ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Long-Term Web Prefetching Algorithms: A Comparison

C.Anuradha, Sundararajan.M, Arulselvi S

Assistant Professor, Dept. of CSE, Bharath University, Chennai, Tamil Nadu, India

Director, Research Center for Computing and Communication, Bharath University, Chennai, Tamil Nadu, India

Co-Director, Research Center for Computing and Communication, Bharath University, Tamil Nadu, India

ABSTRACT: User perceived latency has become a potential problem due to the increase in internet traffic. Web caching is an effective means of reducing user perceived latency. Web prefetching is an attractive solution which relies on web caching to reduce access latency. There are two kinds of algorithms that are currently used for prefetching i.e., linear algorithms and data mining algorithms. Web prefetching reduces user access time, but it requires more bandwidth and an increase in the network traffic. Performance measurement of prefetching techniques is primarily in terms of hit ratio and bandwidth usage. The important factor in deciding a prefetching algorithm is its ability to choose objects to be fetched in advanced. In this paper, a detailed analysis of various existing prefetching algorithms have been proposed.

KEYWORDS: Hit rate, Prefetching, bandwidth consumption, H/B metric, APL characteristic.

I. INTRODUCTION

Web caches are widely used in the current Internet environment to reduce the user-perceived latency of object requests. One example is a proxy server that intercepts the requests from the clients and serves the clients with the requested objects if it has the objects stored in it; if the proxy server does not have the requested objects, it then fetches those objects from the web server and caches them, and serves the clients from its cache. Another example is local caching that is implemented in web browsers.One way to further increase the cache hit ratio is to anticipate future requests and

prefetch these objects into a local cache.[1] On the other hand, prefetching consumes more network bandwidth. Venkataramani et al. mentioned that system resources in terms of bandwidth usage and response time in terms of hit ratio are not directly comparable quantities. In this paper, by considering that network bandwidth usage is cost and hit ratio is the performance of prefetching, we introduce a concept which defines our notion of the value of prefetching by measuring the balance between the response time improvement and the extra amount of system resources consumed by prefetching compared to demand caches. The problem with prefetching is it will increase network traffic.[2-3] It will be useful only if the prefetching is done at idle time i.e., when the network traffic is very low. Intuitively, to increase the hit rate, we want to prefetch those objects that are accessed most frequently; to minimize the bandwidth consumption, we want to choose those objects with longer update intervals. We assume unlimited cache sizes for both on-demand and prefetching cases in this paper. Various algorithms have been currently used for prefetching web objects which are well accepted. In this paper, we are going to discuss what are the various issues involved in choosing a prefetching algorithm. Linear prefetching algorithms which are well known are Popularity, Good Fetch, APL characteristic, Lifetime, H/B greedy, Hit rate greedy, bandwidth greedy and H/B optimal [3]. This paper also discusses some of the data mining algorithms which rely on the anchor texts. Their performance can be measured using the different criteria discussed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

II. EXISTING PREFETCHING ALGORITHMS

2.1. Top 10 Approach

Markatos et al. [4] suggested a "Top Ten" criterion for prefetching web objects, which keeps in cache the ten most popular objects from each web server. Each server keeps records of accesses to all objects it holds, and the top ten popular objects are pushed into each cache whenever they are updated. Thus, those top-ten objects are kept "fresh" in all caches. A slight variance of the "Top Ten" approach is to prefetch the most popular objects from the entire system. Since popular objects account for more requests than less popular ones, Prefetch by Popularity is expected to achieve the highest hit rate [5].

2.2. Domain Top Approach

Seung Won Shin et al. proposes a domain top approach for web prefetching, which combines the proxy's active knowledge of most popular domains and documents.In this approach proxy is responsible for calculating the most popular domains and most popular documents in those domains, then prepares a rank list for prefetching.

2.3. A Keyword based semantic prefetching approach in internet news services

This proposes a key word based sematic prefetching ,in which prediction of future requests are based on semantic references of past retrieved web documents.This technique is applied to internet news services , it finds out sematic preferences by analyzing keywords in URL anchortext of previously accessed documents in different news categories. It assumes a proxy is running behind the web browser keeps track of clients characteristics and find out semantic relation between the documents[6].

