IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Concept of Prefetching and Caching in Web Usage Mining

Pushpraj Singh Chauhan[1] , Sarvottam Dixit[2] , Suresh Jain[3]

[1]Mewar University, Chittorgarh,India
[2]Mewar University, Chittorgarh, India
[3]Medicaps University Indore, India

Available online at: www.ijcseonline.org

Accepted: 12/Nov/2018, Published: 30/Nov/2018

***Abstract:-* :** Since the growth of internet is increasing day by day, hence the amount of data that is storing in Web Server is also increasing rapidly. The growth of number of users of internet is also increasing at a rapid rate, this in turn increasing the Web traffic, so we need some type of strategies or mechanism that can handle this rapid growth of Web traffic. Web Prefetching and caching are techniques that can be used to deal with this increased growth of Web Traffic. Web prefetching and caching are processes that prefetch frequent pages which are likely to be requested in near future and caching is used to store these pages in Proxy Cache Server. Here we have proposed some cache replacement policies by which the hit ratio is likely to get increased. We have proposed novel pre fetching and caching scheme to access frequent data items. It helps in improving pattern analysis, and pattern generation process. Proposed techniques will be useful in E-commerce, Web personalization for customer requirement & satisfaction. This will reduce the user overall access time in future.

***Keywords*—**World wide web, Web log, Web mining, Web usage mining, Web Transactions, Caching policies, LRU, LFU, Web Prefetching, Web caching, Hit Ratio.

I. INTRODUCTION

WWW has a tremendous effect on user,its impact is growing rapidly with the increase in number of users. Due to which huge amount of information is available in the internet,extracting useful information is a big challenge.Various data mining algorithms are available to extract meaningful information from big database.To mined useful data and stored them for further use is based on the concept of web prefetching and caching.

Web Transaction

Web Transaction can be defined as an act of performing some operations over an internet. When a sequence of URLs get combined to complete a process, then it is known as web transaction. A complete Web Transaction means when a user opens its Web browser, access some Web Sites , perform some activities on it. Then all these activities grouped together to make a complete sequence

of operation which is known as Web transaction. Client-Server computing systems in which we use Web browsers as client systems are called web application Systems.

WEB MINING

Web Mining is the application of data mining techniques to extract patterns from the WWW. This is the information that we can extract from Web. There are various tool that are available and can be used to extract information from Weblogs which we can obtained from various Web Servers. As the name itself suggests these informations extracted by web Mining.

WEB CACHING

The strategy of enhancing the performance of web based Systems is known as Web caching. There are three levels on which Web Caching mechanism can be implemented.

They are original Server level, proxy level, client level [1],[2]. In order to minimize the response time of web user's request between user and websites proxy servers can be used. This in turn helps in saving the network bandwidth. Therefore, for achieving a better response time an efficient caching schemes can be built at Proxy Server Level.

Various Cache Replacement Policies are the core/heart of Web caching. Hence, in order to achieve better caching, better Cache replacement Algorithm should be designed. Hence , sometimes cache replacement algorithms can also be called as web caching algorithms. There are various caching policies that are used in the past but they are not efficient.

The main drawback with them is that, they consider one factor at a time while ignores other factors. Due to these limitations an intelligent mechanism or caching policies are required to manage the Web cache contents efficiently. [3],[4],[5]. In the traditional caching policies, some of the web pages never or very seldom gets the request from user

and some web pages get very frequent requests, this can be referred as cache pollution problem. Hence in the past various researchers proposed various cache replacement policies for achieving good performance.

There are combinations of factors that can influence the replacement process. Replacement decision is not an easy task because one factor in a particular situation or environment is more important than other environments [6].

Hence, to determine which web data should be re accessed or which should not be, is a very difficult task for the researchers.

This can be also explained as which web data should be replaced from cache and which (web data) should not be, to improve the hit ratio, alleviate loads, reduce network traffic is a very difficult task [7]. Unfortunately, caching schemes does not improve the cache hit ratio too much. Despite of using caching schemes the hit ratio is limited (40% to 50%) Even if we delimit the cache size[8],[9],[10].

