An Integration of Web Caching and Web Prefetching using Cyclic Behaviour Analysis of Web Sequential Patterns

Srikantaiah K C¹, Krishna Kumar N¹, Venugopal K R¹, L M Patnaik²

¹Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore, India
²Honorary Professor, Indian Institute of Science, Bangalore, India
srikantaiahkc@gmail.com

Abstract. Web caching is the process in which the Web objects are temporarily stored to reduce bandwidth consumption, server load and latency. Web Prefetching is the process of fetching Web objects from the server before it is actually requested by the client. Integration of caching and prefetching can be very beneficial as the two techniques can support each other. By implementing this integrated scheme in a client-side proxy, the perceived latency can be reduced for not one but many users. In this paper, we propose a new integrated caching and prefetching policy called the WCP-CMA which makes use of a profit-driven caching policy that takes into account the Periodicity and Cyclic Behaviour of the Web access sequences for deriving prefetching rules. Our experimental results have shown a 10-15% increase in the hit ratios of the cached objects compared to existing scheme.

Keywords: Periodicity, Sequential Pattern Analysis, Web Caching, Web Log, Web Prefetching.

1 Introduction

A Web object is any entity that is stored in the Web server and requested by the clients (browsers). It can be a Web page (such as HTML, XML, etc.), or an image (GIF, JPEG, PNG, etc.) or a multimedia file like audio or a video file. These Web objects can contain links among them and are generally referred to as hyperlinks in HTML context. Caching can be performed at the client by browsers, or by proxies at client-side or proxies at server-side. Applications of Web caching and Web prefetching are (i) A search engine may cache a website. (ii) Browser can prefetch and cache hyperlinked documents based on their hit rate. (iii) An image-sharing website may prefetch popular images and cache them in the client (iv) Websites such as YouTube can prefetch and cache videos based on user’s viewing history. Web caching is achieved through the techniques (i) Prediction-based Buffer Manager (ii) Proactive Caching (iii) Profit-driven Cache Replacement (iv) Cache Coherence (v) Hierarchical Caching. A Web object \( o_j \) is said to be an implied object whenever it is referenced through a link from the Web object \( o_i \) and the 2-sequence \( (o_i, o_j) \) should be
frequent. The periodicity $P$ between a pair of Web objects $(o_i, o_j)$ is defined as the time period $t$ after which $o_j$ shall be accessed periodically after $o_i$ has been accessed.

**Motivation:** The Internet faces many challenges regarding the quality of service provided to the clients with respect to parameters such as efficient bandwidth usage, latency, hit rate etc. There are many solutions devised over the years to overcome these challenges and Web Caching and Web Prefetching are two such methods. The profit function employed by Integration of Web Caching and Prefetching (IWCP) [1] makes use of the confidence value of the prefetching rules. It completely disregards the possibility that objects with very high profit values may remain in the proxy’s cache indefinitely.

**Contribution:** In this paper, we propose a new approach Web Caching and Prefetching with Cyclic Model Analysis of Web object Sequences (WCP-CMA) which uses the periodicity and cyclic behaviour [2] of 2-sequence Web access patterns instead of confidence for finding out the prefetching rules to alleviate the disadvantage of IWCP.

Several techniques have been proposed for Web prefetching and caching. Eden, et al., [3] proposed a user-interactive prefetching scheme where the user is in control of the prefetching process and decides what and when to prefetch, thus eliminating extra traffic in the network by not prefetching irrelevant data that are not required by the user. Kim, et al., [4] implemented a novel mechanism of display-based prefetching where prediction is done by examining the access rates of all the documents displayed in the browser. Prefetching is done by ordering the access rates of all the predicted objects in decreasing order. Feng, et al., [5] designed a prefetching scheme which makes use of a multi-dimensional probability matrix to compare hit rates and analyze sequence patterns, so that it is possible to prefetch multiple Web pages.

Palli et al., [6] proposed a clustering-based prefetching scheme where a graph-based clustering algorithm identifies clusters of correlated Web pages based on the users’ access patterns and integrates Web caching and prefetching. Huang, et al., [7] developed two techniques, the Access Sequence Miner [8] and the Prediction-based Buffer Manager to prefetch and cache and also dynamically adjust cache and prefetch buffer sizes for efficient utilization of the available memory space. Nigam, et al., [9] analyzed the use of the Markov model on Web caching and prefetching and introduced a mechanism called the Dynamically Nested Markov Model (DNMM) that uses two types of caches called regular cache, which is used for normal caching of the accessed Web pages and prefetch cache, which is used for caching prefetch content.

