IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Improved Integrated Approach of Web Prefetching & Caching using eLRU

Arshi Khan[1*], Pushpraj Singh Chauhan[2]

[1]Computer Science and Engineering Department, Bansal Institute Of Research and Technology, RGPV, Bhopal, India
[2] Computer Science and Engineering Department, Bansal Institute Of Research and Technology, RGPV, Bhopal, India

Available online at: www.ijcseonline.org

Accepted: 06/Jul/2018, Published: 31/Jul/2018

Abstract— World Wide Web emerged as a dominant platform where user interaction with the web is increased rapidly which provides the user interest for accessing of resources from the web servers. To improve the web server performance pre-fetching and caching is used with the association rule mining where rules are applied to predict the user request based on the previous request and where most frequently access pages are pre-fetched and cached. In this paper, a better algorithm eLRU is proposed which is enhanced LRU for predicting the most accessed web pages by replacing least recently and no longer access pages and comparison between eLRU and LRU page replacement is also shown. This paper also represents the association rule approach for mining information and also by applying Apriori algorithm with FP growth algorithm for accessing pre-fetched pages.

Keywords— World Wide Web, Web log file, Web usage mining, Apriori, FP growth, eLRU, LRU, Web Caching, Web Pre-fetching.

I. INTRODUCTION

Information technology is an explosive area for the generation of access amount of information .To extract these data researchers have developed many approaches, so that useful information is mined.

Mechanism to analyze and extract useful and relevant information from large amount of database is termed as data mining. Data mining is exploited in many areas as banking, e-commerce, military and scientific purpose, www etc. Due to increase in rapid growth of internet users world wide web becomes a huge source to extract data and information, for this data mining technique is used which extract useful data including web documents, hyperlinks, web log files etc [1].Thus mining useful data and generate patterns becomes a research area Many algorithms are used to generate patterns and analyze it to enhance the web server performance.

1.1 Association Rule Mining

Mining is used for web based applications which generates patterns from the web data log. [1].Data Mining involves association rules, Decision tree, Pattern discovery and pattern analysis. In data mining when web log data is taken then it will be processed first with the help of pre-processing unwanted data are removed from the web log file through many data pre-processing techniques.

Association Rule is applied in web log data to extract items which are likely to be requested together, if a rule is consider as A->B(S, C), where A is an item and B is also an item with support count and confidence .As an example when this rule is applied in a shopping mall where number of items are to be buy by the users simultaneously. If user buy item A then there is a chance for a item B to be buy, so in a mall items which are linked with each other are kept nearby, this is done to increase the profit rate. Similarly to preserve user request in a cache association mining is applied where items are considered as resources and transactions are considered as URL, if X->Y association rule is applied then URL 'X' is requested by the user then there is a chance for URL 'Y' to be requested. so for this the web caching is done,to properly manage the cache replacement policy eLRU is proposed which gives more hit ratio as compared to LRU

1.2 Decision Tree

Decision tree is a classification tree which is having a set of rules and is generated as a tree. It is having a set of attributes as training set and a test set. They are used to construct a classifier to represent a class with training set object values and can used to predict the objects belong to another set value.Decision tree is used to construct a tree which takes decision for the possible outcomes.It is represented by three nodes- a decision node, a chance node and an end node with branches in a tree.

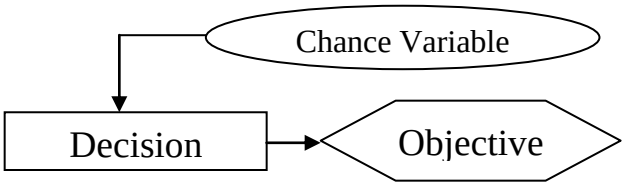

Figure1. Decision tree interface

If web log data is taken and user records are identified then there is a chance to predict next user request. If one url accessed by the user in a session and another url accessed by the user in that same session then there is a chance for that

url to be requested again by the user so that would be put in a cache instead of serving a request a from the server. If sometimes when lower internet speed occurs so local cache works as a best solution where objects which are likely to be requested is put in a local cache.