After prefetching the most likely used Web pages gets stored in the cache. It is the automatic creation of temporary copies of information residing on computers other than host servers in order to make this information readily available to people around the world.

1.7.1 Types of Web Cache:

Caching in simple words is a process to keep the frequently accessed documents in places near to the end user to decrease the user perceived latency.

In Web caching, those items which are frequently occurred are put in Proxy Cache. This is done for increasing the performance of web based applications. Web caching is implemented at three levels main server level, proxy server level and cache level [11]. To enhance the performance proxy cache is implemented in between client and websites by which response time of user is saved and network bandwidth is saved. To manage cache properly cache replacement policies are implemented which is also termed as web caching algorithms [12]. Cache are managed properly because limited space is available in the cache, so content which are likely to be used in the near future is only stored, rest of the web objects are discarded on the basis of page replacement policies. Web objects which are not used in the near future and stored in a cache causes cache pollution problem .In this work we cached those objects in the cache which are used in the future by integrating both web prefetching and caching techniques to increase the hit ratio, reduce network traffic and load on server [13], [14]. In many caching scheme hit ratio varies in a limited range from 30-40% in respect to other replacement policies [15],[16].

Based on the places where the caches are placed web caches can be categorized as below:

1.7.2 Browser Cache:

This type of cache is the one placed in the client's machine. Browser Caches are those created by the popular web browsers such as Google Chrome, Internet Explorer, Netscape, Mozilla etc. on the client machine.

1.7.3 Proxy Server Cache:

Proxy server is a computer system which is found in between the end device and the origin server. Proxy server cache provides the same functionality as the browser cache but on a larger scale. Browser cache is created only for one user but a cache created at the proxy server serves many different users in the same way. Whenever a client request any object, the request first goes to the proxy server where it checks if the object is available in its own cache. If it is available the request is fulfilled by the proxy itself, if not the request is forwarded to the origin server.

1.7.4 Origin Server Cache:

The same caching technique can be deployed at the origin server also to reduce the server load.

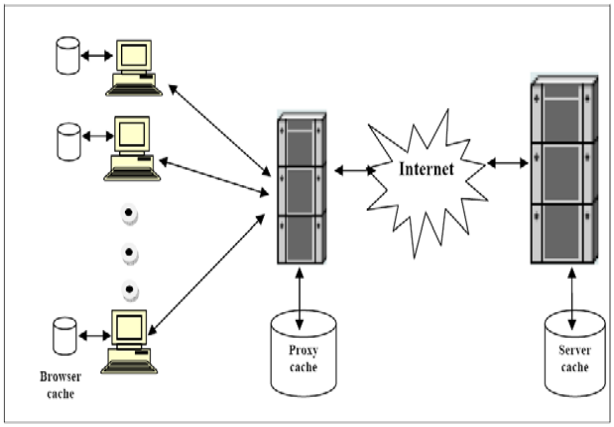

Fig. 1 Web Caching

1.8 WEB PRE-FETCHING

Since the Web is increasing at a rapid rate , hence Web caching mechanism alone can not handle the network latency. Due to this, Web Prefetching is required along with web Caching.

Web prefetching is a process to prefetch the pages in advance to fulfill the user request. To reduce latency web pages are prefetched by proxy server before a client make a request.

Web prefetching is implemented between proxy server and web server, proxy server and client[17], client and web server. Better implementation is between client and proxy server as web pages are stored in advance in a proxy cache so to reduce internet traffic.

1.8.1 Web Pre-fetching Techniques:

In Pre-fetching, we predict the future requests of the users. This is done by studying the Web Log. Web Log helps us in analyzing the user's behavior. Many researchers are interested in predicting the future request based on their past activities.

This is because most people browse and explore the new web pages trying to find new information. We can integrate the Web pre-fetching technique with web caching for improving the Performance. In recent years web prefetching and caching remains a hot topic for researchers and also it gains attention from researchers.

