

A Generalized Model for Characterizing Content Modification Dynamics of Web Objects

Chi-Hung Chi, HongGuang Wang
School of Computing
National University of Singapore
Email: chich@comp.nus.edu.sg

ABSTRACT -- In this paper, we would like to investigate one of the most fundamental questions behind dynamic web caching, the pattern and modeling for the content modification dynamics of web objects. Since content providers cannot guarantee 100% accuracy about the life-span of a web object, a single-valued expire time is found to be too simple to describe its content modification dynamics for risk analysis to reuse copies of the data in the local cache. To capture the statistical features of the content modification dynamics of a web object, a novel mathematical quantity matrix is proposed. The three statistical metrics of measurement are: dynamic degree, predictability index, and safety bound for prediction. All our arguments and models are verified through active monitoring of the content change of real web objects.

1. Introduction

Current research efforts on dynamic web caching mainly focus on the object deposition for partial object caching [TsW00] and on aggressive server-assisted revalidation [YiA99] [DuS00] [YiA01] [YiA02] [YuB99] [NiK02] [DiA99] [TeN02] [CoK01], we would like to investigate one of the most fundamental questions behind dynamic web caching, the pattern and modeling for the content modification dynamics of web objects in this paper. It is observed that although the reusability of a web object should be driven by the content modification dynamics, current cacheability rules in proxy and explicit cache control tags in the HTTP header generated by the content servers are mainly determined by the content generation dynamics instead. Here, we refer the *content generation dynamics* as the frequency pattern of triggering of application execution in a web server to produce the final data sent out to a client upon receiving his web request and the *content modification dynamics* to the actual change pattern of the object content with respect to the previous version for the same request. Through our active monitoring of web object content on Internet, we observe that there is a big discrepancy between the content generation dynamics and modification dynamics

of a web object. Using the former one to determine an object's cacheability and time-to-live (TTL) is also found to be too conservative for content reuse, thus resulting in unnecessary network traffic.

To describe the content modification dynamics of a web object, we also argue that since its content provider cannot guarantee 100% accuracy about the object's life-span, reusing a cached object in proxy even within its TTL period actually involves risk of expired content. As a result, a single-valued expire time is definitely not enough for the proxy cache to make good content reuse decision. In this paper, we would like to address this problem through our novel mathematical quantity matrix to capture the statistical features of the content modification dynamics of a web object. The three key statistical metrics of study are the dynamic degree, predictability index, and the safety bound for prediction. With all these information, a client (either proxy or browser) can estimate the risk of content reuse in its local cache/disk; he can also balance the cost of validating (with retrieval, if necessary) the data with the original content server and the risk level of content inaccuracy that he can accept. In our study, we verify all our arguments and models through active monitoring of the content modification dynamics of real web objects on Internet.

To help our discussion in the paper, we define the following terms precisely here:

- *Life-Span of a Web Object*
It refers to the time period during which the content of an object is "fresh" and is valid to be used.
- *Time-to-Live (TTL) of a Web Object*
It refers to the time period during which a system server (such as proxy or browser cache) can use the content of a copy of an object it stores without contacting the original content server.

Note that while the life-span depends solely on the content nature, the TTL depends on many

other factors such as server storage policy, and requirements for client behavior monitoring.

2. Related Work

To improve the performance of web caching for dynamic objects, researchers also investigate various criteria to decompose a web object into static and dynamic fragments for partial object caching. These include the markup language support (such as ESI [TsW00] [Liu02] [AnJ02] [RaX03] [Xlinc]), dynamic template techniques [DoH97] [BrA02] [ChI98] [ChI00], and "reverse proxy-like" systems in front of the database and web server (e.g. XCache) [DaI01] [DuD02] [XCache]. There are also studies on the possibility of caching web query pages in proxy [LuN00] [LuN01] [LuK02].