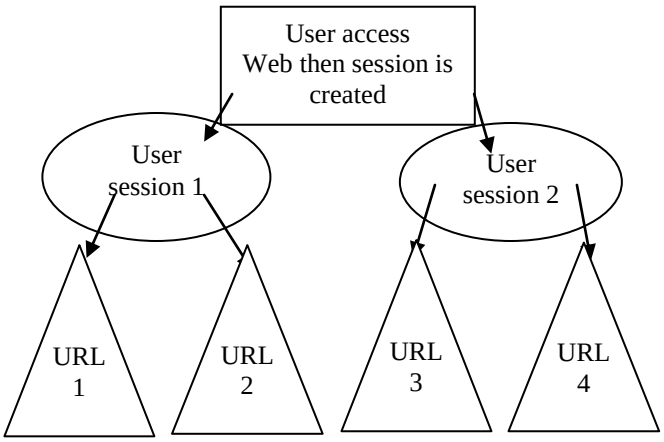

Figure2. Decision tree for url retrieval

Figure.2 represents URL1 ---> URL2 means rule is made where when user request for url1 in a same session then there is a chance for url2 be requested in the same session.
After this patterns are generated by pattern generation algorithms in data mining are used to generate most accessed patterns and help them in pre-fetching and caching. After completion of these steps user can use this information according to their specific needs. In different fields web usage mining show impact as web personalization, e-commerce area, user recommendation sector, pre-fetching the most access pages and cache pages for user navigation, tried to enhance the performance of web with web design improvement.
Mining can be used in various applications shown below:

a. Recommendations: Recommendation is a process to access popular products and services used by the user. To analyze user behaviour based on past and current and to recommend new user to buy frequent access pages and most buy products.

b. Pre-fetching and Caching: To improve the performance of web server and web applications, concept of web usage mining came into existence. Pre-fetching and caching helps immensely for better server performance.

c. Web site design and improvement: Web usage mining provides the ease of use in designing of web applications. Adaptive web sites emerged as one of the application of web site design improvement and provide user feedback to enhance quality of website.

d. Business intelligence: Extraction of business intelligence for website improvement from the web usage data plays an important part for online commercial website. It retains various issues as customer attraction, customer retention, customer departure and cross sales.

1.3 Web Caching

After pre-fetching the most popular pages it is likely to be stored in the cache. Web caching is a concept where caching of frequent occurred objects are put in a proxy cache to increase the performance of web based applications. Implementation of web caching is done based on three levels as main server level, proxy server level and cache level [5, 6]. 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 [7]. Cache are managed properly because limited space is available in the cache, so content which are requested for future request is only stored, rest of the web objects are discarded based on the page replacement policies. It causes pollution problem when web objects stored in a cache are not requested by the user. In this work we cached those objects in the cache which are used in the future by integrating both web pre-fetching and caching techniques to increase the hit ratio, reduce network traffic and load on server [3, 4].

1.4 Web Pre-fetching

Web pre-fetching is a process to pre-fetch the pages in advance to fulfil the user request. To reduce latency web pages are pre-fetched by proxy server before a client make a request.
Web pre-fetching is implemented between proxy server and web server, proxy server and client, 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.