In prefetching and caching the web data is prefetched and cached even before a user requests for it. This in turn helps in minimizing user perceived latency. It has been shown in various researches that when we use caching along with prefetching the performance almost get double as compared to single caching [18],[19],[20],[21],[22].

According to [23],[24] , web caching improves the latency upto 26%, but when we use prefetching and caching together then latency gets improved upto 60%.

Moreover, the cache space is not used optimally. For improving the performance of Web Based System Web Caching can be used, which is a successful solution for improving the efficiency. In Web caching, the Web data which is likely to be visited in near future is kept in the Server (proxy) which is kept nearest to the Web user.

So whenever a user request a web page it can be accessed directly from the Proxy Server, this helps in accessing the Web page quickly and it also reduces the overall response time. Thus, scalability of Web Systems can be improve by web caching. There are three main advantages of Web caching . They are

a. Perceived latency is decreased by Web caching.
b. Network bandwidth usage is reduced by Web caching.
c. Origin servers loads get decreased by Web Caching.

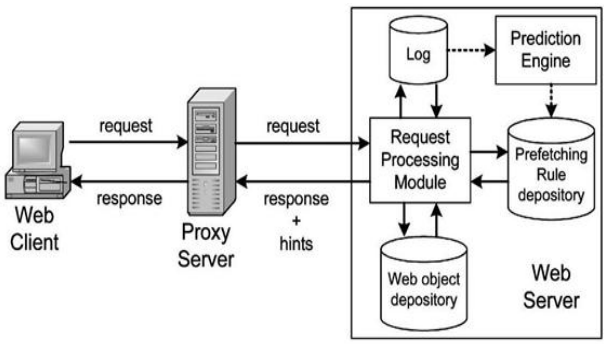

Fig. 2 System Model of Web Prefetching and Caching

II.RELATED WORK

Schloegel et al. represents web log file through graph theory .The paper also describes about web navigational graph using web log files with partition techniques [25].

G. Bejerano et al. proposes a Variable length Markov model (VLMC) is that allow to capture variable length history of user navigation session. In VLMC the probability of next traversed link is chosen and is used to reach the page [26].

J. Borges and M. Levene et al shown Transactions based on Web as Innovations and also examines its factors. Accuracy of Second order Markov model is high as compared to First Order Markov model. In the same way as higher order model is built the accuracy level will be increased [27].

Nanhay Singh, Arvind Pawar, Ram Shringar Raw et al. describe about the comparison analysis of different pattern recognition techniques , filtering, IP address to domain name and by recognizing the bandwidth/Hit comparison for image files to improve the website performance by structure content delivery and presentation [28].

K R Baskaran, Dr. C.Kalarasan, A Sasi Nachimuthu et al. described prefetching using clustering technique with machine learning concept amd it is combine with support vecor machine to give better result for caching. This paer presents SVM is better than clustering technique and LFU used improves bandwidth utilization and access latency. Main aim is to gain high bandwidth utilization, by reducing load on the origin server with high access speed are possible by combining Web caching and prefetching techniques [29].

Ravinder Singh,Bhumika garg et al. proposed a framework for web caching and prefetching together using Dynamic technique which maps into Domain Top approach. Frequent access domains are kept in the list and ranks are calculated. Most poular domains are rank according to the access latency of user [30].

III.PROPOSED WORK

WEB LOG DATA

When user interact with web server then his behaviour is stored in a file which is called as web log file in a textual format.Web log contain include browser information,server information,Proxy server information in different file formats.In our work we have used proxy server web log which contain lots of information,to extract useful information from web log data preprocessing is done where irrelevant data is removed from the web log.

DATA PREPROCESSING

Preprocessing of data is to filter unwanted information from the web log by many data preprocessing techniques like data cleaning,session and user identification and data transformation.Afer preprocessing filter data is used for further pattern generation.

WEB PATTERN DISCOVERY

Pattern discovery is a very important step because it is related with web prefetching and caching.To generate more patterns is necessary for any data mining algorithms as these patterns which are occurred frequently are prefetched and stored in a proxy server cache so that performance of web based application is increased.In our work Apriori and FP growth both algorithms are used to generate patterns.

