Exploiting Inter-reference Time Characteristic of Web Cache Workload for Web Cache Replacement Design

Agung Sediyono[a]

[a]*Informatics Department*
Faculty of Industrial Technology Trisakti University
E-mail: agung@trisakti.ac.id

ABSTRACT

Caching objects in Internet environment is aimed to reduce bandwidth consumption and increase the response time of system in term of user perception. Researcher believes that the performance of web cache is dependent on user access behavior. Therefore, many studies on workload of the internet web cache have been conducted. However, in the web cache environment, the study in inter-reference time of the successive requests is rarely. This paper tries to explore the characteristic of workload of the internet web cache, especially on the inter-reference time (IRT). Based on the correlation test and trend analysis, it can be concluded that IRT is reasonable to be applied as a metric of web cache replacement policy.

Keywords: *inter-reference time, web cache replacement*

1.0 INTRODUCTION

Caching objects in Internet environment is aimed to reduce bandwidth consumption and increase the response time of system in term of user perception. The performance of web cache is measured using the hit ratio (HR) and the byte hit ratio (BHR). HR is calculated from how many user requests that can hit the cache divided by the total requests. Mean while, BHR is calculated from how many bytes that can be hit in the cache divided by total bytes requested. In a web cache with limited cache size, the HR and BHR can be optimized by using cache replacement algorithms. Based on the research conducted by Lindemann & Waldhorst (n.d), it can be concluded that there is no cache replacement algorithm that can outperform at all workloads. The performance of the cache replacement policy is dependent on the behavior or characteristic of the web cache workload.

The research of workload of the web cache was conducted extensively (Breaslau, et al., 1999; Cohen&Kaplan, 1999; and Benevenuto et al., 2005). Breaslau, et al. (1999) concluded that the distribution of the web requests follow a Zipf-like distribution and this model can explain why the performance of the web cache is certain asymptotic properties. Cohen&Kaplan (1999) measured the regularity of the workload and use it to design the optimal cache replacement algorithm. Mean while, Benevenuto et al. (2005) explores the impact of the first timer, included the one timer, on the performance of cache replacement schemes. From these studies, there are many properties of the workload such as the object size, the frequency of references, recency, one timer, first time, type of objects, and inter-reference time of object requests that can influent the performance of the web cache. The last property that is the inter-reference time of the object requested was discussed intensively in memory cache replacement (Phalke&Gopinath, 1995; Jiang&Song, n.d; Takagi&Hiraki, 2004) and outperforms the previous cache replacement schemes, but it was rarely discussed in web cache environment. Tanaka&Tatsukawa (2003) use the inter-reference interval called II-PO for manage the size of the web cache by using the modified 2Q, but they only use one trace log as a testbed. Moreover, they did not explore in depth the characteristic of the inter-reference time of the workload. Therefore, this paper tries to explore the inter-reference time (IRT) characteristic of the workload of web cache and uses the finding to design the cache replacement algorithm. Based on the IRT characteristic of eight web trace logs from three companies: GIA, Telcom, and Peti Kemas Co., it can be concluded that there is a correlated between IRT and the temporal popularity of the web object.

The rest of this paper will be arranged as follows. Related work will be discussed in Section 2. Data collection and analysis is presented in section 3, and then the conclusion and future work are presented in section 4.

2.0 RELATED WORK

Inter-reference time of the successive object requests was extensively discussed and implemented in memory cache replacement (Phalke&Gopinath, 1995; Jiang&Song, n.d; Takagi&Hiraki, 2004). Phalke&Gopinath (1995) explored the behavior of inter-reference gap (IRG) that is the time interval between successive references to the same address. They concluded that the IRG has, in general, a repetitive behavior. Therefore, they applied a *k* order Markov chain to predict the next reference in the future. Based on the experiment, IRG can improve the cache replacement until 37% over the Least Recently Usage (LRU). Mean while, Jiang&Song (n.d) introduce the LIRS cache replacement based on Inter-reference Recency (IRR) Set. IRR uses the number of references of the other objects that is in the inter-reference time of certain object. On the other hand, they use spatial locality instead of temporal locality. They argue that the age of the object in the cache can be