In web information system, the interest on the dynamics of content modification of web objects is due to the huge amount of effort to maintain the most recent information content in the system. This is particularly important to search engines because crawling over the Internet might take weeks to finish. Brewington [BrC00] modeled the change of web content as a renewal process. He proposed an up-to-date measure for indexing a large set of web resource. While his model can help to reduce the bandwidth usage of web crawlers, the relative dynamic nature of contents in the monitored population is not available. Cho [ChM00] proposed and compared several estimators for the modification frequencies of web pages. In his model, he assumed that the change of a web resource can be modeled by a Poisson process, which might or might not always be acceptable [CrB97] [PaF94]. Another limitation of his study is the monitoring time interval, which is chosen to be daily. Such a large time interval is too long to capture the essential modification pattern of web content for proxy caching.

3. Modeling Content Modification Dynamics of Web Objects

In this section, we would like to propose a new quantity matrix with multiple metrics to capture the key features of the life-span distribution of web objects.

3.1. Quantity Matrix for Content Modification Dynamics

There are many features that are related to the content modification dynamics of web objects. A

quantity matrix is a measurement methodology to describe its key features so that effective content delivery services and proxy caching can be based on. In this paper, we propose three measurement metrics for the quantity matrix for web content modification dynamics. They are the *dynamic degree* for the expectation value, *predictability index* for the spreading variance, and the *safety bound* value for prediction. These features will be defined based on the distribution of the probability distribution function of the life-span of web object content.

- *Dynamic Degree:*
This parameter is to measure the central tendency of the life-span of an object. It can be determined based on the mean, median, or mode value of the PDF of an object's life-span value. Most likely, this value will be used directly in the prediction of the life-span of an object. To simplify its usage, the domain of this value ranges from 0 to 100, with the maximum value referring to the most rapid modification dynamics.
- *Predictability Index*
This index is to describe the likelihood for the life-span's PDF to be aggregated around the expectation value. It is used to measure the spreading variance of the life-span's PDF. If the life-span distribution aggregates near the centre or expectation value, it will be appropriate to use the centre value as the predicted value for the next life-span. Here, we define the domain of this predictability index to range from 0 to 1, with larger value implying better predictability.
- *Safety Bound Value*
This parameter is related to the boundary feature of the life-span value. Since the life-span only distributes over a certain range of values, the lower bound of the range can be viewed as the safety bound for prediction. It is safe because all the previously monitored life-span values always pass beyond this boundary.

3.2. Calculation of Quantity Matrix

With the three key metrics defined in the last section, we would like to define their calculation in this section. Below are the formulas we propose. Note that while there are still room for fine-tuning, the basic concept and idea are captured in the formulas. Validation of the result

of the formulas will be given in Section 4 to support our proposal here.

Assume that the PDF of the life-span value of a web object k , $PDF(Obj_k)$ is given by $f(\Delta t)$, where Δt is the life-span between two consecutive content changes:

$$\text{Dynamic Degree } (Obj_k) = \frac{1}{\Delta T_{\text{Typical}} + 1}$$

$$\text{Predictability Index } (Obj_k) = \frac{\Delta T_{\text{Typical}}}{\sigma(\Delta T_{\text{Typical}}) + \Delta T_{\text{Typical}}}$$

$$\text{Safety Bound Value for Prediction } (Obj_k) = \Delta T_{\text{Min}}$$

where ΔT_{Min} is the lower bound value for the non-zero $PDF(Obj_k)$, $\Delta T_{\text{Typical}}$ is the expectation value to represent the central position of $PDF(Obj_k)$, and $\sigma(\Delta T_{\text{Typical}})$ is the standard derivation of $f(\Delta t)$ relative to $\Delta T_{\text{Typical}}$. In our model, the content modification dynamics of a given web object is represented by the probability distribution function $f(\Delta t)$ of the life-span values of the object. The modification interval (or life-span), Δt , is defined by the time period between two consecutive content change of the object. Once $f(\Delta t)$ is known, the three key metrics for the content modification dynamics of a web object given above can be found.

The first metric, the dynamic degree, is a measure to quantify the central tendency of $f(\Delta t)$. Possible statistical values to describe the central tendency of $f(\Delta t)$ include the mean, mode, and median values. The choice of selection depends on the domain of web applications under study as well as the statistical features of interest. For example, the median value is usually chosen when $f(\Delta t)$ is skew; the mode value is usually used to represent the peak position of a typical "mountain-like" distribution. In our definition of dynamic degree, we purposely use the term "central tendency" to give flexibility to the model usage. In the above formula, the central tendency is denoted by the symbol $\Delta T_{\text{Typical}}$.