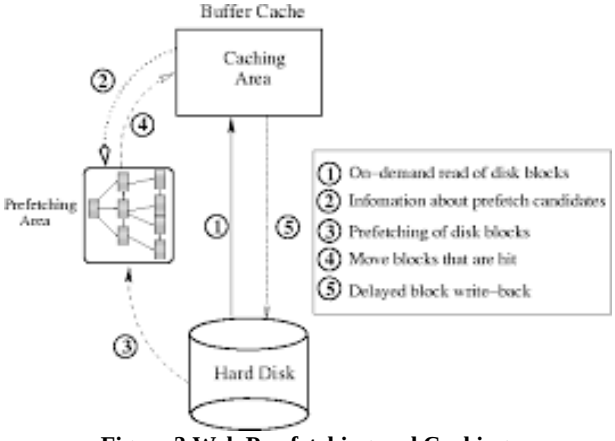

Figure.3 Web Pre-fetching and Caching

II. RELATED WORK

Internet users has become rapidly day by day with the growth of emerging technology, user perceived latency becomes a serious concern for the web service providers. Researchers have been done a tremendous effort to combine various techniques from different domains to resolve this issue. To overcome network latency, prefetching of popular documents is focused by authors. Improvement of web server is increased by integration of pre-fetching and caching concept. It also enhances the running time of application 50% [2].

Garofalakis et al. describes about hypertext, discovering web structure and hyperlink through data mining algorithms and techniques [3]. Also clustering approach [4], based on generalization is introduced which incorporates attribute oriented induction.

Pitkow et al. represents about pattern extraction mechanism in web surfer's path [5]. Build n-gram model [6] for future prediction requests.

Cooley et al. classified web mining for allocation of frequent web pages through data mining algorithms and also neural network is represented [7].

Zhang et al. describes about the hierarchical clustering for large databases which is an efficient approach for generating access patterns from the users. [8]. When user request for the web page first-order Markov models has been used with the clustering technique [9]. In order to reduce the user perceived latency short-term pre-fetching uses Dependency Graph (DG) is used, where graph consist of most access patterns and also Partial Matching (PPM) is implemented..It also provides some drawbacks when pre-fetching policy is not designed cautiously then excessive network traffic is occurred. It also shows optimization of cache space which is not good in this pre-fetching scheme. The long-term pre-fetching describes about the pattern statistics to identify valuable objects by using global access pattern statistics and it has application in Content Distribution Network (CDN), mobile computing environments etc.
Various advantages of web pre-fetching are described in [11, 12], whereas research in web caching [10] grow immensely.

Vakali et al. described about intra-site web pages for web data clustering schemes [14]. It has demerit as complexity of wen increases web clustering performance reduces. If some changes occur in web user's pattern then it has been updated in resultant clusters.

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 [13].

Nanhay Singh et al. proposed two mining algorithms first clustering it through k-means and pre-fetching the patterns by Apriori algorithms of web proxy server [15].

Waleed Ali , Siti Mariyam Shamsuddin, and Abdul Samad Ismail et al. represents about the concept of integrating pre-fetching and caching by improving the web server performance as web objects which are requested in future are to be fetched again by new user, so this objects are pre-fetched and cached to fulfil user request[16].Researchers describes web caching as temporal locality when objects are revisited, and to predict future request from current request pre-fetching represents spatial locality. This paper also represents conventional and investigated approach to integrate web caching and pre-fetching.

Greeshma G. Vijayan1 and Jayasudha J. S et al. represents about the access of internet traffic overloading as large number of users are increasing rapidly which creates web traffic and network bottlenecks, users are accessing many sites which are overloaded and also network links is busy[17], many users didn't have patience so they download the page frequently. Researchers also describes about the pre-fetching and caching technique to reduce the web traffic and network bottlenecks. This paper also presents the web latency to predict web objects requested and also include challenges applied to a mobile environment.

Arun Pasrija et al. Have proposed a system which works in an integrated manner so that certain amount of caching space is to be [18] reserved for pre-fetching. This paper also represents frequent objects which are to be mined in pre-fetching engine objects.

Sonia Setia, Dr. Jyoti, Dr. Neelam Duhan et al. proposed to improve user perceived latency by describing web pre-fetching and caching [19].

K R Baskaran, Dr. C.Kalarasan, A Sasi Nachimuthu et al. describes about pre-fetching technique by using clustering with support vector machine and machine learning concept [20] for the better improvement of caching result .Researchers also compares clustering technique with SVM, with the aim to achieve high bandwidth to combine pre-fetching and caching techniques by reducing web server load. This paper also represents LFU with improved bandwidth use and also access latency is high.