measured by counting the number of references of the other object after the object measured is entered into the cache. LIRS uses two blocks of cache: LIR for low inter reference and HIR for high inter reference. By using this approach that is not depending on the detectable pre-defined regularities in the reference of the workloads, LIRS can improve the LRU performance. Mean while, Takagi&Hiraki (2004) argue that each memory address has own IRG distribution, so that they suggest to make individual probability distribution of each memory block and use the distribution to estimate the next reference in the future. This approach depends on the historical data so that it can introduce the complexity in both memory and computation.

Eventhough IRT has extensively discussed and implemented successfully in memory cache replacement, the research on the inter-reference time for web cache replacement was rarely conducted. Tanaka & Tatsukawa (2003) adopted IRT to be a metric in web cache replacement algorithm. They modified the definition of IRT as an interval time between the time of purge object and time of the miss access of that object. For example, if an object x is referenced at time t_1, t_2, and t_3, and the reference at t_2 is not in the cache, then the original inter-reference interval for object x are t_2-t_1 and t_3-t_2, but Tanaka&Tatsuka (2003) take only t_2-t_1 as a metric called II-PO for cache replacement. A small II-PO implies that if the cache had additional it could have kept the object. They implemented the II-PO metric in the modified 2Q (2Q-Opt). 2Q-Opt uses two caching areas, Q1 and Q2. Q1 is a FIFO queue that keeps objects which are referenced for the first time, and Q2 is a LRU queue that keeps objects whose references counts are more than one. Caching management is conducted by decreasing or increasing the length of Q1 or Q2 vice versa based on the II-PO value so that the total cache size is not change. The drawbacks of this approach are requiring unlimited space for recording purged objects and testing only in one trace log. Therefore, it stills not confident whether is not the result also valid for the other web trace logs. This question is reasonable because based on research conducted by Lindemann & Waldhorst (n.d), it can be concluded that there are no cache replacement algorithm that can fit at all situations. The performance of the cache replacement policy is dependent on the behavior or characteristic of the web cache workload, especially the composition of the object type in the web cache. Therefore, it urges to explore in depth the characteristic of the web cache workload in term of inter-reference time, so that it can be concluded whether is not IRT can be used to be a common metric for the different web cache workloads.

3.0 DATA COLLECTION AND ANALYSIS

Data collection and analysis describes and discusses about the raw data, data processing of raw data, data properties of the processed web caches, and the IRT characteristic and its analysis.

3.1 Raw Data

The raw data for the experiment are collected from three companies: Garuda Indonesia Airways (GIA) that has four web caches, PT Telkom (Telcom) that has three web caches, and PT Peti Kemas (PetiKemas) that has one web cache. The GIA web caches have been collected as long as three weeks from November, 1st till 18th 2008, and the Telcom web caches have been collected for one week from Nopember, 2nd till 8th 2008. Mean while, PetiKemas web cache have been collected for five weeks from June, 26th 2008 till July, 31th 2008.

3.2 Data Processing

Before the web caches workload is explored, the web caches are filtered so that only the cacheable object that will be explored.