WEB PATTERN ANALYSIS

Web pattern analysis is a process of analyzing useful patterns by representing them in a graph, charts and in other forms.

LRU

LRU is least recently used algorithm which replace the page that are least used and manage the cache properly with min page faults and max hit ratio. LRU is based on greedy approach where pages are constantly replaced until optimal replacement is achieved. It is a best algorithm to achieve hit ratio in terms of cache size as compare to LFU and Optimal.

LFU

LFU is a page replacement algorithm that replace the page which less frequent, so that proper management of cache is done**.**

PROBLEM DEFINITION

Caching is used to increase the performance of Web.In existing research prefetched pages are less generated by mining algorithm so that less pages are cached in a proxy cache which in turn increase the page fault rate.To overcome this in our work we integrate both prefetching and caching technique to increase the accuracy by generating more patterns through FP growth algorithm,which is helpful to prefetch more pages and put them in the proxy cache.In addition we used LRU , LFU and Optimal replacement policies to properly manage the cache by replacing the pages which are not used by the user in the future and compare them on the basis of hit ratio with cache size.

IV. RESULT ANALYSIS

In the result section we show how cache replacement policies work.to represent the hit ratio in terms of cache size.Prefetched pages are cached in the proxy server cache to serve the user request.Prefetched pages are generated by apriori and FP growth algorithm.

In this paper 3 cache replacement policies LRU,LFU and Optimal are used and each is compared on the basis of hit rate and cache size.

In our work first we have used web log which gives user information,after applying preprocessing on web log it is filtered to get useful information.Then cluster is formed on the basis of IP location of user,that location is found by mapping IP address to user location.Now frequent patterns are generated through mining alogorithms on he basis of location.After this prefetched urls are stored in a proxy cache,that url which is prefetched is given specific number to have unique identification.This urls is map to unique number which is stored in a cache.

Now this unique number is passed to cache replacement policies to manage the cache according to request made by the user in the future and compare the result by calculating hit ratio in terms of cache size.

LRU(Least Recently Used) used to discard urls which are least recently used and replace them with the new request .Graph to represent the hit ratio by LRU.

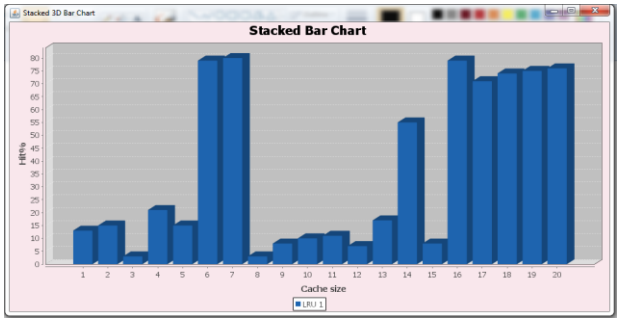

Fig. 4 LRU hit ratio vs Cache size

LFU(Least frequently used) is used to discard urls which are least frequently used and replace hem wih the new request made by the user.Graph to represent hit ratio by LFU.

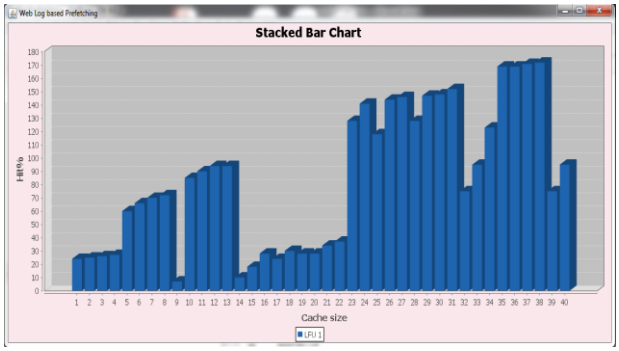

Fig. 3 LFU hit ratio vs Cache size

V.CONCLUSION

In this paper cache replacement policies are used to manage the cache properly with the increase in hit ratio ,it is done to store those pages in the cache which are likely to be requested in the near future based on the prefetch pages generated by Apriori and FP growth algorithms.This paper also shows the comparison of three replacement policies with the hit ratio in terms of cache size. In the future scope other techniques will be implemented to increase the hit ratio with minimum page faults to enhance the performance of web server.