Ravinder Singh, Bhumika garg et al. proposed a framework for web caching and pre-fetching together using Dynamic technique which maps into Domain Top approach [21]. Frequent access domains are kept in the list and ranks are calculated. Most popular domains are rank according to the access latency of user.

Francesco Bonchi Fosca Giannotti Giuseppe Manco Chiara Renso et al. proposed a smart caching model to serve a fast user request. They used data mining algorithms association rule mining and decision tree approaches for web caching [22] to increase the hit ratio in terms of cache size.

III. METHODOLOGY

In the existing works mining of frequent access pages are done through Markov model, Apriori algorithm and FP growth algorithm based on the prediction of pages. Through this the performance of server is likely to improve. In some paper Data mining algorithms are mined with location on the basis of user request. In the proposed work two different Pattern discovery algorithms are used to generate frequent pages on the basis of association rules. We consider FP growth algorithm generates more patterns than Apriori algorithm but when they are used combine better result is evaluated. In the existing work users are clustered according to group, interest and time taken by the server to respond back also with location .In this work to improve the performance, a cache replacement algorithm eLRU is proposed which is least recently and no longer access page replace algorithm as it is an enhanced LRU so that from proxy cache when user request for pages then it will be served through replacement policies to achieve more hit ratio. In this proposed work we use both apriori and FP growth algorithm with frequent item generation and after that pre-fetching and caching is done with the help of page replacement algorithms to improve web server performance and gain in hit ratio,

We can formulate the complete work in the following pseudo code.
Step1 Web log data is collected from the proxy server log files.
Step 2 Apply association rule mining to generate rules.
Step3 Apply pattern generation algorithms to generate patterns based on the association rules.

a. Generated rules predict the most accessed pages are likely to be requested.
B.Pre-fetch the most accessed pages and store them in the proxy cache.

Step4. Apply Cache replacement algorithms to generate more hit ratio and store those pages only in cache which can be request more likely in future and compare the hit ratio by 2 of them.

a. Apply eLRU which discard least recently and no long accessed pages.
b. Apply LRU which discard least recently accessed pages.

Step5. Compare eLRU and LRU on the basis of future page request .
Step 8 Ends

3.1 WEB LOG DATA SET

Whenever there is an interaction between user and the server then there is a generation of text file called as web log file. Web log file contain useful information about the user like IP address, URL accessed, time stamp, number of bytes used, protocol details which are used and requesting method. Web log also contain some irrelevant information, removal of such kind of information and filtered pre-processed data is a necessary step in web mining.
Where web log contains:

- 114.67.87.34.3-IP address
- "-"-it is anonymous user id
- 22/6/2018:05:16:45- Access time of web page
- -700-Yime zone
- GET/HTTP- it is a http request method by the user
- 200-http status code
- 4587-it is number of bytes transmitted

Web log file is a dataset which we used in this work. Web log files are that files which are generated when user and web gets interacted. It contains user behaviour information. Various types of web log files are generated like referrer logs, access logs error logs and with client side cookies also. These web logs are available in various sources and can be downloadable for experimenting.

TYPICAL SOURCE OF DATA

There are many websites from where web log data is available such as:

- E-commerce data
- Web page consist of content and structure
- Profiles of web user

KINDS OF WEB LOG FILES

a. Web Server Logs(Server side)

Server side logs are termed as web server logs which contain information like IP address, timestamp, URL etc.

b. Proxy Server Logs(Proxy side)

Proxy server is that which respond when main server is not working. So the log files generated at that time is proxy server log and it contain information related to proxy server with web server log.

c. Browser Logs(Client side)

This log file is generated at client side when user interact with the server, it stores all information related to web user.