Table 1: *The properties of the web caches under investigation*

	PT Garuda Indonesia Airways (1-18 Nov 2008)				PT Telkom (2-8 Nov 2008)			Peti Kemas (26 Juni - 31 July 2008)
	GIA #1	GIA #2	GIA #3	GIA #4	Telcom #1	Telcom #2	Telcom #3	
# of Request	3,544,156	8,269,922	4,717,459	3,137,920	5,014,879	4,560,189	8,219,840	7,558,496
# of Cachable Request	1,372,801	3,195,265	2,194,430	1,664,758	2,208,864	1,385,718	3,519,394	3,253,394
Request rate daily	**76,266**	**177,514**	**121,912**	**92,486**	**315,552**	**197,959**	**502,770**	**92,954**
% of Cachable Request	**38.73**	**38.64**	**46.52**	**53.05**	**44.05**	**30.39**	**42.82**	**43.04**
Total Size of Cachable Object (MB)	13,957.7	37,774.8	16,557.4	31,971.9	245,888. 8	99,323.9	245,888.8	66,548.2
One Timer	298,121	649,826	417,580	389,405	35,024	16,584	34,894	792,099
% of One Timer	**21.72**	**20.34**	**19.03**	**23.39**	**1.59**	**1.20**	**0.99**	**24.35**
# of Distinct Request	379,746	839,080	532,241	487,728	594,061	408,016	923,313	1,030,870
Percentage of Object Type								
- image	**41.194**	**40.596**	**40.015**	**40.768**	**57.664**	**61.111**	**60.438**	**58.888**
- audio	0.008	0.005	0.034	0.025	0.092	0.251	0.111	0.044
- video	0.035	0.000	0.059	0.141	0.097	0.264	0.147	0.136
- Text	**30.112**	**28.418**	**30.051**	**35.984**	**31.626**	**28.261**	**28.352**	**30.747**
- Application	**9.867**	**13.913**	**11.720**	**10.375**	**7.273**	**6.596**	**6.855**	**6.763**
- Others	18.785	17.067	18.120	12.707	3.249	3.517	4.098	3.421

To filter the cacheable objects, this paper adopt the rule that was also used by Casilari & Trivino-Cabrera (2008). The rule is the web request that contain the '?', 'cgi', or 'cgi-bin' will be discarded from the web cache log, and only those request with a cacheable response code, that is, 200 (OK), 203 (Partial), 206 (Partial Content), 300 (Multiple Choices), 301 (Move), 302 (Redirect), and 304 (Not Modified) will be used in the experiment. The IRT is counted from top one thousand of the popular objects, and then the IRT characteristics are plotted using MATLAB 7.01.

3.3 Data Properties

The properties of the web caches workload are presented in Table 1. From the Table 1, it can be described that the cacheable requests are below 53 % of total web requests. The percentage of one timer is different among three companies, but for the cache in the same company the one timer is nearly equal. In all web caches, the object type is dominated by application, image, and text. More over, the composition of object type contained in the cache in the same company is nearly equal. The important property that is related to IRT is the web request rate that shows the density of web request. From Table 1, It can be described that all web cache have different web request rate

3.4 Analysis of IRT

Analysis of IRT will be described as follows. First, the age of request and the IRT value for one of object is identified. Then, the IRT values for all web caches are compared and the characteristic of the IRT for each object type in each web caches is explored in depth. Finally, we try to make a conclusion from those evidences.

Tabel 2 : *The correlation between IRT and other metrics*

	Correlation between IRT and		
Web Cache	Elapse time	Frequency	Object Size
GIA#1	0.0320	-0.1368	-0.0472
GIA#2	**-0.1080**	-0.1193	-0.0757
GIA#3	-0.0592	-0.1763	-0.0652
GIA#4	-0.0789	-0.1087	0.0414
Telcom#1	-0.0160	-0.1613	N/A
Telcom#2	-0.0113	-0.1436	**-0.1436**
Telcom#3	-0.0085	-0.1846	-0.0455
PetiKemas	-0.0247	-0.1549	**-0.1549**

In this paper, the age of request is defined as a time of last request of certain object is subtracted by time of the first request of that object. The age of request represent how long the object is interested by the users. From Figure 1, it can be seen that the age of request, in general, is constant to frequency of reference for all web caches. This constant is varied among the web cache. On the other hand, if the age of request is constant for all frequencies, the IRT is linier in logarithmic of the frequency of request. If a frequency of request can be used in LFU so IRT does. Mean while, the successive sequences of the IRT are varied and for the first and last sequence the IRT are, in general, bigger than in the middle sequence. More over, the characteristic of frequency follows Zipf-like with certain parameter value (see Figure 2). This finding is conformance with Breaslau, et al. (1999).

