Sites Visited by a Random Walk With an Infinite Variance", *J. Math. Phys.*, Vol. 11, No. 4, pp. 1307-1312, April 1970.

[Klei76] Kleinrock, L, *Queuing Systems: Volume I*, John Wiley and Sons, 1976.

[KwMR95] Kwan, T. T., McGrath, R. E. and Reed, D. A., "NCSA's World Wide Web Server: Design and Performance". *IEEE Computer*, November 1995.

[MeAD94] Menasce D., Almeida V., and Dowdy L., *Capacity Planning and Performance Modeling*, Prentice Hall, New Jersey, 1994.

[LTWW94] Leland, W. E., Taqqu, M. S., Willinger, W. and Wilson, D. V. , "On the Self-Similar Nature of Ethernet Traffic (Extended Version)", *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, February 1994.

[Mand83] Mandelbrot, B. B., *The Fractal Geometry of Nature*, W. H. Freedman and Co., New York, 1983.

[PiRe94] Pitkow, J. G., and Recker, M. M., "A Simple Yet Robust Caching Algorithm Based on Dynamic Access Patterns", *Proccedings of the Second International WWW Conference*, Chicago, Illinois, October 1994, pp 1039-1046.

[PeGr95] Peterson D. and Grossman R., "Power laws in large shop DASD I/O activity", in Proceedings of CMG95 International Conference, Computer Measurement Group Inc., Nashville, December 1995.

[Spir76] Spirn, J., "Distance String Models for Program Behavior", *IEEE Computer*, November 1976.

[Thie89] Thiébaut, D., "On the Fractal Dimension of Computer Programs and its Application to the Prediction of the Cache Miss Ratio", *IEEE Transactions on Computers*, Vol. 38, July 1989.

[ThWS92] Thiébaut, D., Wolf, J. L. and Stone, H. S., "Synthetic Traces for Trace-Driven Simulation of Cache Memories ", *IEEE Transactions on Computers*, Vol. 41, No. 4, pp. 388-410, April 1992.

[VMHK83] Voldman, J., Mandelbrot, B., Hoevel, L. W., Knight, J. and Rosenfeld, P. L., "Fractal Nature of Software-Cache Interaction", *IBM Journal of Research and Development*, Vol. 27, No. 2, 1983.