DIFFERENT WEB LOG FORMATS

Web log files which are generated have some file extensions which are listed below.

a. Common log file format:

It is comma separated files which is a straight line record in a standard text log file format.

b. Microsoft IIS log format:

It is a standard format for storing web log files and contains more information than NCSA file. It is a non-customizable format of ASCII.

Fetch Most Occurring Pages

After applying Apriori and FP growth algorithm the frequent pages which are requested in future by the user is pre-fetched and cached in a proxy server cache.

Caching the Web Pages

To reduce the latency pre-fetched pages are stored in a proxy server cache so that when in future user made a same request then it is send from the proxy server cache then main server.

3.2 CACHE REPLACEMENT ALGORITHMS

Cache replacement algorithm is mainly used to discard the pages which are no longer in use and manage the cache to store those pages that are used in the near future. If page which is requested is not in the cache then page fault occurs then that page is to be replaced with the existing page to put it in the cache. various cache replacement policies are used to manage the pages in the cache .In this paper an enhanced version of LRU known as eLRU is used which is likely to discard least recently pages and no longer access pages so that when user made a new request then there is a chance to replace those pages which are not of any use. eLRU replacement are used so that pages which are requested by the user is properly stored in the cache with minimum number of page faults and maximum hit ratio.

3.2.1 eLRU

eLRU is enhanced LRU, least recently and no long access used page algorithm which replace the page that are least used and manage the cache properly with min page faults and max hit ratio. eLRU 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 LRU, LFU and Optimal. eLRU is proposed to increase the web server performance by replacing the web pages which are of not interest for user. It is very helpful for those applications when same request is made by the user then that request is pre-fetched and cached so that less time will be taken to access the request from he cache as compare to server access time because of this the web server performance is increased greatly.

3.2.2 LRU

LRU is least recently page replacement algorithm that replaces the page which is less recently used, so that proper management of cache is done.

PROBLEM DEFINITION

To increase the performance of web based application caching becomes a very important technique which reduces the network traffic, frequent access content are replicated on proxy cache to reduce end user retrieval and also server load .In existing research pre-fetched 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 pre-fetching and caching technique to increase the accuracy by generating more patterns and is helpful to pre-fetch more pages and put them in the proxy cache.For this we proposed eLRU algorithm which is enhanced LRU to discard least recently and no longer access pages as compare to LRU 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.

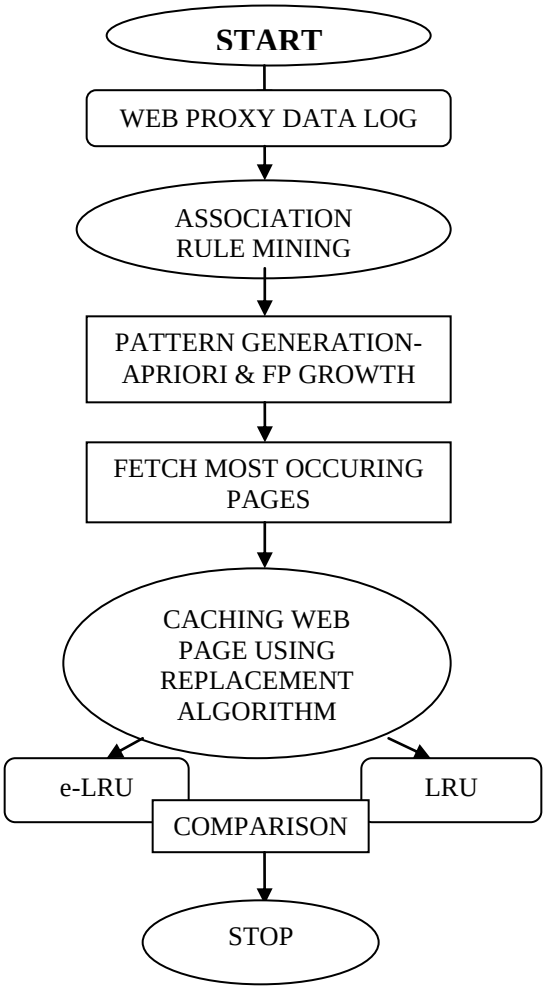

Figure.4 Algorithm flow chart

IV. RESULTS AND DISCUSSION

For the experimental work dataset from the ircache.net website is collected and is used for experiment which is available at the proxy server installation ftp://ircache.net.