Based on Table 2, the correlation between IRT and elapsed time, or IRT and object size, or IRT and frequency of references is weak, only below 3,5 % of the variation is related. It means that the IRT can be used to be one of the metrics implemented in the web cache replacement policy.

The characteristic of IRT versus the popularity of the objects is shown in figure 4. From this characteristic it can be concluded that each web cache has its own IRT characteristic that is linier fashion. There is a correlation between IRT and the popularity of objects. The IRT tends increases by decreasing of popularity of the objects. Figure 5 and 6 shows the characteristic of IRT for each object type in several web cache workloads. From this figure, we can conclude that each workload has own characteristic. However, the characteristic of IRT for each object follows the linier fashion to the popularity of the objects.

The research conducted by Lindemann& Waldhosrt (n.d) argues that the performance of web cache replacement is dependent on the workload, especially the composition of the objects contained in the web cache workload. Based on table 3, there are a few strong correlations between object types in each web cache workload. It means that this evidence supports Lindemann& Waldhosrt (n.d) conclusion because each object type has own characteristic for each web cache workload. Figure 7, 8, and 9 shows the difference between weak and strong correlation related to Table 3.

Table 3: *The Correlation Between Objects*

		Image	Text
GIA#1	Application	*0.054923*	*-0.01379*
	Image	1	**0.728943**
GIA#2	Application	0.318007	0.442942
	Image	1	0.400633
GIA#3	Application	0.648293	0.597376
	Image	1	0.4659
GIA#4	Application	0.551053	-0.16296
	Image	1	**0.958899**
Telcom#1	Application	0.59086	**0.755819**
	Image	1	0.523466
Telcom#2	Application	0.566894	0.55498
	Image	1	0.55498
Telcom#3	Application	**0.727655**	0.591803
	Image	1	**0.727966**
PetiKemas	Application	0.425899	0.425899
	Image	1	0.455599

4.0 CONCLUSION AND FUTURE WORK

This paper shows the characteristic of IRT for eight web cache workloads from three companies. Based on the correlation test, it can be concluded that IRT can be used to be a metric in determining object replacement, beside other attributes such as frequency of reference, object size, and elapsed time. Eventhough each web cache workload has own IRT characteristic, it can be generalized that the characteristic of IRT tends proportional to the popularity of objects for all web cache workloads. This fact makes more confident to apply the IRT as a metric for designing the replacement method in limited size of the web cache.

The future work is to implement IRT for a metric in determining object replacement in web cache, and compare to the previous web cache replacement schemes, especially that close to LFU family.

ACKNOWLEDGEMENT: Special thank to Garuda Indonesia Airways, PT Telkom, and PT Peti Kemas that have prepared and given the trace log of the web cache for this research.

REFERENCE

Benevenuto, Fabricio et al. (2005). Web cache replacement policies: properties, limitation, and implications. Proceeding of the Third Latin American Web Congress (LA-WEB'05).

Breaslau, Lee et al. (1999). Web caching and zip-like distributions: evidence and implications. IEEE INFOCOM, Vol XX No. Y.

Casilari, F.J. & Trivino-Cabrera, A. (2008). A windows based web cache simulator tool. Conference of SIMUT Tools, Marsielle, France.

Cohen, Edith & Kaplan, Haim. (1999). Exploiting regularities in web traffic patterns for cache replacement. STOC'99 Atlanta GA, USA

Jiang, Song & Zhang, Xiaodong. (n.d.). LISRS: an efficient low Inter-reference recency set replacement policy to improve buffer cache performance. IEEE explorer.

Lindemann, Cristoph & Waldhosrt, O.P. (n.d). Evaluating the impact of different document types on the performance of web cache replacement schemes. http://www4.cs.uni-dortmund.de/~Lindemann/

Phalke, Vidyadhar & Gopinath, Bhaskarpillai. (1995). An inter-reference Gap model for temporal locality in program behavior. SIGMETRICS'95 Ottawa, Ontario, Canada.