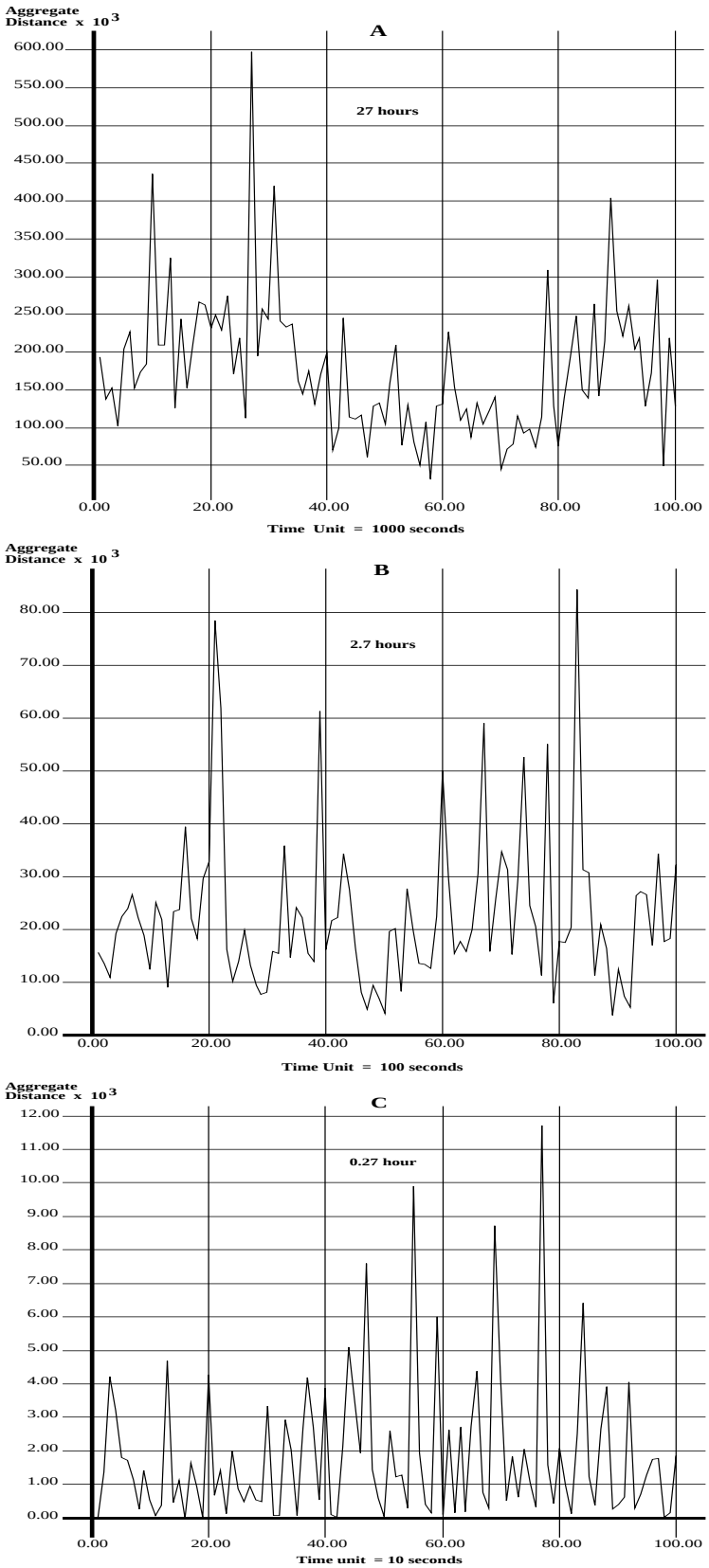

Figure 1: Aggregate Distances for Different Time Units

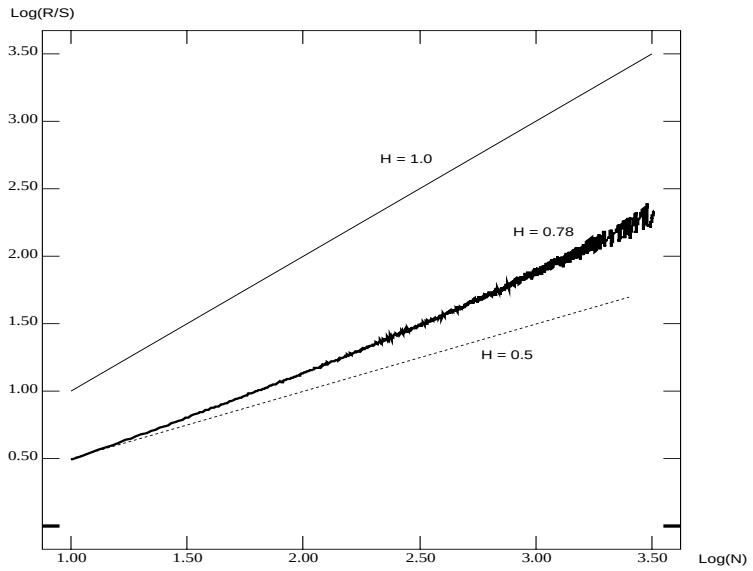

Figure 2: R/S Plot: graphical analysis of one day

