Modeling Dynamic Web Content *

Antonio Barili[1], Maria Carla Calzarossa[1], and Daniele Tessera[2]

[1] Dipartimento di Informatica e Sistemistica
Università di Pavia
I-27100 Pavia, Italy
{abarili,mcc}@unipv.it
[2] Dipartimento di Matematica e Fisica
Università Cattolica del Sacro Cuore
I-25121 Brescia, Italy
d.tessera@dmf.unicatt.it

Abstract. Web sites have become increasingly complex and offer a large variety of services and contents. The proliferation of dynamic Web contents opens up new challenging performance and scalability issues. This paper addresses the characterization of dynamic Web contents by studying their update process. We identify parameters and metrics that describe the properties of Web pages and we derive an analytical model that captures their behavior. The model is then validated against simulation experiments.

1 Introduction

Web sites have become increasingly complex and offer a large variety of services and contents. Static pages, that is, pages whose contents seldom change, have been largely replaced by dynamic pages whose contents change with high rates and can be customized according to the user's preferences. Many sites rely on dynamic pages to provide users with dynamic, interactive and personalized contents. A dynamic page is typically generated by a combination of servers, that is, a front-end Web server, an application server and a database server, and involves computation and bandwidth requirements.

The proliferation of dynamic Web contents opens major performance and scalability issues and new challenges for efficient content delivery. In [1] and [2], the performance of different dynamic Web technologies is compared. The study presented in [2] shows that the overheads of dynamic contents generation significantly reduce the request peak rate supported by a Web server.

Caching at the server side can reduce server resource demands required to build the page, thus reducing the server side delays to make information available to the users. Proxy caching and content replication can reduce the latency to retrieve Web pages at the expense of very complex consistency and update

* This work has been supported by the Italian Ministry of Education, Universities and Research (MIUR) under the FIRB-Perf project.

management policies. These issues have been largely addressed in the literature and various caching and replication schemes have been proposed (see e.g., [3–9]).

An accurate characterization of Web pages is very beneficial for developing efficient caching policies. In spite of its importance, few studies focus on the behavior of Web pages. In [10] the dynamics of server content are studied by considering the process of creation and modification of HTML files of a large commercial news site. This study shows that the server content is highly dynamic. In particular, the files tend to change little when they are modified and modification events tend to concentrate on a small number of files. The stochastic properties of the dynamic page update patterns and their interactions with the corresponding request patterns are studied in [11]. This study shows that the pages of highly dynamic sport sites are characterized by a large burst of updates and by a periodic behavior. Moreover, the stochastic properties can differ significantly from page to page and vary over time. In [12], the object change characteristics are used to classify the objects at the server side. The four identified categories take into account how frequently objects change and whether their changes are predictable. In [13], the workload of a personalized Web site is studied by analyzing the document composition, the personalization behavior and the server side overheads. A methodological approach to evaluate the characteristics of dynamic Web contents is presented in [14]. As a result of this methodology, models for independent and derived parameters are obtained. These models rely on the analysis of measurements collected at six representative news and e-commerce sites. The analysis shows that the sizes of the objects can be captured by exponential or Weibull distributions, whereas the freshness times are distributed according to a Weibull or a bimodal distribution.

In this paper, we study the characteristics and properties of Web pages and we propose an analytical model that captures their update process. This model helps in better understanding the behavior of the workload of Web servers and can be used to optimize the generation and the delivery of dynamic pages and to develop consistency management policies that take into account the page characteristics.

Our study is more general than the studies described in [11] and [14] in that we predict the behavior of Web pages as a function of the behavior of their constituting objects rather than deriving the models from the observed behavior of the pages.

In what follows we assume that a Web page consists of a collection of static and dynamic objects (or fragments) whose number does not vary over time. Each object corresponds to a portion of the page and can be treated independently of the other objects. A Web page can then be seen as consisting of two components: contents and layout. The contents are associated with the objects. The layout can be seen as a special object that specifies how the contents have to be displayed. A page changes whenever its contents change, that is, the contents of one or more of its objects change. Note that our study focuses on the properties and behavior of the Web pages and it is out of the scope of this paper to model how the users access the pages.

The paper is organized as follows. Section 2 introduces the parameters that characterize dynamic Web contents and describes the analytical model that captures their update process. Section 3 presents a few simulation experiments aimed at validating the model. Finally, Section 4 summarizes the paper and outlines future research directions.

2 Analytical Model

The characterization of dynamic Web content is based on the analysis of the properties of the Web pages which are in turn derived by studying the properties of their constituting objects.