In the proposed framework, figure.5 shows first web log data is pre-processed by different pre-processing schemes by removing unwanted data like .gif and other irrelevant

information which is of no use. Pre-processing step filtered the data and give the necessary information for further processing.

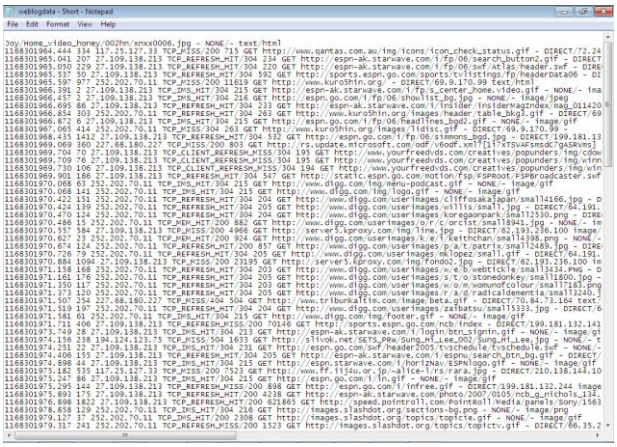

Figure.5 Web log

After that k-means clustering is done to cluster same type of urls. Then by applying FP growth algorithm frequent urls are generated. FP growth generates more pattern then Apriori algorithm. Thus FP growth is best as compared to Apriori for pre-fetching frequent pages.

After this the figure.6 shows that predicted pages are fetched from the server and the most accessed pages are stored in cache of proxy server.

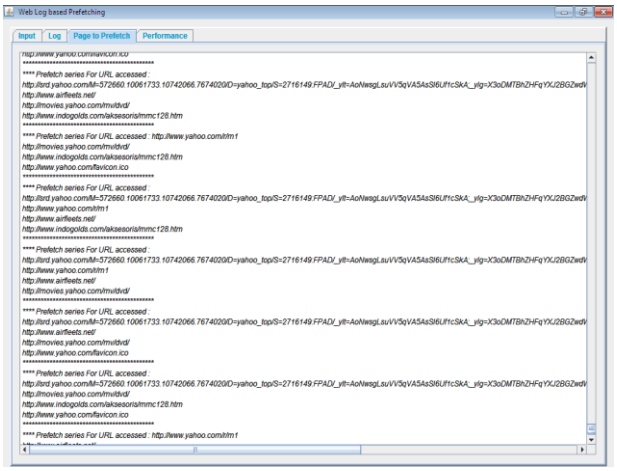

Figure.6 Pre-fetch Pages

In FP growth more association rules is applied as compare to apriori so it generates more number of patterns but access time is more in FP growth. It shows

FP generates more patterns. After pattern generation pages are pre-fetched and stored in a cache. Then cache replacement algorithm proposed eLRU is applied to represent the hit ratio in terms of cache size. Pre-fetched pages are cached in the proxy server cache to serve the user request. Pre-fetched pages are generated by apriori and FP growth algorithm.

Then cache replacement policies LRU, e-LRU are used and is compared on the basis of hit rate and cache size.

In our work frequent patterns are generated through mining algorithms .After this pre-fetched urls are stored in a proxy cache, then cache replacement policies is applied 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.

Figure.7 shows 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.