2.4 . Semantic Link Prefetcher algorithms

A Semantic Link Prefetcher was proposed to utilize the
semantic link information associated with the current Web page hyperlinks to predict the Web objects to be prefetched during the limited view time interval of the current Web page. A transparent and speculative algorithm was proposed for content based Web page prefetching with the assumption that textual information in both the visited pages and the followed links were influential in determining the preferences of a user. A novel non-intrusive Web prefetching system has been proposed to avoid the interference between prefetch and demand requests by effectively utilizing only the spare resources on the servers and network. The system was deployed without making any modifications to the Web browser, HTTP protocol and the network. In another paper, a client-based Web prefetching system was proposed that used detection theory to determine the threshold value for selecting the Web documents to be prefetched. [8]

2.5. PPM Models

PPM models were commonly used in Web prefetching for predicting the user's next request by extracting useful knowledge from historical user requests. Factors such as page
access frequency, prediction feedback, context length and conditional probability influence the performance of PPM models in prefetching. An online PPM model based on non compact suffix tree was implemented in [9] that used maximum entropy principle to improve the prefetching performance. A novel PPM model based on stochastic gradient descent was proposed that defined a target function to describe a node‟s prediction capability and then selected a node with maximum function value to predict the next most probable page.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

2.6. Markov Models

Markov models were effectively used in Web prefetching by utilizing the information gathered from Web logs. In [13] different techniques were presented for intelligently selecting the parts of different order Markov models to create a new model with reduced state complexity and improved prediction accuracy[11][12]. Three schemes of pruning (support, confidence and error) were presented to prune the states of All-K th order markov model. A Markov–Knapsack approach was proposed that combined Multi-Markov Web-application centric

2.7. Prefetch by Lifetime

Prefetching objects leads to extra bandwidth consumption, since in order to keep a prefetched object "fresh" in the cache, it is downloaded from the web server whenever the object is updated. Starting from the point of view of bandwidth consumption, it is natural to choose those objects that are less frequently updated. Prefetch by Lifetime [7], as its name indicates, selects m objects that have the longest lifetime, and thus intends to minimize the extra bandwidth consumption.

2.8. Good Fetch

Venkataramani et al. [7] proposed a Good Fetch criterion that balances the access frequency and update frequency (or lifetime) of web objects. In the Good Fetch algorithm, the objects that have the highest probability of being accessed during their average lifetime are selected for prefetching. Assuming the overall object access rate to be a, the frequency of access to object i to be pi , and the average lifetime of this object to be li, the probability that object i is accessed during its lifetime can be expressed as

$$P_{goodfetch} = 1 - (1 - p_i)^{ali}$$

The Good Fetch algorithm prefetches a collection of objects whose $P_{goodfetch}$ exceeds a certain threshold. The intuition behind this criterion is that objects with relatively higher access frequencies and longer update intervals are more likely to be prefetched, and this algorithm tends to balance the hit rate and bandwidth in that it increases the hit rate with a moderate increasing of bandwidth usage. Venkataramani et al. [10] argued that this algorithm is optimal to within a constant factor of approximation. However, it could behave inefficiently under some specific access-update patterns.

It is observed that when an object is prefetched, the hit rate is increased from $ap_i l_i /(ap_i l_i + 1)$ to 1 which is 1/ f(i) times the bandwidth of object under on-demand caching. Prefetching an object leads to the same relative increase on its hit rate and
bandwidth consumption. A prefetching algorithm uses the hit rate and bandwidth of on-demand caching as a baseline for comparison and as they are constants, the H/B metric is equivalent to:

$$(\frac{H}{B})_{pref} = \frac{Hit_{pref}}{BW_{pref}} = \frac{\sum_i p_i h_i}{\sum_i b_i}$$