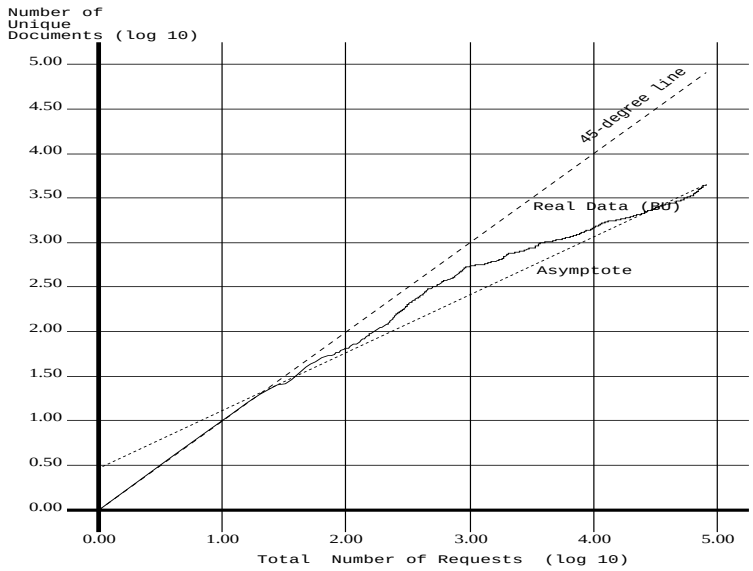

Figure 3: Growth of the accumulated number of unique documents

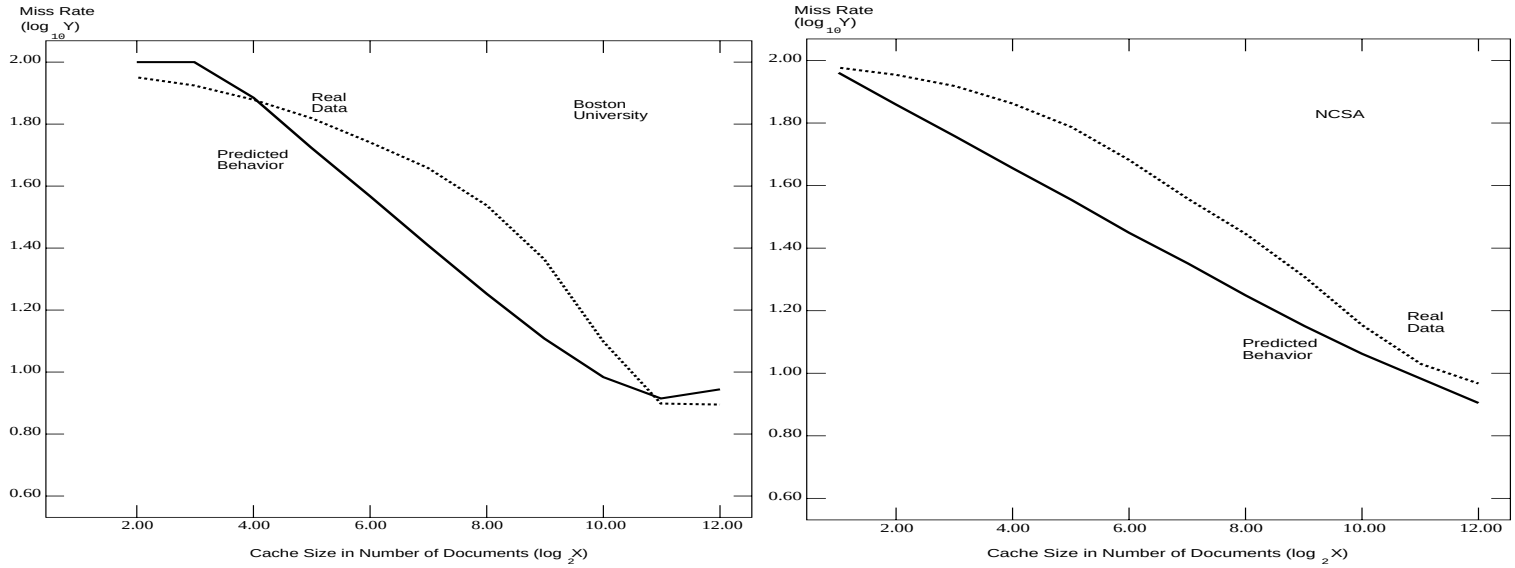

Figure 4: Simulated and Predicted Miss Ratios