Being a quantitative measurement parameter, the dynamic degree is a non-negative value ranging from 0 to 100. A larger value of dynamic degree corresponds to a smaller value of the life-span of an object; a smaller value of the dynamic degree implies that the object of study is very static. In other words, this dynamic degree is inversely proportional to (or a reciprocal function of) $\Delta T_{\text{Typical}}$. Under the active monitoring process for the life-span history of a web object, there is an intrinsic limitation on its resolution. Within one

cycle period of monitoring, only one life-span value can be reported, independent of the actual value of the modification frequency. With the typical life-span ranging from zero to infinite, the formula for the dynamic degree will give values from 0 to 100.

The second metric, the predictability index, is a measure for the confidence level for the expectation value of the distribution to be the next life-span value. It tries to reflect the spreading variance of the PDF. The value of the predictability index ranges from 0 to 1, with larger value corresponding to a smaller derivation from the expectation value of the distribution. This implies that the predictability index is inversely proportional to $\sigma(\Delta T_{\text{Typical}})$. If two distributions have the same derivation, the less dynamic one should also be more predictable. Hence, a better measure will be $\sigma(\Delta T_{\text{Typical}})/\Delta T_{\text{Typical}}$ instead of $\sigma(\Delta T_{\text{Typical}})$. Furthermore, it is possible for the derivation to be zero, and this might map to the predictability index of infinity. To map the predictability index of this extreme situation to 1, we propose to shift the denominator by 1. The other extreme situation is when the derivation is very large. This will cause the predictability index to approach zero, which implies that the expectation value of the distribution is not suitable to be used as the next predicted life-span value.

The last metric, the safety bound for prediction, is another measure for the spreading variance of the PDF of the life-span value of a web object. Based on the history of previous life-span values, it covers the range of the distribution within which all monitored life-span values actually occur. Thus, its lower bound gives the minimum predicted life-span value that an object is still fresh whereas its upper bound gives the maximum life-span value that the object is likely to be stale. In proxy caching, researchers are more interested in the lower bound value than the upper bound value because the former one is directly related to the accuracy of the TTL setting of an object. As a result, we take the lower bound value as the safety prediction value in our model and represent it by ΔT_{Min} . Since the life-span of an object is non-negative, ΔT_{Min} (or the safety bound value for prediction) is also non-negative.

About the probability distribution function of the life-span of web objects, statistical approach is recommended. In statistics, the histogram of a random variable is usually chosen to summarize the distribution graphically. Curve fitting of the

histogram will then be used to get the proper distribution model. While there can be multiple distributions that can satisfy our need, we found that the gamma distribution is actually a fairly good choice [ChZ03] (and results in the later part of this section also support our argument here). Let us assume the PDF of the life-span of a web object be a gamma distribution and our monitoring cycle period is one minute, we have the following:

$$f(\Delta t) = \text{Gamma}(\Delta t; \alpha, \beta, \mu) = \frac{(\Delta t - \mu)^{\alpha-1} e^{-(\Delta t - \mu)/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

where α is the shape parameter, μ is the location parameter, β is the scale parameter, and $\Gamma(\alpha)$ is the gamma function with the following formula.

$$\Gamma(\alpha) = \int_0^\infty \Delta t^{\alpha-1} e^{-\Delta t} \Delta t$$

And the statistics to characterize the central tendency and the derivation of gamma distribution are:

$$\Delta T_{\text{Typical}} = \text{Mode}(\text{Gamma}(\Delta t; \alpha, \beta, \mu)) = (\alpha - 1)\beta + \mu$$

$$\sigma(\Delta T_{\text{Typical}}) = \sqrt{(\alpha + 1)\beta^2}$$

Substituting these values into the definitions of the quantity metrics gives:

$$\text{Dynamic Degree} = \frac{100}{\alpha\beta + \mu - \beta + 1}$$

$$\text{Predictability Index} = \frac{(\alpha - 1)\beta + \mu}{(\alpha + \sqrt{\alpha + 1} - 1)\beta + \mu}$$