Where h_i and b_I are hit rate and bandwidth of object , respectively.
Consider the H/B value of on-demand caching:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

$$(\frac{H}{B})_{demand} = \frac{Hit_{demand}}{BW_{demand}} = \frac{\sum_i p_i f(i)}{\sum_i \frac{s_i}{l_i} f(i)}$$

What a prefetching algorithm does is choose an appropriate
subset of objects from the entire collection – for each of those prefetched objects, say object i, we simply change the corresponding f(i) term to 1 in equation (2). Our H/BGreedy algorithm aims to select a group of objects to be prefetched, such that the metric would achieve a better value than that obtained by any other algorithm. Since the object characteristics such as access frequencies, lifetimes, and object sizes are all known to the algorithm, it is possible that, given a number n, we could select n objects to prefetch such that H/B reaches the maximum value for the given number ✳ and this maximum value will be greater than the value obtained using any existing algorithm. This optimization problem can be formalized as finding a subset of size n from the entire collection of objects, such that H/B is maximized[14][15].

2.10 Hit Rate-Greedy prefetching

Sometimes it is desirable to maximize the overall hit rate given the number of objects to prefetch, Jiang et al. claimed that Prefetch by Popularity achieves the highest possible hit rate. However a special form of our Objective-Greedy algorithms would actually obtain higher hit rate than Prefetch by Popularity.

$$H\_contr(i) = p_i(1 - f(i)) = \frac{p_i}{ap_i l_i + 1}$$

2.11 Bandwidth-Greedy prefetching

Another optimization problem in prefetching is to minimize the excessive bandwidth consumption, given the number of objects to prefetch[16]. Intuition may suggest that Prefetch by Lifetime has the least bandwidth usage . However, by applying our Objective-Greedy principle with bandwidth as the objective, we get an algorithm that results in even less bandwidth consumption. Using analogous reasoning to that for the Hit Rate-Greedy algorithm, the extra bandwidth contributed by prefetching object I is:

$$B\_contr(i) = \frac{s_i}{l_i}(1 - f(i)) = \frac{s_i}{ap_i l_i^2 + l_i}$$

III. H/B-OPTIMAL PREFETCHING

The optimal prefetching approach that achieves the highest H/B value given the number of objects to prefetch. The goal is to select a subset of m objects to prefetch, such that the
objective H/B is maximized[15].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

H/B-Optimal Prefetching	Meaning
i, j	Object
$p_i(1 - f(i))$	Hit rate contribution by prefetching i
$\frac{s_i}{l_i}(1 - f(i))$	Bandwidth contribution by prefetching i
$\sum_{i \in S} p_i f(i)$	On-demand hit rate
$\sum_{i \in S} \frac{s_i}{l_i} f(i)$	On-demand bandwidth
S	Entire set of objects

IV. CONCLUSION

Simulations have shown that H/B-Greedy prefetching beats Prefetch by Popularity, Good fetch, Prefetch by APL, and Prefetch by Lifetime in terms of H/B for any number of prefetched objects, as expected. This is because the object selection in H/B-Greedy prefetching is always guided by the objective H/B. The gain by H/B-Greedy over Good Fetch (APL) prefetching is more significant when the number of prefetched objects is relatively small compared to the total number of objects. H/B-Greedy, and H/B-Optimal initially ascend to some highest values as the number of prefetched objects increases, and then gradually descend and converge with other curves. The reason for this scenario is that when the prefetched number is relatively small, the algorithms that use balanced metrics (Good Fetch (APL), H/B-Greedy, and H/B-Optimal) have better chances to choose those objects that contribute the largest goodness for the balanced metric. Specifically, Good Fetch (APL), H/B-Greedy, and H/B-Optimal have better chances to choose those objects that have the largest values of , the largest values of increase factor, and the best contribution to improving respectively[17-20]