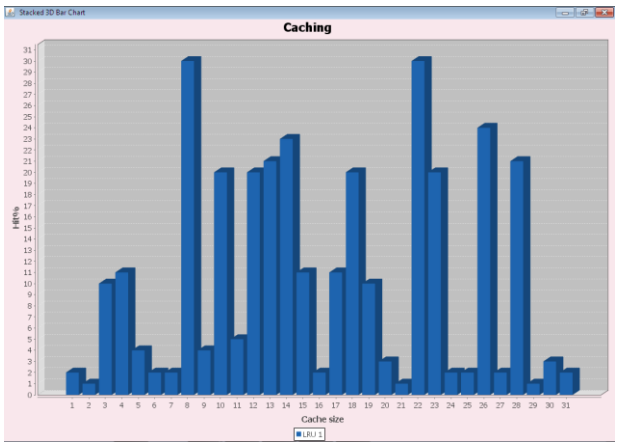

Figure.7 LRU Hit Ratio

Figure.8 shows eLRU (enhanced Least recently and no longer access page algorithm) is used to discard urls which are least recently and not used from a longer time and replace them with the new request made by the user .Graph to represent hit ratio by eLRU.

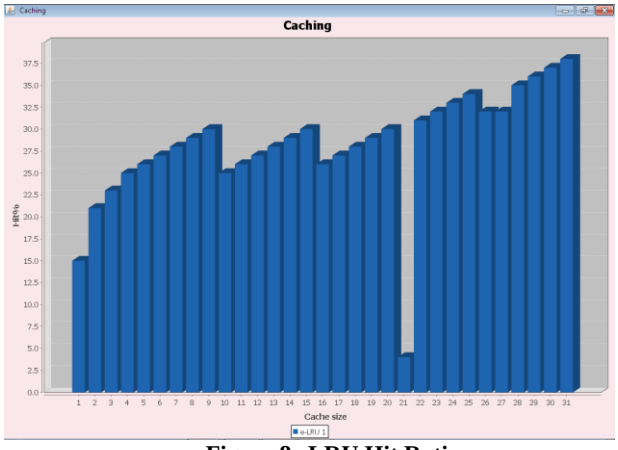

Figure.8 eLRU Hit Ratio

elru show the hit ratio of the algorithms. Hit ratio is increased by 10% to 50% when using eLRU approach. Above graph also shows that eLRU performs better than LRU.

To improve the performance of caching various cache replacement algorithms like LRU, LFU and Optimal are studied and LRU is implanted and compared with proposed approach eLRU. When most access pages are stored in proxy server of cache then that pages are applied to replacement algorithms to gain the more hit ratio because when user request the same pages then it would be cached through LRU, eLRU algorithms.

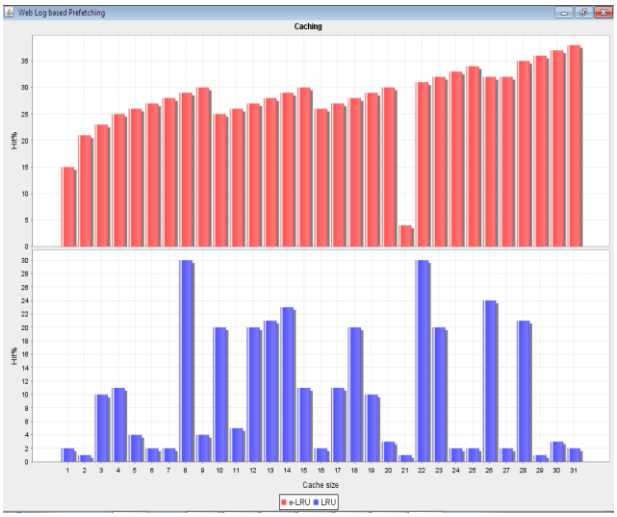

Figure.9 Comparison between eLRU and LRU

Figure.9 shows that hit ratio with respect to cache size which is increased rapidly by using eLRU algorithm approach. Hit ratio is more in eLRU as cache size increase hit ratio also improves as compare to LRU.