$$\text{Safety Bound Value for Prediction} = \mu$$

The above calculation for the quantity matrix uses the mode value as the central location for the life-span distribution. If we use the mean value instead of the mode value, the formulas will be transformed into the followings:

$$\Delta T_{\text{Typical}} = E(\Delta t) = \alpha\beta + \mu$$

$$\sigma(\Delta T_{\text{Typical}}) = \sqrt{\alpha\beta^2}$$

$$\text{Dynamic Degree} = \frac{100}{\alpha\beta + \mu + 1}$$

$$\text{Predictability Index} = \frac{\alpha\beta + \mu}{(\alpha + \sqrt{\alpha})\beta + \mu}$$

$$\text{Safety Bound Value for Prediction} = \mu$$

4. Verification of Quantity Matrix for Content Modification Dynamics of Objects

In this section, we would like to use real data to verify our proposed quantity matrix for content modification dynamics given in the last section.

With consistent results obtained for hundreds of objects under active monitoring for content change, six typical objects are randomly selected from our dataset for presentation here and their data are used for curve fitting. To study the correctness of the model, we calculate the mean square error (MSE) between the observed and predicted values in each monitoring interval. If the MSE is less than or equal to 10^{-3} level, it is generally agreed that the fitted curve can approximate the actual distribution of the life-span's PDF of the observed object. In our curve fitting process, Gamma distribution is used as the basic curve family. And the parameters of the curve are determined by fine-tuning the value of the maximum likelihood estimation (MLE).

The PDF of the life-span values for each sample object is shown in Figure 1, with the parameters of its associated best-fitted gamma distribution given in Table 1. The table shows that the Gamma distribution is indeed a reasonable PDF to describe the content modification dynamics of web objects. With proper curve-fitting, all the mean square errors are about one order of magnitude less than the threshold 10^{-3} . The values for the three key metrics of the quantity matrix are also given in Table 2. The table shows that although BBC1 and GEO are more dynamic than the others, their predictability indexes are actually the worst. On the contrary, TER is the least dynamic but it is the most predictable one. It also has the largest safety bound for life-span prediction. Another interesting observation is that although YAHOO has similar dynamic degree as GRA does, its higher predictability index suggests its less variance nature. Finally, the typical life-span values of the six sample URLs are also given in the last column of Table 2. These values can be approximated to be the TTLs of the objects for proxy caching

Table 1: Parameters of the Best-Fitted Gamma PDFs for the Life-Span of Six Sample Objects

Sample URL	Curve Fitting Parameters			Mean Square Error of Curve Fitting
	Theta (μ)	Shape (α)	Scale (β)	
BBC1	0	1.75	2.7	2.50×10^{-4}
TER	57	12	0.273	1.52×10^{-4}
GRA	12	17	0.18	2.19×10^{-4}
YAHOO	11	17	0.24	3.37×10^{-4}
GEO	0	1.78	0.77	9.21×10^{-4}
BBC2	0	1.5	6.4	1.35×10^{-4}