Takagi, Masamichi & Hiraki, Kei. (2004). Inter-reference Gap Distribution replacement: an improved replacement algorithm for set-associative caches. ICS'04 Saint Malo, France.

Tanaka, Atsuhiro & Tatsukawa, K. (2003). Interreference interval for purge objects: a metric for design and analysis of web caching algorithm. Internet System Research Laboratories, NEC Corp., Japan.

Zhang, Junbiao et al. (n.d). Web caching framework: analytical models and beyond. C&C Research Laboratory, NEC USA, Princeton, New Jersey

Appendix

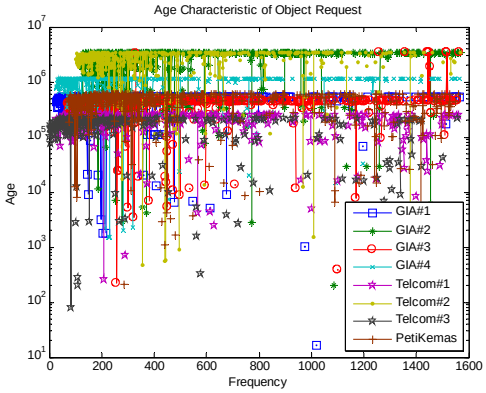

Figure 1 The characteristic of the age versus frequency of reference

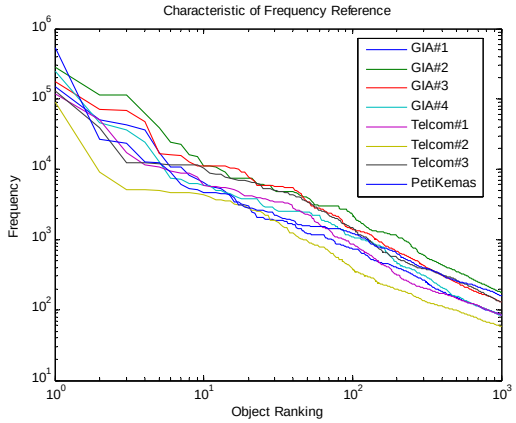

