ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

A Survey on Intelligent Web Caching Technique using Data Mining

Swathikaa RK
Research scholar
PSGR Krishnammal College for Women
Coimbatore, India
swathikaa8@gmail.com

Vijaya MS
Associate Professor
PSGR Krishnammal College for Women
Coimbatore, India
msvijaya@psgrkc.com

***Abstract*—Web caching plays a very important role for improving the performance of many Web-Based systems. This paper presents some approaches for improving the efficiency of web cache memory by assessing the usage patterns of users. Based on the web log file, the pages that are frequently used together are mined. The data structures are used to store the pages in cache memory as a tree. This is aimed at improving hit ratio. Bayesian algorithm is combined with page replacement algorithms to determine the replacement decision when there is a miss during a search. The data structure that results in maximum hit ratio and minimum web latency is to be determined so that the cache memory is used efficiently. Page Replacement Technique is used along with Bayesian techniques. The results obtained from each combination are compared to find out the efficient data structure.**

***Keywords*— web caching, web log mining, Bayesian techniques**

I. INTRODUCTION

Web server plays an important role between users and web sites in reducing the response time of user requests and saving the network bandwidth. In the web server, some frequently used web objects, which are likely to be revisited, are stored on web cache memory for retrieving them later shortly. Thus, web caching is one of the most successful solutions for improving the performance of web-based systems. However, the difficulty in determining the web objects that will be re-visited in the future is a problem faced by existing conventional web caching techniques. In this paper, association rule mining and Bayesian network technique is used to enhance the performance of conventional web caching approaches such as Least-Recently-Used (LRU) and Greedy-Dual-Size (GDS). BN can be intelligently incorporated with conventional Web caching techniques to form intelligent and effective caching approaches.

II. STORING WEB OBJECTS USING DATA STRUCTURES

The Web log file is mined using data mining tool to determine the usage patterns of users. Tree structures are

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

formed with frequently used pages occupying the root and top nodes. The pages that are accessed less frequently occupy bottom nodes or root nodes.

Web latency is determined for linear, block-wise search as it is done in the cache memory. Also, latency is determined by employing tree structures for the web log entries. The latency is noted in nanoseconds so as to visualize minute difference in the performances.

The data structure that takes minimum time to fetch an entry is considered to be efficient and when a miss occurs, page is to be replaced. Page replacement is done by considering the intelligent approaches to calculate the aging values of each page.

III. USING DATA MINING APPROACHES

The web log file is mined using association algorithms to determine the usage patterns of the users. Based on the patterns, the web objects are stored in data structures. Bayesian approaches in data mining are used for calculating age values to predict which page will likely be replaced. Following are the approaches to be used.

A. *Bayesian Network-Greedy-Dual-Size Approach (BN-GDS)*

The proposed policy integrates BN classifier with GDS for improving the performance in terms of the byte hit ratio of GDS, and so called BN-GDS. In the proposed BN-GDS, GDS is enhanced by incorporating the accumulative scores or probabilities $W(g)$ that object g will be revisited in the future depending on BN classifier as shown in following equation:

$$K(g) = L + W(g) * (C(g) / S(g))$$

Instead of object frequency, the probabilities or membership scores of belonging to the class with objects may be revisited are accumulated and incorporated into caching priority of Web object. Hence, the scores predicted by BN can contribute effectively in improving caching priority, compared with the priority with just frequency factor.

B. *Bayesian Network-Least-Recently-Used Approach (BN-LRU)*

LRU policy is the most common proxy caching policy among all the web caching algorithms. However, LRU policy suffers from cold cache pollution, which means unpopular objects remain in the cache for a long time. In other words, in LRU, a new object is inserted at the top of the cache stack. If the object is not requested again, it will take some time to be moved down to the bottom of the stack before removing it from the cache. For reducing the cache pollution in LRU, BN classifier is combined with LRU to form a new algorithm called BN-LRU.

The proposed BN-LRU works as follows. When the Web object g is requested by user, BN predicts the class of that object either be revisited again or not. If the object g is classified by BN as object to be revisited again, the object g will be placed on the top of the cache stack. Otherwise, the object g will be placed in the middle of the cache stack.

Hence, BN-LRU can efficiently remove the unwanted objects early to make space for the new Web

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 7, January 2016

objects. By using this mechanism, the cache pollution can be reduced and the available cache space can be utilized.

C. *Bayesian Network-Dynamic Aging Approach(BN-DA)*

When user visits Web object g, the trained BN classifier can predict the probability of belonging g to the class with objects may be revisited. Then, the probabilities of g are accumulated as scores $W(g)$ used in cache replacement decision as show in the equation:

$$K(g) = L + W(g)$$

The object with lower scores is a better choice for replacement. In the proposed BN-DA policy, L is a dynamic aging (DA) factor to prevent cache pollution and improve the performance in implementation for longer periods of time.

IV. DISCUSSION

BN-GDS was able to make better balance between HR and BHR than other algorithms. On the other hand, BN-LRU and BN-DA achieved the best BHR among all algorithms, and better HR compared LRU and NNPCR-2 in most proxy datasets. Satisfactory experiment results are achieved by the efficiency of integration which uses web cache prediction method and web cache replacement algorithm together in order to get better performance. Access latency is reduced while using SVM pre-fetching because of increase in percentage of cache hits. Implementing neuro-fuzzy networks can significantly improve performance.

V. CONCLUSION

Web caching is an important technology which can improve content availability, reduce network latencies and address increasing bandwidth demands. Unfortunately, the cache hit ratio is not improved much with caching schemes. Despite with a cache of infinite size, the hit ratio are still limited only at the range from 40% to about 50%, regardless of the caching scheme. The decision-tree strategy outperforms all best fixed strategies. In this paper, some approaches are discussed and the combination of which, better results can be achieved.

REFERENCES

[1] Qiang Yang and Haining Henry Zhang, Web-Log Mining for Predictive Web Caching, IEEE transactions on knowledge and data engineering, Volume 15, Issue 4, Page: 1050-1053 (2003).

[2] W. Ali, and S. Shamsuddin, *Intelligent Client-Side Web Caching Scheme Based on LRU Algorithm and Neuro-Fuzzy System*, Advances in Neural Networks, Page: 70-79 (2009).

[3] Francesco Bonchi, Fosca Giannotti, Giuseppe Manco and Chiara Renso, *Data Mining for Intelligent Web Caching,* CNUCE-CNR – Institute of Italian National Research Council.

[4] R. Kosala and H. Blockeel., *Web mining research: A Survey*, ACM SIGKDD Explorations, Volume 2, Issue 1, Page: 1–15 (2000).