REFERENCES

[1] H.T. Chen, Pre-fetching and Re-fetching in Web caching systems: Algorithms and Simulation, Master Thesis, TRENT UNIVESITY, Peterborough, Ontario, Canada(2008).

[2] T.Chen, "Obtaining the optimal cache document replacement policy for the caching system of an EC Website", European Journal of Operational Research.181(2),(2007), pp. 828. Amsterdam.

[3] T. Koskela, J. Heikkonen, ,and K. Kaski, (2003). "Web cache optimization with nonlinear model using object feature", Computer Networks journal, elsevier , 43(6), (2003), pp. 805-817.

[4] J. Cobb, and H. Elaarag, "Web proxy cache replacement scheme based on back-propagation neural network", Journal of System and Software, 81(9), (2008), pp. 1539-1558.

[5] R. Ayani, Y.M. Teo, and Y.S. Ng, "Cache pollution in Web proxy servers", International Parallel and Distributed Processing Symposium (IPDPS'03), 22-26 April 2003, pp.7.

[6] A.K.Y. Wong, " Web Cache Replacement Policies: A Pragmatic Approach", IEEE Network magazine, 20(1), (2006), pp.28–34.

[7] I. R. Chiang, P. B.Goes, and Z. Zhang, "Periodic cache replacement policy for dynamic content at application server", Decision Support Systems, Elsevier, 43 (2), (2007), pp. 336-348.

[8] H.k. Lee, B.S. An, and E.J. Kim, "Adaptive Prefetching Scheme Using Web Log Mining in Cluster-Based Web Systems", 2009 IEEE International Conference on Web Services (ICWS), (2009), pp.903-910.

[9] L. Jianhui, X. Tianshu, Y. Chao. "Research on WEB Cache Prediction Recommend Mechanism Based on Usage Pattern", First International Workshop on Knowledge Discovery and Data Mining(WKDD), (2008), pp.473-476.

[10] A. Abhari, S. P. Dandamudi, and S.Majumdar , "Web Object-Based Storage Management in Proxy Caches", Future Generation Computer Systems Journal , 22(1-2), (2006). pp. 16-33.

[11] H. Elaarag and S. Romano, "Improvement of the neural network proxy cache replacement strategy", Proceedings of the 2009 Spring Simulation Multiconference,(SSM'09), San Diego, California, (2009), pp: 90.

[12]. Koskela, J. Heikkonen, ,and K. Kaski, (2003). "Web cache optimization with nonlinear model using object feature", Computer Networks journal, elsevier , 43(6), (2003), pp. 805-817

[13] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. In CHI-96, Vancouver, 1996.

[14] W. Ali, and S.M. Shamsuddin, "Intelligent Client-side Web Caching Scheme Based on Least recently Used Algorithm and Neuro-Fuzzy System", The sixth International Symposium on Neural Networks(ISNN 2009), Lecture Notes in Computer Science (LNCS), Springer-Verlag Berlin Heidelberg , 5552, (2009), pp. 70–79.

[15] W. Tian, B. Choi, and V.V. Phoha,"An Adaptive Web Cache Access Predictor Using Neural Network". Proceedings of the 15th international conference on Industrial and engineering applications of artificial intelligence and expert systems: developments in applied artificial intelligence, Lecture Notes In Computer Science(LNCS), Springer- Verlag London, UK 2358, (2002).450-459.

[16] M.S.Chen, J. Hart, and P.S. Yu. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering, 8(6):866- 883, 1996.

[17] Rabinovich M, Spatsheck O. Web caching and replication. Addison Wesley; 2002.

[18] T. M. Kroeger, D. D. E. Long, and J. C. Mogul, "Exploring the bounds of web latency reduction from caching and prefetching", Proceedings of the USENDC Symposium on Internet Technology and Systems, (1997), pp. 13-22.

