ABSTRACT

Caching objects in Internet environment is aimed to reduce bandwidth consumption and increase the response time of system in term of user perception. In limited size of web cache, it needs to manage the objects in web cache so that the hit ratio and byte hit ratio are optimal. For solving this problem, researcher propose many web cache replacement policies such as LRU, LFU, GDS, GDSF, GD*, and IRT. In evicting object from web cache, each policies uses recency, frequency, size, size and frequency, or inter-reference time of object requested by user. The last metric, inter-reference time, has been explored intensively in previous research in memory cache replacement. This research proposes to apply perfectly dynamic average of inter-reference time (PDA-IRT), an extended version of DA-IRT, as a metric in evicting object from the web cache. Based on experiment using web trace log from three companies, it can be concluded that PDA-IRT can improve DA-IRT in term of HR and BHR.

Keywords: inter-reference time, perfectly dynamic average of inter-reference time, web cache replacement

1 INTRODUCTION

Caching objects in Internet environment is aimed to reduce bandwidth consumption and increase the response time of system in term of user perception. In an unlimited size of web cache, there is no problem in caching objects. Otherwise, in the limited size of web cache, it needs to manage objects in web cache so that the performance of web cache can be maximized. Mostly, 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. At the beginning, the researcher adopted the memory cache replacement such as least recency usage (LRU) and least frequent Usage (LFU) to be implemented in web cache environment. 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 relies 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). 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 influence the performance of the web cache. Based on these properties, researcher proposes many web cache replacement policies such as variance of LRU and LFU, Greedy Dual Size (GDS) and its variance, and inter-reference time (IRT). The last approach, IRT, was discussed intensively in memory cache replacement (Phalke&Gopinath, 1995; Jiang&Song, 2002; 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 after purged object (II-PO) for managing the size of the web cache by using the modified 2Q, but they use only one trace log as a testbed.

Based on IRT characteristic of eight web trace logs from three companies: GIA, Telcom, and Peti Kemas Co., it can be concluded that there is a correlation between IRT and the temporal locality of the web object. Therefore, Sediyono (2009) implements the dynamic average of inter-reference time (DA-IRT) as a metric for web cache replacement policy. However, the performance of DA-IRT relies on the behavior of the user requests. This paper tries to improve DA-IRT by recording
2 RELATED WORK

Even though IRT has extensively discussed and implemented successfully in memory cache replacement (Phalke&Gopinath, 1995; Jiang&Song, 2002; Takagi&Hiraki, 2004), the research on the inter-reference time for web cache replacement was rarely conducted. Tanaka & Tatsukawa (2003) and Sediyono 2009) adopted IRT to be a metric in web cache replacement algorithm. Tanaka & Tatsukawa (2003) 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 \( t_3-t_1 \) and \( t_2-t_3 \), 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 two queues (2Q) called 2Q-Optimal (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 still not confident whether or 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.

Sediyono (2009) implements IRT approach called a dynamic average of IRT. This approach is obtained based on the fact that there is a correlation between the temporal locality and the average of IRT. Based on this research, it can be concluded that the stronger correlation results a stronger performance. But, there is an open question that has to be answered such as what are the other dominant parameters that can improve the web cache performance, how about the impact of maintaining IRT value of the purged objects on the web cache performance.

3 PERFECTLY DYNAMIC AVERAGE OF IRT

In this section it will be described why the dynamic average of IRT (DA-IRT) can be used as a metric in evicting object in the web cache, how this metric will be implemented so that the computation time will be efficient, and why the perfectly dynamic average of IRT (PDA-IRT) is needed. Finally, this section is closed by presenting the pseudo code of PDA-IRT.

Rationale

Based on the characteristic of IRT average, it can be shown that there is a strong correlation between IRT average and the temporal locality (Sediyono, 2009). Therefore, it is reasonable to implement inter-reference time as a metric for web cache replacement policy. However, if this metric will be implemented in web cache replacement
policy, it is calculated from previous set of log trace. If this approach will be taken, it is not only takes time to compute the average of IRT but also there is difficulty to determine how long the previous log trace will be taken. Therefore, the IRT average will be approximated by the dynamic average of IRT (DA-IRT). In this approach, the average of IRT is calculated on every object in web cache by dividing the cumulative of the IRT by the frequency of reference (Sediyono, 2009) (see Equation 1)

\[
IRT_{ni} = \frac{IRT_{ni-1} + (t_{ni} - t_{ni-1})}{f_{ni}} \quad \text{Eq. 1}
\]

where \(IRT_{ni}\) is a dynamic average of inter-reference time of the reference \(n\) at time \(i\), \(t_{ni} - t_{ni-1}\) is inter-reference time of reference \(n\) at time \(i\), \(f_{ni}\) is frequency of reference \(n\). Notable, for the first timer object, the average of IRT is assumed equal to the first reference time and placed into the web cache based on LRU policy among the first timer object.

Object with a largest average IRT value will be evicted first from the web cache. This approach assumes that object with smaller average IRT value will be accessed sooner, so that it has to be kept in the web cache in order to be hit.

