ISSN 2047-3338

A Survey on Web Cache Replacement Technology

Mohammed Salah Abdalaziz Khaleel[1], Saif Eldin Fattoh Osman[2] and Hiba Ali Nasir Sirour[3]

[1]Faculty of Computer Studies, International University of Africa, Khartoum, Sudan

[2,3]Emirates College of Technology, Faculty of Computer, Studies IUA, Khartoum, Sudan

[1]mohammedkh33@hotmail.com, [2]saifefatoh@hotmail.com, [3]hibaalinasir@hotmail.com

***Abstract*— Loading a document, an image, a voice recording, or a video on the Internet can cause jamming signals and overload on the network traffic and performance. These happen as a result of the speedy growth of the browsing web site pages. Web caching policies are popular techniques that have played a key role in improving performance by increasing the hit ratio. There are more web cache policies that are used to solve the problems and increase the performance of the internet like web cache replacement techniques. This paper highlights some of the policies and compares them, using the hit ratio, performance matrix, latency time and byte hit ratio.**

***Index Terms*— Intelligent Caching, Proxy Cache, Removable Policies, Web Caching Performance, Hit Ratio, Byte Hite Ratio and Latency Time**

I. INTRODUCTION

THE rapid development of the Internet needs to be met by improving the techniques to increase the performance of browsing and decrease the overload on the network traffic. These need depend on the cache memory in any site like (client, proxy, and original server).

To overcome this situation, web caching techniques are used. Web cache reduces the high traffic over the internet so that users can access the web content faster. The main purpose of cache is to place the copy of an object near to the client, so that the web user can access the object easily, without the request going back to the proxy or original web server. This reduces the overhead on the network while increasing the availability of the object.

There are different points where a cache can be set up browser, proxy server or close to server. When a user request a web page, first it is checked in the cache, if the requested web page is available then it is sent back to the user. If the web page is not found in the cache, then the request is redirected to the web server, which sends the response to the client between cache store and the requested web page .Because of the limited memory size of the cache, it is impossible to save all objects in the memory.

Most proxy caching use the Least Frequently Used (LFU) method that is deletes the least used object to free up space. Other methods include the Size and Greedy Dual Size (GDS where objects with large sizes are deleted, and the Least Recently Used (LRU) where objects have not been used recently are deleted. However, all these techniques are not efficient in Proxy caching, since they only consider just one factor at a time and ignoring other factors

This paper tries to describe benefits of these techniques and compare them.

II. WEB CACHE TECHNIQUES

Many web cache techniques are used to increase the availability of web sites and decrease network traffic that directly increases the cache performance. There are several algorithms used for web cache replacement.

These algorithms are divided into two classes, frame work algorithms (immeasurable) and application studies (measurable).

Framework algorithms:

A) Integration of Least Recently Used Algorithm and Neuro-Fuzzy System into Client-side Web Caching

A reference [1] proposes splitting client-side web cache to two caches, short-term and long-term. Initially, a web object is stored in short-term cache, and the web objects that are visited more than the pre-specified threshold value will be moved to long-term cache. Other objects are removed by Least Recently Used (LRU) algorithm as short-term cache becomes full. More significantly, when the long-term cache saturates, a neuro-fuzzy system is employed in classifying each object stored in it into either cacheable or uncatchable. The old uncatchable objects are candidates for removal from the long-term cache. By implementing this mechanism, cache pollution can be mitigated and cache space can be utilized effectively. Experimental results have revealed that the proposed approach can improve the performance up to 14.8% and 17.9% in terms of hit ratio (HR) compared to LRU and

Least Frequently Used (LFU). In terms of byte hit ratio (BHR), the performance is improved up to 2.57% and 26.25%, and for latency saving ratio (LSR), the performance is improved up to 8.3% and 18.9%, compared to LRU and LFU. Although the simulation results have proven that work helps in improving the performance in terms of the hit ratio (HR), the performance in terms of the byte hit ratio (BHR) is not good enough since the cost and size of the predicted objects in the cache replacement process were not taken into consideration. Moreover, the training process requires long time and extra computational overhead.