Let us model a Web site as a set $\mathcal{P}$ of N different pages $\{P_i,\ i = 1,\ \ldots, N\}$. Each page is a set of n_i distinct objects drawn from the set $\mathcal{O}$ of all possible objects $\{o_j,\ j = 1, \ldots, M\}$, where M denotes the number of different objects. An object o_j may belong to one or more pages and, in general, there is a many–to–many correspondence between $\mathcal{P}$ and $\mathcal{O}$. Note that the set $\mathcal{P}$ can be seen as a subset of $\mathcal{P}(\mathcal{O})$, where $\mathcal{P}(\mathcal{O})$ is the power set of $\mathcal{O}$.

Objects are associated with different content types and exhibit different characteristics in terms of how often and to what extent their content changes. Our study focuses on the update process of the objects, that is the sequence of their update events.

Even though objects may exhibit predictable or unpredictable update patterns, we focus on unpredictable updates as their study is more challenging. To model the update process of object o_j we consider the sequence $\{u_{o_j,k},\ k \geq 0\}$ of its update events and we denote with $t_{o_j,k}$ the instant of time of occurrence of the event $u_{o_j,k}$. In what follows we refer to the k-th update event of object o_j as the k-th generation of the object. The time–to–live of the k-th generation of object o_j, that is, the time between two successive updates, is defined as:

$$ttl_{o_j}(k) = t_{o_j,k+1} - t_{o_j,k}$$

We then introduce the counting process $\{X_{o_j}(t),\ t \geq 0\},\ j = 1, \ldots, M$ that models the number of update events occurred to object o_j up to time t. If the $\{ttl_{o_j}(k),\ k \geq 0\}$ are independent and identically distributed random variables, then $X_{o_j}(t)$ is a renewal process.

Once we have characterized the update process of the objects, we study the update process of page P_i as the superposition of the update processes of its constituting objects. For this purpose, we introduce a counting process $\{Y_i(t),\ t \geq 0\},\ i = 1, \ldots, N$ that models the number of updates of page P_i up to time t due to the updates of its objects:

$$Y_i(t) = \sum_{o_j \in P_i} X_{o_j}(t)$$

Let $\tau_{i,s}$ denote the instant of time of the s-th update event of page P_i, that is, its s-th generation. As the contents of a page change whenever the contents

of any of its constituent objects change, we obtain that $\{\tau_{i,s},\ s \geq 0\}$ is equal to $\bigcup_{o_j \in P_i} \{t_{o_j,k},\ k \geq 0\}$, with $\tau_{i,s+1} > \tau_{i,s}$. Even if the update processes of the constituent objects of page P_i are modeled by renewal processes, the page update process is not necessarily a renewal process.

The time–to–live of s-th generation of page P_i is then given by:

$$TTL_i(s) = \tau_{i,s+1} - \tau_{i,s}$$

We can predict the time–to–live of page P_i, provided that the expectations of the time–to–live of its constituent objects are known. Indeed, it is known (see [15]) that the superposition of n mutually independent renewal processes approximates a Poisson process as n goes to infinity.

An interesting metric that can be easily derived is the rate–of–change $R_i(t,T)$ of page P_i over the interval $[t, t+T]$, defined as the ratio between the number of update events occurred in the interval and its length. As the updates of page P_i are the superposition of the updates of its constituent objects, the rate–of–change $R_i(t,T)$ is obtained from the rate–of–change $r_{o_j}(t,T)$ of the objects:

$$R_i(t,T) = \sum_{o_j \in P_i} r_{o_j}(t,T)$$

This metric describes how fast the contents of a page change. Note that static objects are characterized by a rate–of–change equal to zero over any interval of length T.

As a refinement, we study the probability that page P_i sampled at time t will be up–to–date at time $t+T$:

$$D_i(t,T) \ = \ Prob\ [\ (Y_i(t+T) - Y_i(t)) = 0\ ]$$

This metric allows us to compute the expected time–to–live of a page, an information very useful to implement efficient page and object caching strategies. Note that under the assumptions of independent and stationary increments of the update process of page P_i and of mutually independent update processes of its constituting objects, we obtain that $D_i(t,T)$ is independent of t and we can write it as:

$$D_i(t,T) \ = \ D_i(T) \ = \ Prob\ [\ Y_i(T) = 0\ ] \ = \ \prod_{o_j \in P_i} Prob\ [\ X_{o_j}(T) = 0\ ]$$

A closed form of $D_i(T)$ can be obtained, provided that the probability distributions of the $X_{o_j}(T)$ are known. For example, if each $X_{o_j}(T)$ is a Poisson process with rate λ_{o_j}, then $D_i(T) \ = \ e^{-\Lambda_i T}$, where $\Lambda_i \ = \ \sum_{o_j \in P_i} \lambda_{o_j}$.