## 2 Problem Definition and Mathematical Model

### 2.1 Problem Definition

Given the Web log $W$ and the set of Web Objects $O=\{o_i; 1 \leq i \leq n\}$ in a Web server. Our objective is to design an efficient algorithm WCP-CMA that integrates Web caching and Web prefetching in a client-side proxy system and utilizes the periodicity and cyclic behaviour of the 2-sequence patterns to determine when the object should be placed or removed from cache to improve hit rate and increase
spatial locality. In our simulation, the time window is assumed to be constant for all the Web objects.

2.2 Mathematical Model

2.2.1 Finding Prefetching Rules using Periodicity

Consider the Web log $W$ and a database $SP$ consisting of 2-sequence patterns which are mined from $W$. Consider $\langle o_i, o_j \rangle$ as a 2-sequence pattern in $SP$ where object $o_i$ is accessed at time $T_1$ and object $o_j$ is accessed at time $T_2$ after $o_i$ is accessed. The time interval between $T_1$ and $T_2$ is $x$. Trend Distribution Function (TDF) \cite{2} $f(x)$ is used to find the number of occurrences of the 2-sequence at the time interval $x$.

Generalized Periodicity Detection (GPD) \cite{2} function determines whether a certain sequence (2-sequence in this case of Web access sequences) has periodicity and computes the value of the periods if periodicity exists based on regression. The regression model is in the form of $y = ax + b$ where $a$ is the slope of the TDF curve and $b$ is the intercept. GPD function $g(x)$ is given by

$$g(x) = \frac{f(x)}{y}$$

Now the mean error between two values of GPD functions, $g(x_i)$ and $g(x_{i+P_1})$ is given by

$$\text{error}(P_1) = \frac{\sum_{i=1}^{n-P_1} |g(x_{i+P_1}) - g(x_i)|}{\sum_{i=1}^{n-P_1} |g(x_i)|}$$

Among these errors, the value of $P_1$ that gives the minimum error is the Periodicity $P$, i.e.,

$$P = \arg \min \{\text{error}(P_1)\} \quad 1 \leq P_1 \leq \frac{n}{2}$$

Trend Modelling (TM) \cite{2} is applied to find out the frequency of the trends using polynomial regression which is of the form:

$$(ax + b) = \sum_{i=0}^{m} z_i (x \mod P)^i$$

where $Z = \{z_0, z_2, \ldots, z_m\}$ is the set of coefficients of polynomial regression. Finally, the cyclic behaviour which gives the stopping criteria $C$ is determined using TM and GPD recursively to get Periodicity and a polynomial. The rules for prefetching the Web object $o_j$ in the sequence $\langle o_i \rightarrow o_j \rangle$ can be found out using the Periodicity of this sequence.

2.2.2 Profit Function

Consider reference rate $\delta_i$ of object $o_i$ is the rate of reference of a Web object within a given time window $w$. The Profits of a Web object $o_i$ is dependent on the number of references ($\delta_i w$) to $o_i$ and Periodicity $P_i$ and threshold of Cyclic Behaviour $C$. The profit value of $o_i$ depend on $F_i$. The duration in which the prefetched object $o_i$ must remain in the cache is determined using the profit function given by:
The WCP-CMA algorithm is shown in Table 1.

Table 1. WCP-CMA Algorithm

<table>
<thead>
<tr>
<th>Purpose:</th>
<th>To perform integrated Web Prefetching and Caching in client-side proxy by using Cyclic Model Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input:</td>
<td>Rule Table, Web object ( o_i ), Reference Rate ( \delta_i ), Time window ( w ), Frequency ( F_i ), Periodicity ( P_i ) and Cyclic Behaviour ( C_i ) of ( o_i ) to ( o_i ), Cache and Linked List ( L )</td>
</tr>
<tr>
<td>Output:</td>
<td>Cache and Linked List ( L )</td>
</tr>
</tbody>
</table>

\[
\text{Profit}(o_i) = \delta_i w + \frac{C_i - (P_i \times F_i)}{C_i} \tag{5}
\]