B) Latency Reduction in mobile environment

In reference [2] an integrated caching and pre-fetching technique to reduce latency in mobile environment. The proposed model consist of bandwidth monitoring agent to find out current bandwidth usage, a pre-diction module to predict the number and the list of rules to be pre-fetched and a pre-fetch module to pre-fetch the web page and store them in a pre-fetch area. Simulation results show that the browser implemented in a mobile environment maintains almost constant web traffic even id pre-fetching is done and latency is reduced up to 40- 70%.

C) Performance Improvement of Least-Recently-Used Policy in Web Proxy Cache Replacement Using Supervised Machine Learning

Reference[3]enhance traditional Web caching polices using supervised machine learning techniques such as a support vector machine, a naïve Bayesian classifier (NB), a decision tree (C4.5) and Size . It trained from Web proxy logs files to predict the object that would be revisited.HR increased by 30.15% and BHR increased by 32.43%.

D) Traing and Simulation of NeuralNet Works for Web Proxy Cache Replacement

In [4] NNW are trained to classify cacheable objects from real world data sets using information known to be important in web proxy caching, such as frequency, recency and size. In simulation, the final NN achieve HR that are 86.60% of the optimal in the worst case and 100% of the optimal in the best case. BHR are 93.36% of the optimal in the worst case and 99.92% of the optimal in the best case.

E) A Novel Cooperative Caching Scheme for Wireless Ad Hoc Networks: Group Caching

Study [5] proposed a novel cooperative caching scheme called Group Caching (GC) which allows each mobile host and its 1-hop neighbors form a group. The caching status is exchanged and maintained periodically in a group. By using the proposed Group Caching, the caching space in mobile hosts can be efficiently utilized and thus the redundancy of cached data is decreased and the average access latency is reduced. The authors evaluate the performance of the Group Caching by using NS2 and compare it with the existing schemes such as Cache Data and Zone Cooperative. The experimental results show that the cache hit ratio is increased by about 3%~30% and the average latency is reduced by about 5%~25% compared with other schemes.

F) Improving Performance of World Wide Web by Adaptive Web Traffic Reduction

In [6] an adaptive hybrid algorithm has been developed for reducing web traffic. Intelligent agents are used for monitoring the web traffic. Depending upon the bandwidth usage, user's preferences, server and browser capabilities, intelligent agents use the best techniques to achieve maximum traffic reduction. Web caching, compression, filtering, optimization of HTML tags, and traffic dispersion are incorporated into this adaptive selection. Using this new hybrid technique, latency is reduced to 20 – 60 % and cache hit ratio is increased 40 – 82 %.

Measurable algorithm:

A) Distributed Web caching: A Dynamic clustering Approach

Study [7] gave a solution for scalability and robustness of Distributed web caching System and for load balancing and Metadata manageability. The study had refined the technique using proxy server clusters with Dynamic allocation of requests. They devised an algorithm for Distributed Web Cache concepts with clusters of proxy server based on geographical regions. It increases the scalability by maintaining metadata of neighbors collectively and balances load of proxy servers dynamically to other less congested proxy servers, so system doesn't get down unless all proxy servers are fully loaded and higher robustness of system is achieved. This algorithm also guarantees data consistency between the original server object and the proxy cache objects using semaphore.

B) An Integrated Model for Next Page Access Prediction

Paper [8] provides an improved prediction accuracy and state space complexity by using novel approaches that combine clustering, association rules and Markov models. The three techniques are integrated together to maximize their strengths. The integration model has been shown to achieve better prediction accuracy than individual and other integrated models.

C) Review LRU Algorithm to Implement Proxy Server with Caching Policies

In [9] a technique to remove the problem of cold cache pollution is proposed which is proved mathematically that it is better than the existing.

D) A Least Grade Page Replacement Algorithm for Web Cache Optimization

A new algorithm called least Grade Replacement (LGR) is proposed in [10] by considering recency, frequency, perfect-history and size in replacing policy. The 2-way and 4-way set associative caches were used to determine the optimal recency coefficients. The cache size was varied from 32k to 256k in the simulation. The results showed that the new algorithm (LGR) is better than LRU and LFU in terms of Hit Ratio (HR) and Byte Hit Ratio (BHR).