Moreover, under the same assumptions of independent and stationary increments of the update process of page P_i and of mutually independent update processes of its constituting objects, the rate-of-change $R_i(t,T)$ of page P_i does not depend on t and is then equal to $R_i(T)$.

For each page P_i, we can also compute the average number of updates per object in the interval $[t, t+T]$. This metric is obtained as the ratio between the number of update events occurred in the interval and the number of objects n_i of page P_i. Under the assumption that none of the objects has been updated more than once in the considered time interval, this metric measures the fraction of the objects of page P_i that are out–of–date after T time units.

Once we have modeled the update process of the pages, we can focus on the page size. Let $s_{o_j,k},\ k \geq 0$ denote the size in bytes of the k-th generation of object o_j. As we assume that the size of an object varies only in correspondence to an update event, we obtain that $s_{o_j}(t) = s_{o_j,k}$ for $t_{o_j,k} \leq t < t_{o_j,k+1}$.

Page size at time t is then given by:

$$S_i(t) = \sum_{o_j \in P_i} s_{o_j}(t)$$

Since an object can belong to more than one page, we can study the degree of object sharing, that is the fraction of objects shared across pages. This metric provides a measure of content reusability across pages. Let $\tilde{M}$ be the total number of objects belonging to the N pages:

$$\tilde{M} = \sum_{i=1}^{N} n_i$$

As M counts the number of distinct objects, we obtain that $\tilde{M} \geq M$. The ratio between $\tilde{M}$ and M measures the degree of object sharing. This ratio is equal to one for Web sites with no sharing at all across pages. The larger the degree of sharing the larger the probability of multiple page updates due to a single object update.

Similarly, we introduce a metric that measures the content sharing, that is the fraction of bytes shared across pages. Note that the degree of object sharing is a property of the Web site and depends on the page composition, whereas the content sharing varies with time as a consequence of the object updates. The content sharing is defined as:

$$cs(t) = \frac{\sum_{i=1}^{N} S_i(t)}{\sum_{j=1}^{M} s_{o_j}(t)}$$

This metric can be used to assess the benefits of caching at the object level. When no content is shared across pages, the value of $cs(t)$ is equal to one, hence, the benefit of caching at the object level is rather limited. As the values of $cs(t)$ become larger, caching at the object level become more and more beneficial.

3 Simulation Experiments

To validate the analytical model introduced in Section 2, we have performed a few simulation experiments. These experiments are aimed at assessing whether the

time–to–live of the pages obtained by simulation probabilistically supports the theoretical model, that is, the goodness–of–fit between the empirical distribution function of the time–to–live $TTL_i(s)$ of page P_i and the exponential distribution. For this purpose, we have applied the Anderson–Darling (A^2) test [16]. Note that the hypothesis of an underlying exponential distribution is tested at 0.05 significance level. Since it is known that this test applied to large data sets fails, we based our analysis on random samples of the update events occurred to each page during the simulation interval. Moreover, to assess the sensitivity of the results, we have analyzed the time–to–live of the pages as a function of their composition, namely, varying the number of objects constituting each page and the distribution of their time–to–live.

An obvious result obtained is that, whenever we compose in a page any number of objects, each with update events modeled by a Poisson process, the time–to–live of the page is exponentially distributed. We have further investigated the behavior of $TTL_i(s)$ when the time–to–live of the constituting objects is characterized by distributions other than the exponential. We have performed various experiments with $ttl_{o_j}(k)$ characterized by a uniform distribution. In particular, we focused on a uniform distribution over the time interval $(90, 110)$, that is, a distribution characterized by a mean and a standard deviation equal to 100 and 5.77 time units, respectively. Figure 1 shows the empirical distribution function of the time–to–live of a page consisting of 5 objects and its corresponding exponential model. The mean value of the empirical distribution is equal to

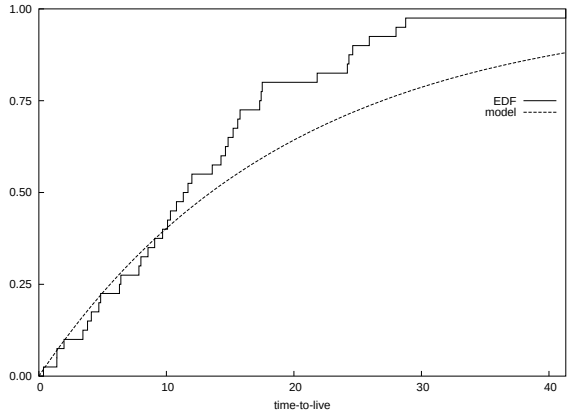

Fig. 1. Empirical distribution function (EDF) and distribution of the exponential model of time–to–live of a page consisting of 5 objects.

