International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 1 Issue 1 March 2014

Enhancing the Web Pre-Fetching at Proxy Server using Clustering

Monti Babulal Pal
Department of Computer Science & Engineering,
Shri Vaishnav Institute of Technology and Science Indore, India
monti_pal@yahoo.co.in

ABSTRACT:- *Web caching is used to reduce the network traffic by caching web pages at the proxy server level. Now a day's caching alone is not sufficient because of World Wide Web has evolved rapidly from a simple information-sharing mechanism. This mechanism offer only static text and images to a rich assortment of dynamic and interactive services, such as video/audio conferencing, e-commerce and distance learning. Web is demanding to improve the cache performance. If we use the prefetching technique with caching then the performance of cache is improved. Prefetching fetches objects that are likely to be accessed in the near future and store them in advance thus the response time of the user request is reduced. In this paper, our main objective is to give a new framework to improve performance of web proxy server using web usage mining and prefetching scheme. Further, we cluster the user according to their access pattern and usage behavior with the help of K-Means algorithm and then Apriori algorithm is applied to generate rules for prefetching pages. This cluster based approach is applied on proxy server web log data to test the results using LRU and LFU prefetching schemes.*

KEYWORDS*: - Proxy server, Caching, Prefetching, Data mining techniques, LRU, LFU.*

1. INTRODUCTION

Internet has become a popular medium to share and exchange the information. This results in demand of applications that generates high network traffic which puts the limited network infrastructure under excessive loads. One of the possible solution is to increase network infrastructure but that will not only increase the economic cost but will also increase the demand of more network hungry applications demanding more bandwidth, there by achieving the same status as before. So, the solution lies in efficient use of existing network resources. Proxy Servers are popular applications which are used to facilitate clients for accessing web. These servers help in reducing the network traffic and perceived lag by storing copies of web objects accessed in local temporary memory storage area, known as cache and to provide to other users the same page on demand. The cache of proxy servers is limited and therefore there is a need of replacement of pages as soon as cache becomes full. Various algorithms such as LRU (Least Recently Used), LFU (Least Frequently Used) etc[6]. are implemented for removing the stale pages and replacing them by newly requested pages. The question of what to store in cache and for how long, to keep that web object in cache is very pertinent. The caching at proxy could be improved by pre-fetching the pages that may be requested by the users in near time and by following a web caching policy that synchronizes with the pre-fetching technique. The work done in this paper addresses the integration of these two related issues.

2. PREVIOUS WORK

Now days, internet has evolved to provide complex user specific dynamic services. These services have put a great demand on the limited network infrastructure that provides support for these services. The higher demand for services and the limited network infrastructure has caused an inferior experience for the Internet users in terms of higher latency[9, 10, 11]. The simplest solution to this problem is to add network resources and increase the total bandwidth of the network. However, such a solution only encourages the development of applications that consume a higher network bandwidth and deliver a richer user experience thus causing network congestion. Web caching and pre-fetching are the techniques used for enhancing the performance of proxy server. The various approaches proposed on these techniques on the basis of log analysis and applying data mining techniques can be classified in to following categories:

2.1 Clustering based Approach

Web-access prediction model with high accuracy by proposing a method which is termed as cut-and-pick method which is based on basic assumption that same user typically visits multiple related website which forms clusters [2, 4]. The main challenge, as proposed by author in Web Proxy log is transaction identification of each user as proxy log contains information of more than one website and also the information is interleaved with respect to time. The author introduces a more advanced approach for transaction identification between transactions by deciding boundaries between them and then by selecting right reference sequence between them. In this approach the author construct a STG (Site Traversal Graph) from the access pattern of user obtained from preprocessed proxy log to identify the closely related websites for client IP address. While constructing the STG the specific information regarding webpage is discarded and only website information is used. Then by using associate rule mining we prune the less related websites for each client IP and hence obtain a cluster of closely related websites for each user. In this approach a session based method (known as cut-and pick) which is used which ensures that the time difference between two subsequent requests for a client IP is not more than threshold (t) and each request is arranged by the time stamp of the entry in log. Finally, author compares its result with previous approaches (known as fixed time interval & server approach) and shows that his proposed cut-and-pick method is efficient.

2.2 Sequential Data Mining based Approach

Sequential mining approach to mine frequent web access pattern from the raw log of Web server data. In this approach the preprocessed web logs are arranged in access sequence of individual user sequence resulting in access sequence database known as Web Access Sequence Database (WASD), so that sequential mining can be done on it [1, 7].