E) SEMALRU: An Implementation of modified web cache replacement algorithm

Reference [11] proposed a SEMALRU replacement policy by combining the semantic content and recency of web pages. It outperformed other policies in terms of Page Hit Ratio, Byte Hit Ratio and number of replacement as demonstrated in the text. The policy was tested in a simulated environment with the related and unrelated set of user access pattern. The parameters pertinent to cache replacement algorithms are computed and the results showing the improvement in the efficiency of the algorithm are furnished.

F) Web Proxy Caching Object Replacement: Frontier Analysis to Discover the "Good-Enough" Algorithms

Reference [12] proposed and used the data envelopment analysis (DEA) as a technique that can be used to enhance the trace-driven simulation experiments that constitute the common methodology to study the object replacement strategies in web caching. The DEA model clearly showed that the cache size plays a crucial role in improving the performance of all the algorithms, for all the performance metrics under study. When the cache size increases, there is a general convergence of the efficiency scores towards the unity.

G) A scheme for adaptive web caching based on multi-level object classification

Reference [13] proposed a caching scheme which utilized multi-level class information. A MLR (Multinomial Logistics Regression) based classifier is constructed using the information from web logs. Simulation results confirm that the model has good prediction capability and suggest that the proposed approach can improve the performance of the cache substantially.

H)Approximation approach to performance evaluation of Proxy Cache Server systems

A modification of the performance model of Proxy Cache Servers to a more powerful case when the inter-arrival times and the service times are generally distributed was proposed in [14]. The paper described the original proxy cache server model where the arrival process is a Poisson process and the service times are supposed to be exponentially distributed random variables. Then they calculate the basic performance parameters of the modified performance model using the well-known Queuing Network Analysis (QNA) approximation method. The accuracy of the new model is validated by means of a simulation study over an extended range of test cases.

I) Impact of One-Timer/N-Timer Object Classification on the Performance of Web Cache Replacement Algorithms

The work [15] presented a technique to classify whether a cached object is a One-Timers (OT) referenced only once or not. Statistical analysis of the workload shows that as much as 76% of objects are One-Timers (OT), Caching OT objects usually degrade the performance of all Web cache replacement algorithms .Simulation shows that classification may significantly enhance the performance of replacement algorithms with respect to the HR, the BHR and the DSR.

J) The effect of caching on a model of content and access provider revenues in information-centric networks

A game between an Internet Service Provider (ISP) and content provider (CP) on a platform of end-user demand was considered in [16]. A price-convex demand-response is motivated based on the delay-sensitive applications that are expected to be subjected to the assumed usage-priced priority service over best-effort service. The authors considered two-sided market with multi-class demand wherein one class (that under consideration herein) is delay-sensitive. Both the Internet and proposed Information Centric Network, encompassing Content Centric

III. APPLICATIONS STUDIES

There are standard performance metrics to evaluate the Performance of Web caching techniques. Hit Ratio (HR), Byte Hit Ratio (BHR). These can calculate as follows:

$$HR = \frac{\sum_{i=1}^{n} \partial i}{n} \quad (1)$$

$$BHR = \frac{\sum_{i=1}^{n} b_i \partial i}{\sum_{i=1}^{n} b_i} \quad (2)$$

n: total Number of requests

∂i: 1 if the request i is in the cache

∂i: 0 otherwise

bi: size in bytes

IV. RESULTS DISCUSSION

The following figures present a discussion of the study results in terms of Latency Time, Hit Ratio (HR) and Byte Hit Ratio (BHR).

Fig. [1] describes the comparison between *Algorithms* depend on Latency time in Integration of Least Recently Used Algorithm and Neuro-Fuzzy System into Client-side Web Caching [1] study the latency time percentage 18.90%, in Latency Reduction which in mobile environment[2] study the latency time percentage 70%, in A Novel Cooperative Caching Scheme for Wireless Ad Hoc Networks: Group Caching [5] study the latency time percentage 25% and in Improving Performance of World Wide Web by Adaptive Web Traffic Reduction [6] study the latency time percentage 60%.