13.085 time units and its standard deviation is equal to 8.980 time units. Even though the distribution is slightly right skewed and its median (equal to 11.321 time units) is smaller than the mean, the A^2 test applied to this distribution does not indicate a good fit to the exponential model. The value of A^2 is equal to 1.596 and is larger than the corresponding critical value at 0.05 significance

level, that is 1.323. This could be due to the small number of objects in the page. For this experiment, we have computed $Prob\ [\ Y_i(T) = 0\ ]$ for $T = 5,\ 10$ and 15 time units. The values obtained are equal to 0.757, 0.544 and 0.375, respectively. As expected, the probability of having no update events in an interval of length T steadily decreases as T increases.

Table 1 summarizes the statistical values of the experiments with time–to–live of the objects uniformly distributed over the time interval $(90, 110)$. As can be seen, as the number of objects per page increases, the empirical distribution function accurately reflects the analytical model of $TTL_i(s)$. For pages with 10 objects or more, the A^2 test at 0.05 significance level indicates a good fit to the exponential model.

# objects	mean	st. dev.	A^2	critical value
2	56.988	26.002	4.992	1.323
5	13.085	8.980	1.596	1.323
10	10.275	8.733	0.405	1.323
20	4.680	4.135	0.415	1.323
50	1.541	1.402	0.312	1.323
100	0.826	0.868	0.167	1.323

Table 1. Statistical summary of the experiments with time–to–live of the objects uniformly distributed over the interval $(90, 110)$.

Similar conclusions can be drawn in the case of pages constituted by objects whose time–to–live is described by a uniform distribution with a standard deviation of an order of magnitude larger. There is a significant goodness–of–fit of the empirical distribution functions and the corresponding exponential models even with pages consisting of as few as 5 objects, where the value of A^2 is equal to 0.413, that is much smaller than the corresponding critical value (equal to 1.323). The same conclusions hold for pages with objects characterized by a time–to–live distributed according to a Weibull distribution. The A^2 test applied to pages with 5 objects indicates a good fit to the exponential model, whereas the test applied to pages with 2 objects rejects the exponential model. The values of A^2 are equal to 0.510 and 1.486, respectively.

To assess how these conclusions are influenced by the type or parameters of the distributions, we simulated the case of objects exhibiting a "quasi–deterministic" behavior, that is ttl_{o_j} distributed according to a uniform distribution over the time interval $(99, 101)$. In this case, the A^2 test, even at 0.01 significance level, rejects the exponential assumption for pages with a number of objects as large as 100. The value of A^2 is equal to 4.758, whereas the critical value at 0.01 significance level is equal to 1.945.

Finally, we have studied the behavior of pages consisting of a mix of objects. These pages are composed by "slow" varying and "fast" varying objects

with different distributions of their time–to–live. Figure 2 shows an example of the empirical distribution function and of the corresponding exponential model for an experiment with pages consisting of 10 objects. The time–to–live of the objects are distributed according to either Weibull, exponential or uniform distributions. The good fit that can be visually perceived is confirmed by the A^2

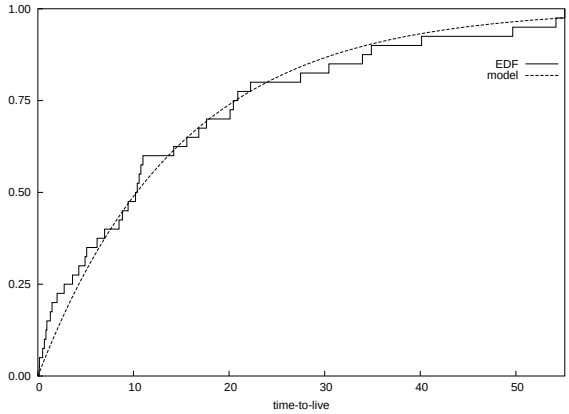

Fig. 2. Empirical distribution function (EDF) and distribution of the exponential model of time–to–live of a page consisting of 10 objects with time–to–live characterized by "mixed" distributions.

test whose value is equal to 0.623. In general, we have noticed that when the fraction of "fast" objects is dominant in a page, the time–to–live of the pages best fits an exponential distribution even when the number of objects per page is small, whereas the same conclusions cannot be drawn for pages where the fraction of "slow" varying objects is prevalent.