[19] U. Acharjee, Personalized and Artificial Intelligence Web Caching and Prefetching. Master thesis, University of Ottawa,Canada(2006). A Survey of Web Caching and Prefetching.

[20] Y.f. Huang and J.M. Hsu, "Mining web logs to improve hit ratios of prefetching and caching". Knowledge-Based Systems, 21(1), (2008), pp. 62- 69.

[21] G. Pallis, A. Vakali, and J.Pokorny, "A clustering-based prefetching scheme on a Web cache environment", Computers and Electrical Engineering, 34(4), (2008). pp.309-323.

[22] W. Feng, S. Man, and G.Hu, "Markov Tree Prediction on Web Cache Prefetching", Software Engineering, Artificial Intelligence(SCI), Springer- Verlag Berlin Heidelberg, 209,(2009). pp. 105–120.

[23] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1), 1999.

[24] Zhang T., Ramakrishnan R., and Livny M., "Birch: AnEfficient Data Clustering Method for Very Large Databases." In Proceedings of the ACM SIGMODConference on Management of Data, pages 103-114, Montreal, Canada, June 1996.

[25] Schloegel K, Karypis G, Kumar V. Parallel multilevel algorithms for multi-constraint graph partitioning. In: Proceedings of 6th international Euro-Par conference. September 2000. p. 296– 310.

[26] G. Bejerano, "Algorithms for Variable Length Markov Chain Modelling, Bioinformatics, vol. 20, pp. 788-789, Mar. 2004.

[27] J. Borges and M. Levene. "Evaluating Variable Length Markov Chain Models for Analysis of User Web Navigation Sessions", IEEE Transactions on Knowledge and Data Engineering,19 (4), pp. 441-452, April 2007.

[28] Nanhay Singh, Arvind Panwar and Ram Shringar Raw. Enhancing the performance of Web Proxy Server using Cluster Based Pre-fetching technique. IEEE 2013.

[29] Study of Web Pre-Fetching With Web Caching Based On Machine Learning Technique " (K R Baskaran, Dr. C.Kalarasan, A Sasi Nachimuthu) (2013).
[30] Hybrid Approach for Performance of WebPage Response through Web UsageMining"(Ravinder Singh,Bhumika garg) (2014).

Authors Profile

Mr. Pushpraj singh Chauhan has completed his BE from RGPV University in CSE in 2004. He has completed his M-Tech (IT) from SOIT –RGPV.He is Currently pursuing his Ph.D. from Mewar University Chittorgarh. He has guided various M-Tech Students and published more than 20 research papers in reputed International Journals and Conferences. His area of research is Web Mining, Machine Learning, Network Security.

Dr. sarvottam Dixit is an M.E. (Computer Science), Ph.D. (Material Science) from "Agra University" (now called "Dr. B. R. Ambedkar University"),India and has done Post Doctoral Research work in Astro-Physics at TIFR, Mumbai. He is currently working as Professor in Department of Computer Science and Engineering, Mewar university, Chittorgarh (Rajasthan). He is a member of various Computer Societies. He has published more than 35 research papers in reputed International Journals and Conferences. His main research work focuses on Cryptography Algorithms, Big Data analytics and Computational Intelligence based education.

Dr. Suresh Jain received the B.E. Degree in Civil Engineering, from Maulana Azad National Institute of Technology, Bhopal, India in May 1986,ME.in Computer Engineering from Shri Govindram Sakseria Institute of Technolgy & Science, Indore,India in April. 1988 and the Ph.D. Degree in Computer Science from Devi Ahilya University Indore ,India in March 2007.He is presently the Professor & Head (CSE) in Medicaps University Indore, he was a Professor in Institute of Engineering & Technology, Devi Ahilya University, Indore. His research Interests are in Machine Learning, Grammatical Inference , Artificial Intelligence, Graphics andMultimedia, Database Applications.He is a Senior Member of the ComputerSociety of India and ISTE societies and Life Member of the Indian Society for Technical Education.