WCP-CMA-Proxy(o_i){
    // Whenever object \( o_i \) has been requested from Client
    If \( o_i \notin \text{Cache} \) Then
        Send \( o_i \) from Cache to the Client
        Increment \( F_i \)
        Compute \( \text{Profit}(o_i, F_i, C_i) \)
        If \((F_i \times P_i) > C_i\) Then
            Delete \( o_i \) from \( L \) and Cache
        Endif
        Read prefetching rules corresponding to \( o_i \) and store into \( H \)
    Else
        Read \( o_i \) from Server and send to Client
        Store \( o_i \) in Cache
        Read prefetching rules corresponding to \( o_i \) and store into \( H \)
        Compute \( \text{Profit}(P_i, 0, C) \)
    Endif
    For Each object \( o_j \in H \) do
        If \( o_j \notin \text{Cache} \) Then
            Insert \( o_j \) and its \( F_j \) and \( C_j \) into \( L_j \)
            Compute \( \text{Profit}(P_j, 0, C) \)
        Else
            Update \( o_j(P, C) = \text{min}(\text{Old}_P, P_j), \text{min}(\text{Old}_C, C_j)\)
            Compute \( \text{Profit}(P_j, F_j, C) \)
        Endif
    Endfor
    Sort \( L(o_i) \) according to decreasing order of \( \text{Profit}(o_i) \)
    While(Number of items in \( L(o_i) \) > Cache size) do
        Delete Last item from \( L(o_i) \)
    Endwhile
    For Each item \( L(o_i) \) do
        If \( o_i \notin \text{Cache} \) Then
            Read \( o_i \) from Server and Store in Cache
        Endif
    Endfor
}
3 Experimental Results

The algorithm WCP-CMA has been implemented using JSP (Java Server Pages). The various operations pertaining to the proxy and the server have been implemented using Java in a Pentium Dual Core processor environment, with a 2GB Memory and 100 GB HDD. Apache Tomcat has been utilized as the server, while Google Chrome Web browser has been utilized as the client.

A synthetic data set comprising of 200 Web objects with 559 links amongst them has been used in the simulation. The WCP-CMA algorithm has been implemented in the proxy in the client-side and will be triggered on every request from the client. The cache size is the maximum number of Web objects that the proxy’s cache can store at a time. The cache size should be as small as possible for the efficient running of the algorithm. The algorithm WCP-CMA has been simulated for different cache sizes and has been compared with IWCP.

Cache Hit Ratio is defined as the percentage of Cache Hits with respect to the total number of transactions that have occurred over a certain period of time. If the implied Web object is found in the proxy’s cache itself, it is a “hit”, otherwise it is a “miss” (object has to be fetched from the Web server). In Figure 1, it can be seen that the hit ratios are higher while using WCP-CMA instead of IWCP. Also, the hit ratios of the two schemes differ more in the region between 4 and 6 and tend to become equal as size of the cache increases.

Figure 1 clearly shows that WCP-CMA increases the hit ratios of the cached Web objects by 10-15% compared to IWCP. This is due to the fact that the Web objects cached using IWCP tend to remain in the proxy’s cache for duration longer than necessary. It would suffice to have the implied Web objects cached only up to the threshold of their cyclic behaviour. Since IWCP utilizes the confidence of Web access sequences and not their periodicity and cyclic behaviour, it would naturally result in lower hit ratios than that of WCP-CMA. This is because the new implied objects

![Fig. 1. Cache Size vs. Hit Ratio](image-url)
which have lower profit values than the previously cached objects may not be allowed to be cached even though they may help in preventing at least one miss. Hence, some of the implied objects cached using IWCP not only remain in the cache for a large duration of time and never be accessed, but also cause other useful implied objects to be mercilessly evicted from the cache.

4 Conclusions

In this paper, we have implemented a novel integrated Web caching and prefetching policy called WCP-CMA that makes use of Cyclic Model Analysis of the Web access sequences for deriving prefetching rules. Our experimental results have shown a 10-15% increase in the hit ratios of the cached objects compared to IWCP. Also, the execution time of WCP-CMA is significantly lower than that of IWCP and WCP-CMA saves more network and the proxy’s local resources. Hence, it is obvious that our approach is more efficient than IWCP. We have taken only small values of periodicity \( P \) and cyclic behaviour \( C \) for our implementation. Further, WCP-CMA algorithm can be simulated against to other parameters like Delay, execution Time and effects of different values of Periodicity and Cyclic Behaviour.

References

2. Ding-An, Chiang., Cheng-Tzu, Wang., Shao-Ping, Chen., Chun-Chi, Chen.: The Cyclic Model Analysis on Sequential Patterns. IEEE Transactions on Knowledge and Data Engineering, 21(11), 1617--1628 (2009)