4 Conclusions

Performance and scalability issues have become crucial due to the large increase and popularity of dynamic Web contents. Hence, the development of efficient consistency management policies and the delivery of dynamic contents play a key role. Within this framework, an accurate description of the behavior and of the properties of dynamic Web pages is very beneficial. The characterization proposed in this paper focuses on the analysis of the properties of Web pages as a function of the properties of their constituting objects. We model the update process of the pages and we introduce metrics that describe the degree of sharing of the contents across pages. These metrics are very useful to assess the benefit of caching at the object level other than at the page level. Simulations performed varying the composition of the pages and the distribution of the time–to–live of the objects have shown that, as the number of objects per page increases, the distribution of the time–to–live of the page best fits the exponential model.

However, there are cases where these results do not hold. These cases correspond to pages consisting of a mix of "slow" varying and "fast" varying objects and of objects exhibiting a "quasi–deterministic" behavior. All these results provide very good insights to understand the behavior of dynamic content and to develop efficient caching and consistency policies.

As a future work, we plan to extend our models to take into account correlations among object update processes and time-varying change rates. Moreover, experimental data will be used for validation purposes.

Acknowledgments

The authors wish to thank the anonymous referee for the valuable comments.

References

1. Apte, V., Hansen, T., Reeser, P.: Performance Comparison of Dynamic Web Platforms. Computer Communications **26** (2003) 888–898
2. Titchkosky, L., Arlitt, M., Williamson, C.: A Performance Comparison of Dynamic Web Technologies. ACM SIGMETRICS Performance Evaluation Review **31** (2003) 2–11
3. Zhu, H., Yang, T.: Class-based Cache Management for Dynamic Web Content. In: Proc. of the Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Volume 3. (2001) 1215–1224
4. Bhide, M., Deolasee, P., Katkar, A., Panchbudhe, A., Ramamritham, K., Shenoy, P.: Adaptive Push-Pull: Disseminating Dynamic Web Data. IEEE Transactions on Computers **51** (2002) 652–668
5. Cohen, E., Kaplan, H.: Refreshment Policies for Web Content Caches. Computer Networks **38** (2002) 795–808
6. Yin, J., Alvisi, L., Dahlin, M., Iyengar, A.: Engineering Web Cache Consistency. ACM Transactions on Internet Technology **2** (2002) 224–259
7. Fei, Z.: A New Consistency Algorithm for Dynamic Documents in Content Distribution Networks. Journal of Parallel and Distributed Computing **63** (2003) 916–926
8. Mikhailov, M., Wills, C.E.: Evaluating a New Approach to Strong Web Cache Consistency with Snapshots of Collected Content. In: Proc. of the Twelfth ACM International Conference on World Wide Web (WWW'03). (2003) 599–608
9. Datta, A., Dutta, K., Thomas, H., VanderMeer, D., Ramamritham, K.: Proxy-Based Acceleration of Dynamically Generated Content on the World Wide Web: An Approach and Implementation. ACM Transactions on Database Systems **29** (2004) 403–443
10. Padmanabhan, V.N., Qiu, L.: The Content and Access Dynamics of a Busy Web Site: Findings and Implications. In: Proc. of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. (2000) 111–123
11. Challenger, J.R., Dantzig, P., Iyengar, A., Squillante, M.S., Zhang, L.: Efficiently Serving Dynamic Data at Highly Accessed Web Sites. IEEE/ACM Transactions on Networking **12** (2004) 233–246

12. Mikhailov, M., Wills, C.E.: Exploiting Object Relationships for Deterministic Web Object Management. In: Proc. of the Seventh International Workshop on Web Content Caching and Distribution. (2002)
13. Shi, W., Wright, R., Collins, E., Karamcheti, V.: Workload Characterization of a Personalized Web Site and Its Implications for Dynamic Content Caching. In: Proc. of the Seventh International Workshop on Web Content Caching and Distribution. (2002) 1–16
14. Shi, W., Collins, E., Karamcheti, V.: Modeling Object Characteristics of Dynamic Web Content. Journal of Parallel and Distributed Computing **63** (2003) 963–980
15. Feller, W.: An Introduction to Probability Theory and Its Applications - 2nd Edition. Volume II. Wiley (1971)
16. D'Agostino, R.B., Stephens, M.A., eds.: Goodness–of–Fit Techniques. Marcel Dekker (1986)