Consequently, the best of studies is the lowest latency time percentage is study [1] and the worst study with the highest latency time percentage is study [2].

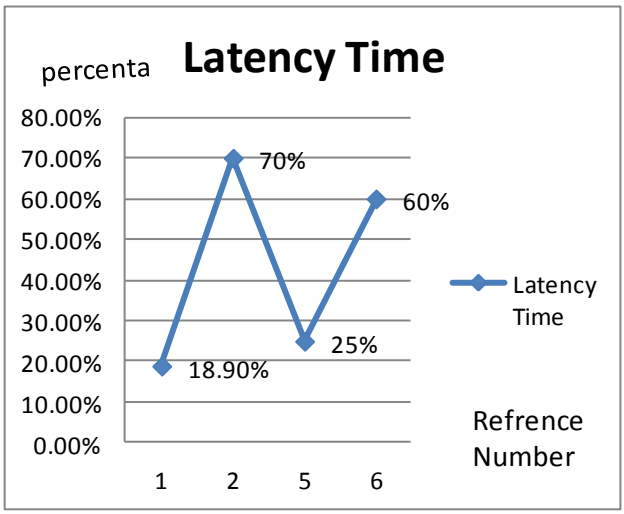

Fig. 1: Comparison depend on Latency time

Fig. 2 describe the comparison between studies using Hit ration, in Integration of Least Recently Used Algorithm and Neuro-Fuzzy System into Client-side Web Caching [1] study the Hit Ration percentage 17.9%, in Performance Improvement of Least-Recently-Used Policy in Web Proxy Cache Replacement Using Supervised Machine Learning [3] study the Hit Ration percentage 30%, in Traing and Simulation of Neural Net Works for Web Proxy Cache Replacement [4] study the Hit Ration percentage 100%,in A Novel Cooperative Caching Scheme for Wireless Ad Hoc Networks: Group Caching [5] study the Hit Ration percentage 30% and in Improving Performance of World Wide Web by Adaptive Web Traffic Reduction [6] study the Hit Ration percentage 82%.

Consequently, the best of studies is the highest Hit Ration Percentage study [4] and the worst study with the lowest Hit Ration percentage both study [5] and [3].

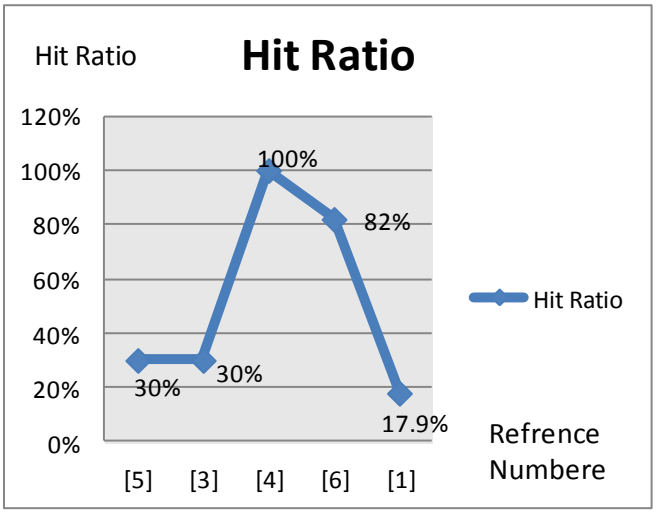

Fig. 2: Comparison depend on Hit Ration

Fig. 3 describe the comparison between studies using Byte Hit Ration, in Integration of Least Recently Used Algorithm and Neuro-Fuzzy System into Client-side Web Caching [1] study the Byte Hit Ration percentage26.25%, in Performance Improvement of Least-Recently-Used Policy in Web Proxy Cache Replacement Using Supervised Machine Learning [3] study the Byte Hit Ration percentage 32.43% and in Traing and Simulation of NeuralNet Works for Web Proxy Cache Replacement [4] study the Byte Hit Ration percentage99.92%. Consequently, the best of studies is the highest Hit Ration Percentage study [4] and the worst study with the lowest Hit Ration percentage both study [1].