In certain situation, DA-IRT allows a long popular object is replaced by a shortly popular object, and then a long popular object will be back to the web cache as a first timer object. This state is not fair, so it needs to record the DA-IRT value after object is evicted from the web cache, and to use the recorded DA-IRT value if the object is referred. This policy is called perfectly dynamic average of IRT (PDA-IRT), analogy of perfect LFU.

Implementation

PDA-IRT is implemented using linked list data structure. The data attribute saved in the web cache are object size, lastly referred time, cumulative IRT, and frequency of reference. The cumulative IRT and frequency of reference of the object will be updated if the object is referred. Every updating the average of IRT, the linked list has to be sorted descendent by the value of average IRT. If the object is evicted from the web cache it is saved in the list of purged object. The algorithms of the PDA-IRT as follows:

Input: X the object requested by user

Process:
If Object X is in Cache
Add cumulative IRT by current IRT
Increase the frequency by one
Calculate the DA-IRT
Sort descendent by DA-IRT value
Calculate HR, BHR
Else
While there is no the space of cache for X
Evict the object with largest DA-IRT value and save it into the list of purged objects.
If object X in the list of purged objects
Fetch X from the list of purged objects and enters into the web cache
Else
Enter object X into the cache using LRU policy among the first timer

Output: HR, BHR

4 METHODOLOGY

The methodology used in this research is experimental-based methodology. The experiment is conducted by simulating the web cache replacement policy and using the web trace log as an input. This section describes and discusses about the evaluation criteria and the data preparation for the web cache simulation.

Evaluation Criteria

The criteria of evaluation is used to asses the performance among web cache replacement policies. Based on the previous research, the criteria of evaluation for the web cache replacement are Hit Ration (HR) and Byte Hit Ratio (BHR). The HR is ratio between the number of referred objects and the number of requested objects. Meanwhile BHR is ratio between the number of byte of referred objects and the number of byte of requested objects. The formulation of HR and BHR as follows:

\[
HR = \frac{\sum \text{hit}}{\sum \text{request}} \quad \text{Eq. 2}
\]

\[
HBR = \frac{\sum \text{byte_hit}}{\sum \text{byte_requested}} \quad \text{Eq. 3}
\]

Data Preparation

This section discusses about data testbed for simulation such as the raw data, data processing, and data properties.
Raw Data

The raw data for the experiment are collected from three companies: Garuda Indonesia Airways (GIA), PT Telkom (Telcom), and PT Peti Kemas (PetiKemas). 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 November, 2nd till 8th 2008. Meanwhile, PetiKemas web cache have been collected for five weeks from June, 26th 2008 till July, 31st 2008.

Data Processing

Before the web caches workload is explored, the web caches is filtered so that only the cacheable object that will be explored. To filter the cacheable objects, this paper adopts the rule that was also used by Casilari & Trivino-Cabrera (2008). The rule is the web request that contain the ‘?’ or ‘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.

Data Properties

The properties of the web caches workload are presented in Table 1. From the Table 1, it can be concluded that the cacheable requested objects are below 53 % of total web requests. The percentage of one timer is different among three companies, but the one timer is nearly equal for the cache in the same company. 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.

Simulation

The simulation is conducted using computer program in C# language. The web cache replacement policies compared in the experiment are LRU, LFU, GDS(1)-GDS having cost value one (1)-, DA-IRT, and PDA-IRT itself. The size of web cache is varied in range 20, 40, 60, 80 percent of the total size of distinct requests in testbed. These policies are implemented in same testbed and then HR and BHR of each web cache replacement policy and web cache size are calculated.

5 EXPERIMENT RESULT AND ANALYSIS

In general, increasing web cache size tends increasing either hit ratio or byte hit ratio. This characteristic confirms to the previous research. GDS(1) outperforms all web cache replacement policies in terms of HR, but poor performance of BHR. This result is also conformity with previous research. The HR of DA-IRT outperforms LRU and LFU for Petikemas and GIA#1 testbed, while for Telkom#2 testbed DA-IRT doesn’t outperform all of web replacement policies observed in this research. In general, it can be concluded that the performance of DA-IRT relies on the behavior of user request. By recording the DA-IRT value for purged objects, the performance of web cache can be improved regardless the characteristic of user access (see Fig. 1, 2, and 3). Moreover, the significant improvement occurs in the Telkom#2 testbed where the LFU outperforms the DA-IRT policy. However, implementation of PDA-IRT has a cost in saving purged object information. Therefore, even though there is an improvement in performance, it must be calculated the cost that will be paid. For example, in case of GIA#1, implementation of PDA-IRT is costly because the performance improvement of web cache replacement is insignificant.
CONCLUSION AND FUTURE WORK

By recording and reusing the DA-IRT value of the purged objects in PDA-IRT policy, the performance of PDA-IRT in term of HR and BHR can be increased. The significant improvement is occurred in the condition where the LFU outperforms the PDA-IRT policy.

It must be explored in depth in order to find out the other properties of user behavior so that the implementation of IRT and its variance outperform other policies. Other way, by expecting the next referred object, the timing of evicting objects can be controlled in order to optimize the web cache replacement performance.

ACKNOWLEDGMENT: 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