Figure 2. The characteristic of referenced frequency

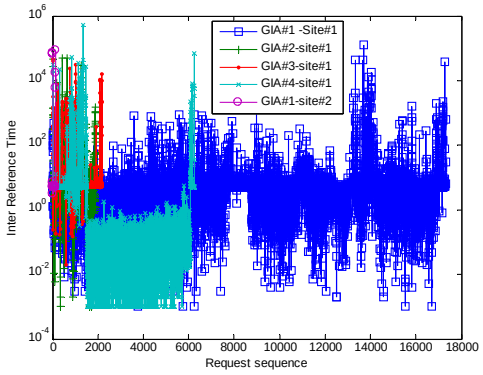

Figure 3 The IRT value for certain sequence of request

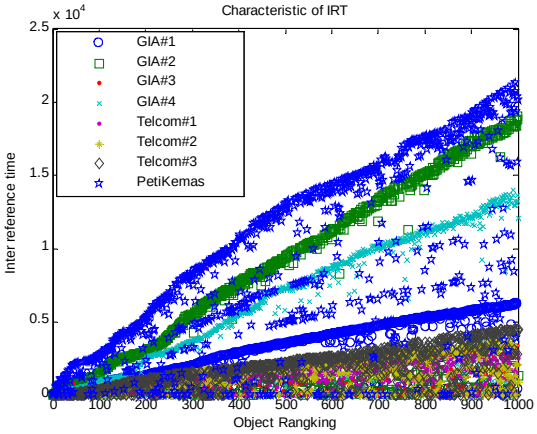

Figure 4. IRT characteristics

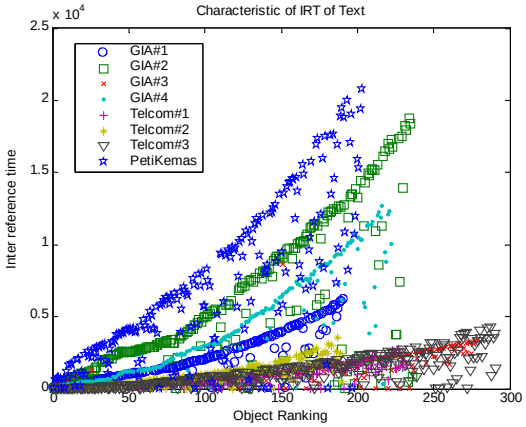

Figure 5. IRT characteristic of text object

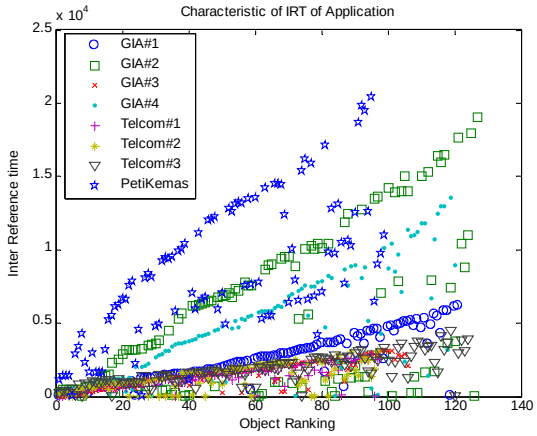

Figure 6. The characteristic of IRT for application objects

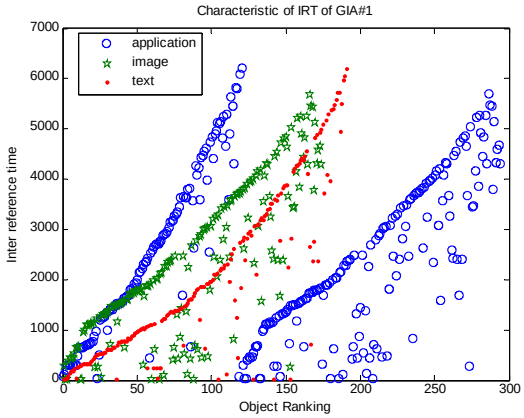

Figure 7. The characteristic of IRT for objects in GIA#1

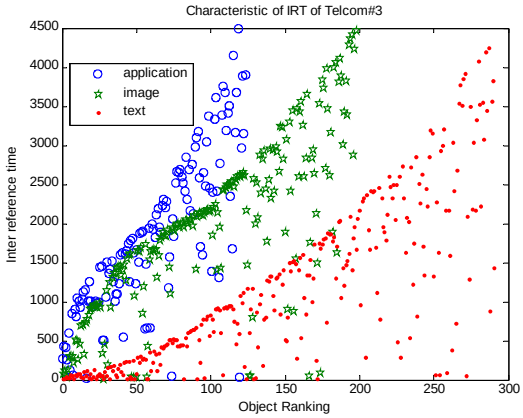

Figure 8. The characteristic of IRT for objects in Telcom#3

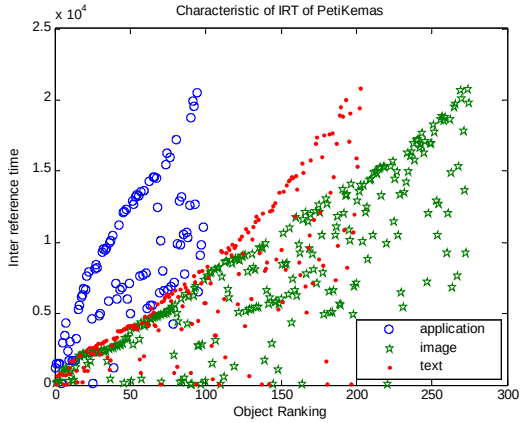

Figure 9. The characteristic of IRT for objects in PetiKermas