V. CONCLUSION AND FUTURE SCOPE

Web caching algorithm is implemented through two approaches based on association rule and decision tree which aimed to increase the hit ratio in terms of cache size. In this paper eLRU is implemented which is an enhanced version of LRU and works as better algorithm by replacing the pages which are not in use for future request. Hit ratio is likely to increase when association rules are applied to cache a object in terms of pre-request. This paper also presents the performance graph between eLRU and LRU in terms of hit ratio to increase the web server performance.

In future scope association rule mining and decision tree both are integrated to generate more optimized result by using clustering strategy.

References

[1] Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Jaideep Srivastava, Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan, ACM SIGKDD, Jan 2000.

[2] K. Chinen and S. Yamaguchi An Interactive Prefetching Proxy Server for Improvement of WWW Latency. In Proceedings of the SeventhAnnual Conference of the Internet Society (INEt'97),Kuala Lumpur, June 1997.

[3] Garofalakis M. N., Rastogi R., Sheshadri S., and Shim K.,"Data mining and the Web: past, present and future." InProceedingsof the second international workshop on Web information and data management, ACM, 1999.

[4] Fu Y., Sandhu K., and Shih M., "Clustering of Web Users Based on Access Patterns" International Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), San Diego, CA, 1999.

[5] Pitkow J. and Pirolli P. Mining longest repeating subsequences to predict www surfing.In Proceedings of the 1999 USENIX Annual Technical Conference, 1999.

[6] Z.Su, Q. Yang, Y. Lu, and H. Zhang. Whatnext: A prediction system for web requests using n- gram sequence models. In Proceedings of the First International Conference on Web Information System and Engineering Conference, pages 200-207, Hong Kong, June 2000.

[7] Phoha V. V., Iyengar S.S., and Kannan R., "Faster Web Page Allocation with Neural Networks," IEEE Internet Computing, Vol. 6, No. 6, pp. 18-26, December 2002.

[8] 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.

[9] Cadez I., Heckerman D., Meek C., Smyth P., and Whire S., "Visualization of Navigation Patterns on a Website Using Model Based Clustering." Technical Report MSRTR-00-18, Microsoft Research, March 2002.

[10] Podlipnig S, Boszormenyi L.A survey of Web cache replacementstrategies. ACM Compute Surveys 2003;35 (4):374–98.

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

[12] TengWG, Chang CY, Chen MS. Integrating Web caching and Web prefetching in client-side proxies. IEEE Trans Parallel Distributed Syst 2005;16(5):444–55.

[13] 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.

[14] Vakali A, Pokorny J, Dalamagas T.An overview of Web data clustering practices. In: Proceedings of the EDBT Workshops 2004. Heraklion, Crete; 2004. p. 597–606.

[15] Nanhay Singh, Arvind Panwar and Ram Shringar Raw Enhancing the performance of Web Proxy Server using Cluster Based Pre-fetching technique. IEEE2013.

[16] A Survey of Web Caching and Prefetching"(Waleed Ali Siti Mariyam Shamsuddin, and Abdul Samad Ismail)(2011)

[17] A Survey On Web Pre-Fetching and Web Caching Techniques in a Mobile Environment"(Greeshma G. Vijayan1 and Jayasudha J. S.) (2012)

[18] Survey on Improving the Performance of Web by Evaluation of Web Prefetching and Caching Algorithms" (Arun Pasrija) (2013)

[19] Survey of Recent Web Prefetching Techniques" (Sonia Setia, Dr. Jyoti, Dr. Neelam Duhan) (2013)

[20] Study of Web Pre-Fetching With Web Caching Based On Machine Learning Technique " (K R Baskaran, Dr. C.Kalarasan, A Sasi Nachimuthu) (2013)

[21] Hybrid Approach for Performance of Web Page Response through Web Usage Mining"(Ravinder Singh,Bhumika garg) (2014)

[22]Data Mining for Intelligent Web Caching Francesco Bonchi Fosca Giannotti Giuseppe Manco Chiara Renso CNUCE-CNR – Institute of Italian National Research Council Via Alfieri 1, 56010 Ghezzano (PI) Italy.