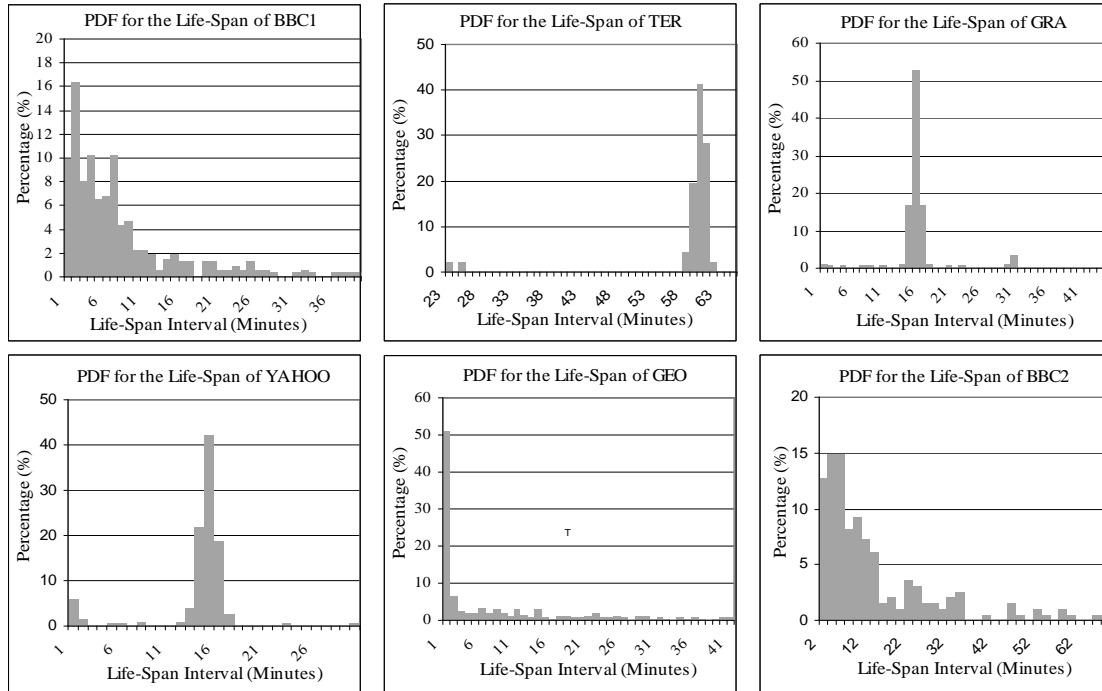


Figure 1: Probability Distribution Function for the Life-Span of Six Sample Objects

Table 2: Values of the Three Key Metrics for the Content Modification Dynamics of the Six Sample Objects

Sample URL	Quantity Matrix			$\Delta T_{Typical} = (\alpha-1)\beta + \mu$
	Dynamic Degree	Predictability Index	Safety Bound for Prediction (μ)	
BBC1	33.00	0.3120	0	2.03
TER	1.64	0.9839	57	60.00
GRA	6.30	0.9512	12	14.88
YAHOO	6.31	0.9358	11	14.84
GEO	62.50	0.3185	0	0.60
BBC2	23.81	0.2402	0	3.2

5. Conclusions

In this paper, we address one of the most fundamental questions in dynamic web caching, the modeling for the content modification dynamics of web objects. With detail monitoring data of the content change of web objects, we show that the cost of object validation without object body fetching is actually not cheap. Current simple heuristics to approximate content generation dynamics to content modification

dynamics are also found to be quite inefficient – the penalty is either the unnecessary consumption of network bandwidth or the potential of retrieving outdated object content. Instead of the single-valued expire time, we propose a quantity matrix to quantify the content modification dynamics of web objects. All the claims and insights are supported and verified with detail experimental data. These results are very important because it opens new ways to improve proxy caching and to reduce efforts to monitor web objects for content updating.

References

- [AnJ02]Anton, J., Jacobs, L., Liu, X., Parker, J., Zeng, Z., Zhong, T., "Web Caching for Database Applications with Oracle Web Cache," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, June 03-06, 2002.
- [BrA02]Brabrand, C., Moller, A., Olesen, S., Schwartzbach, M.I., "Language-Based Caching of Dynamically Generated HTML," *Journal of World Wide Web*, 5(4), 2002, pp. 305-324.
- [ChI98]Challenger, J., Iyengar, A., Dantzig, P., "A Scalable and Highly Available System for Serving Dynamic Data at Frequently