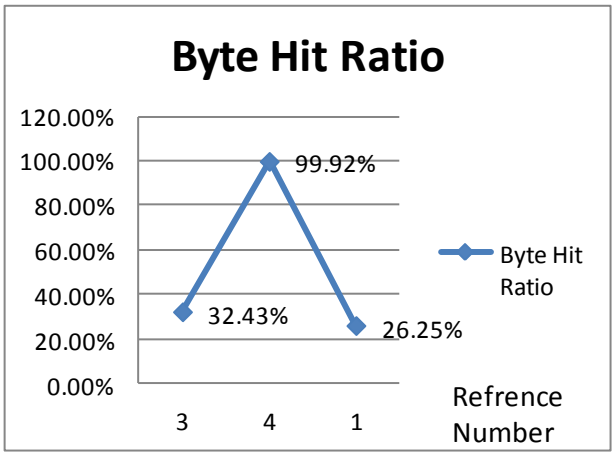

Fig. 3: Comparison depend on Byte Hit Ration

V. CONCLUSION

In this survey of intelligent web caching replacement, many studies have shown that the intelligent approaches are more efficient and adaptive in Web caching environment compared to other technologies.

According to the discussions in section IV, Training and Simulation of Neural Networks for Web Proxy Cache Replacement [4] is the best with the highest Hit Ratio and Byte Hit Ratio.

REFERENCES

[1] S.M.S.m. Waleed Ali prowalid, "Integration of Least Recently Used Algorithm and Neuro-Fuzzy System into Client-side Web Caching.

[2] J.S.Greshma, "Latency Reduction in mobile environment," Elsevier, 2012.

[3] S.S. Waleed Ali, Norbahiah Ahmad, "Performance Improvement of Least-Recently-Used Policy in Web Proxy Cache Replacement Using Supervised Machine Learning," Int. J. Advance, vol. 6, 2014.

[4] H.E. Jake Cobb, "Traing and Simulation of NeuralNet Works for Web Proxy Cache Replacement," 2006.

[5] Y.W.T.a.Y.K. Chang, "A Novel Cooperative Caching Scheme for Wireless Ad Hoc Networks: GroupCaching", 2007

[6] "Improving Performance of World Wide Web by Adaptive Web Traffic Reduction," 2006.

[7] G.K. RAJEEV TIWARI, LALIT GARG, "Robust Distributed Web caching: A Dynamic clustering Approach," vol. 3 No.2, 2011.

[8] J.L. F. Khalil, H. Wang, "An Integrated Model for Next Page Access Prediction", 2009.

[9] J.K.G. JITENDRA SINGH KUSHWAH, BRIJESH PATEL#3, "Review LRU Algorithm to Implement Proxy Server with Caching Policies," International Journal of Engineering Science and Technology (IJEST), vol. 3 No.10, 2011.

[10] B.H. Naizheng, Chen, "A Least Grade Page Replacement Algorithm for Web Cache Optimization," Knowledge Discovery and Data Mining, 2008. WKDD 2008. First International Workshop on, pp. 469-472.

[11] D.N.A.G. K. Geetha, Monikandan S, "SEMALRU: An Implementation of modified web cache replacement algorithm," IEEE, 2009.

[12] E.M. Georgios Kastaniotis, Vasileios Dimitsas, Christos Douligeris, Dimitris K. Despotis, "Web Proxy Caching Object Replacement: Frontier Analysis to Discover the "Good-Enough" Algorithms," IEEE, 2008.

[13] G.P. Sajeev, and M.P. Sebastian, "A scheme for adaptive web caching based on multi-level object classification," Intelligent and Advanced Systems (ICIAS), 2010 International Conference on, pp. 1-6.

[14] T. Berczes, "Approximation approach to performance evaluation of Proxy Cache Server systems," 2009.

[15] S.M. Abid, and H. Youssef, "Impact of One-Timer/N-Timer Object Classification on the Performance of Web Cache Replacement Algorithms," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, pp. 208-211.

[16] F.K.G.K.T.M.P.a.S. Fdida, "The effect of caching on a model of content and access provider revenues in information-centric networks", 2013.