REFERENCES

1. E.P. Markatos, C.E. Chironaki, A Top 10 Approach for Prefetching the Web, Proc. INET'98:Internet Global Summit, July 1998.
2. Udayakumar R., Khanaa V., Kaliyamurthie K.P., "High data rate for coherent optical wired communication using DSP", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) 4772-4776.
3. Y. Jiang, M.Y. Wu, W. Shu, Web Prefetching: Cost, Benefits and Performance, 11th World Wide Web Conference (WWW), 2002.
4. Selva Kumar S., Ram Krishna Rao M., Deepak Kumar R., Panwar S., Prasad C.S., "Biocontrol by plant growth promoting rhizobacteria against black scurf and stem canker disease of potato caused by Rhizoctonia solani", Archives of Phytopathology and Plant Protection, ISSN : 0323-5408, 46(4) (2013) pp.487-502.
5. Bin Wu , Ajay D. Kshemkalyani, Objective-Optimal Algorithms for Long-term Web Prefetching, Computers, IEEE transactions, Jan 2006.
6. Udayakumar R., Khanaa V., Kaliyamurthie K.P., "Optical ring architecture performance evaluation using ordinary receiver", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp. 4742-4747.
7. A.Geetha, Universal Asynchronous Receiver / Transmitter (UART) Design for Hand Held Mobile Devices, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN: 2231-5381, pp 25-26, Volume 3 Issue 1 No1 – January 2012.
8. Zhijie Ban, Zhimin Gu, and Yu Jin, "An online ppmprediction model for web prefetching", In proceedings of the 9th annual ACM international workshop on Web information and data management, Lisbon, Portugal, 2007.
9. Subha Palaneeswari M., Abraham Sam Rajan P.M., Silambanan S., Jothimalar, "Blood lead in end-stage renal disease (ESRD) patients who were on maintainence haemodialysis", Journal of Clinical and Diagnostic Research, ISSN : 0973 - 709X, 6(10) (2012) pp.1633-1635.
10. B.Vamsi Krishna, A New Technique for Elimination of Harmonics Using Three Phase Shunt Active Filter, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN (Online): 2278 – 8875,pp 5596-5602, Vol. 2, Issue 11, November 2013
11. Mukund Deshpande and George Karypis, "SelectiveMarkov Models for Predicting Web Page Accesses", ACM Transactions on Internet Technology, Vol. 4, No. 2, pp.163-184, May 2004
12. Udayakumar R., Khanaa V., Kaliyamurthie K.P., "Performance analysis of resilient ftth architecture with protection mechanism", Indian Journal of Science and Technology, ISSN : 0974-6846, 6(S6) (2013) pp. 4737-4741
13. Alexander P. Pons, "Improving the performance ofclient web object retrieval", Journal of Systems and Software, vol.74, No.3, 2005.

   a. *Venkataramani, P. Yalagandula, R. Kokku, S. Sharif, M. Dahlin, The Potential Costs and Benefits of Long-term Prefetching, Computer Communications, 25(4):367-375, 2002.*

14. M. Crovella, P. Barford, The Network Effects of Prefetching, Proc. IEEE Infocom, pp.12321239, 1998.
15. P. Deolasee, A. Katkar, A. Panchbudhe, K. Ramamritham, P. J. Shenoy, Adaptive Push-pull: Disseminating Dynamic Web Data, Proc. WWW Conference 2001, pp. 265-274, 2001.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

16. F. Douglis, A. Feldmann, B. Krishnamurthy, J. C. Mogul, Rate of Change and Other Metrics: A Live Study of the World Wide Web, Proc. USENIX Symp. on Internet Technologies and Systems, pp. 147-158, 1997.
17. A.Geetha, Face Recognition Using OPENCL, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering , ISSN (Print) : 2320 – 3765, pp- 7148-7151, Vol. 3, Issue 2, Febuary 2014.
18. B.Mehala, Mrs.Anitha Sampath Kumar, Design and Implementation of Resonant Circuit Based On Half-Bridge Boost Rectifier with Output Voltage Balance Control, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN (Online): 2278 – 8875,pp 9370-9378, Vol. 3, Issue 5, May 2014
19. B.Vamsi Krishna, Starting Inrush Current Control of Three-Phase Induction Motors for Dispersed Generating Systems, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN (Online): 2278 – 8875,pp 6411-6422, Vol. 2, Issue 12, December 2013