- Accessed Web Sites," *Proceedings of ACM/IEEE Supercomputing*, 1998.
- [ChI00] Challenger, J., Iyengar, A., Witting, K., Ferstat, C., Reed, P., "A Publishing System for Efficiently Creating Dynamic Web Content," *Proceedings of the INFOCOM 2000*, pp. 844-853.
- [ChM00] Cho, J., Garcia-Molina, H., "Estimating Frequency of Change," Technical report, *Stanford University*, 2000, <http://dbpubs.stanford.edu:8090/pub/1999-22>.
- [CoK01] Cohen, E., Kaplan, H., "Refreshment Policies for Web Content Caches," *Proceedings of the IEEE INFOCOM Conference*, 2001.
- [CrB97] Crovella, M., Bestavros, A., "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, 1997, pp. 835-846.
- [DaI01] Daniela, R., Iyengar, A., Dias, D., "Web Proxy Acceleration," *Journal of Cluster Computing*, 4,(4), October 2001, pp. 307-317.
- [DiA99] Dilley, J., Arlitt, M., Perret, S., JIN, T., "The Distributed Object Consistency Protocol: Version 1.0," *Technical Repprt HPL-1999-109, Hewlett-Packard Laboratories, Palo Alto, CA*, 1999.
- [DoH97] Dougliis, F., Haro, A., Rabinovich, M., "HPP: HTML Macro-Preprocessing to Support Dynamic Document Caching," *Proceedings of the USENIX Symposium on Internet Technologies and Systems*, December 1997, pp 83--94.
- [DuD02] Dutta, K., Datta, A., Meer, D.V., Suresha, Thomas, H., Ramamritham, K., "Proxy-Based Acceleration of Dynamically Generated Content on the World Wide Web: An Approach and Implementation," *Proceedings of ACM SIGMOD*, June 2002.
- [DuS00] Duvvuri, V., Shenoy, P., Tewari, R., "Adaptive Leases: A Strong Consistency Mechanism for the World Wide Web," *Proceedings of the IEEE INFOCOM*, 2000.
- [Liu02] Liu, X., "Developing High Performance Applications with Oracle 9i As Web Cache and ESI," Oracle Corp., 2002.
- [LuK02] Luo, Q., Krishnamurthy, S., Mohan, C., Pirahesh, H., Woo, H., Lindsay, B., Naughton, J.F., "Middle-tier Database Caching for e-Business," *Proceedings of SIGMOD 2002*.
- [LuN00] Luo, Q., Naughton, J.F., Krishnamurthy, R., Cao, P., Li, Y., "Active Query Caching for Database Web Servers," *Proceedings of WebDB 2000*.
- [LuN01] Luo, Q., Naughton, J.F., "Form-Based Proxy Caching for Database-Backed Web Sites," *Proceedings of VLDB 2001*.
- [NiK02] Ninan, A., Kulkarni, P., Shenoy, P., Ramamritham, K., Tewari, R., "Cooperative Leases: Mechanisms for Scalable Consistency Maintenance in Content Distribution Networks," *Proceedings of World Wide Web Conference*, May 2002.
- [RaX03] Rabinovich, M., Xiao, Z., Dougliis, F., Kalmanek, C., "Moving Edge Side Includes to the Real Edge: the Clients," *Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems*, March 2003.
- [TeN02] Tewari, R., Niranjana, T., Ramamurthy, S., "WCDP: A Protocol for Web Cache Consistency," *Proceedings of the 7th WCW*, Boulder, Colorado, August 2002.
- [TsW00] Tsimelzon, M., Weihl, B., Jacobs, L., "ESI Language Specification 1.0," 2000, <http://www.esi.org>.
- [XCache] Xcache. <http://www.xcache.com/home/default.asp?c=32&p=165>
- [Xlinc] Xlinc. <http://www.w3.org/TR/xlinc>
- [YiA99] Yin, J., Alvisi, L., Dahlin, M., Lin, C., "Volume Leases for Consistency in Large-Scale Systems," *IEEE Transactions on Knowledge and Data Engineering*, 11(4), 1999, pp. 563-576.
- [YiA01] Yin, J., Alvisi, L., Dahlin, M., Iyengar, A., "Engineering Server Driven Consistency for Large Scale Dynamic Web Services," *Proceedings of the Tenth International World Wide Web Conference*, May 2001.
- [YiA02] Yin, J., Alvisi, L., Dahlin, M., Iyengar, A., "Engineering Web Cache Consistency," *ACM Transactions on Internet Technologies*, Vol. 2, No. 3, August 2002.
- [YuB99] Yu, H., Breslau, L., Shenker, S., "A Scalable Web Cache Consistency Architecture," *Proceedings of ACM SIGCOMM*, 1999.