A Data structure known as WAP (Web Access Pattern) is devised to register access sequences and corresponding counts compactly for eliminating the expansive support count. WAP tree and also maintains linkages for traversing prefix with respect to the same suffix pattern efficiently. Then a graph known as WAP (Web Access Pattern) tree is constructed to mine the frequent subsequence of web access pattern for each user (Client IP) recursively, by scanning the WASD twice only. In first pass, it determines the set of frequent events and next scan WAP-mine builds a data structure, called WAP tree, using frequent events, to register all count events for further mining. Then WAP-mine recursively mines the WAP –tree using conditional search to find all the frequent Web Access Pattern.

2.3 Web Caching and Pre-fetching Approach

It is stated that the popularity of web objects varies considerably as the user interest changes with respect to time and hence it is difficult to select popular objects beforehand and predict a popularity threshold [3, 12]. Integration of web caching and web Pre-fetching schemes leads to better performance of web cache [13]. Web caching scheme does reduce the network traffic and overall bandwidth consumption but it results in low hit rate. In this approach, N^{th} order Markov Model is used to predict user's next request. Prediction by Partial Match (P.P.M.) Model is discussed for predicting future user request. Author obtains user's session from web server access log and then presents an algorithm to find out similar sessions. Then frequently requested web pages within that web site are identified web caching. It is proposed that for identifying a user access session for a very large web server access file, temporal parameters are taken into account and then user's session is built to find popular web pages within website [5].

3. OVERVIEW OF PREFETCHING

Web caching is used to reduce the network traffic by caching web pages at the proxy server level. Web caching is not sufficient for the present and future scenarios because web technology has growing rapidly from a straightforward information sharing device. This device presents only static text and images to a rich set of dynamic and interactive services. Users face unwanted delay in the response from the server side. If we use the prefetching with caching then the performance of cache is improved. Prefetching retrieve web objects that are probable to be accessed in the near future and store them in advance.
Web prefetching system takes advantage of the spatial locality of the web objects trying to calculate a user's browsing sequence. That is, if object *x* has a hyperlink to object *y*, the likelihood that y will be accessed. Given *x* has been accessed already will increase considerably, for this reason and approach should be to prefetch the objects a user can access from another one.

The combination of these two techniques without any scheduling and scheming might cause major performance degradation to each other. For example, to provide the proxy with rich information, a web server may intentionally send all probable prefetching hints with a range of levels of confidences to the proxy creating a blockage in the proxy itself. One of the best solutions to describe a prefetching rule

is through the web logs mining and analysis. Fig. 1 shows the web prefetching scheme diagram in which a web objects repository store the web logs. Web objects is collect and go through the process module and then to the prefetching engine who generate the prefetching rule. The prefetching rules are the accountable to recognize what to store and what to prefetching each time a user request a resource on the server. Next section describes the proposed framework.

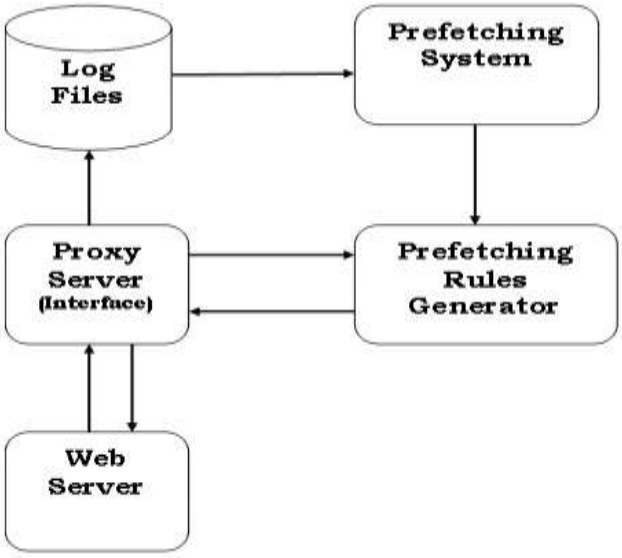

Fig.1. Web Prefetching Scheme

4. PROPOSED ARCHITECTURE

Web caching is used to reduce the network traffic by caching web pages at the proxy server level but nowadays caching alone is not sufficient because World Wide Web has evolved rapidly from a simple information-sharing mechanism to dynamic and multimedia data. To improve the performance researcher shows that combination of prefetching with caching approach is good. In this paper, we give a new framework for web prefetching in which we combine prefetching and caching techniques to improve the performance of proxy server. This section describes the architecture of the proposed framework as shown in fig 2. Following subsection describes the component of the proposed system.

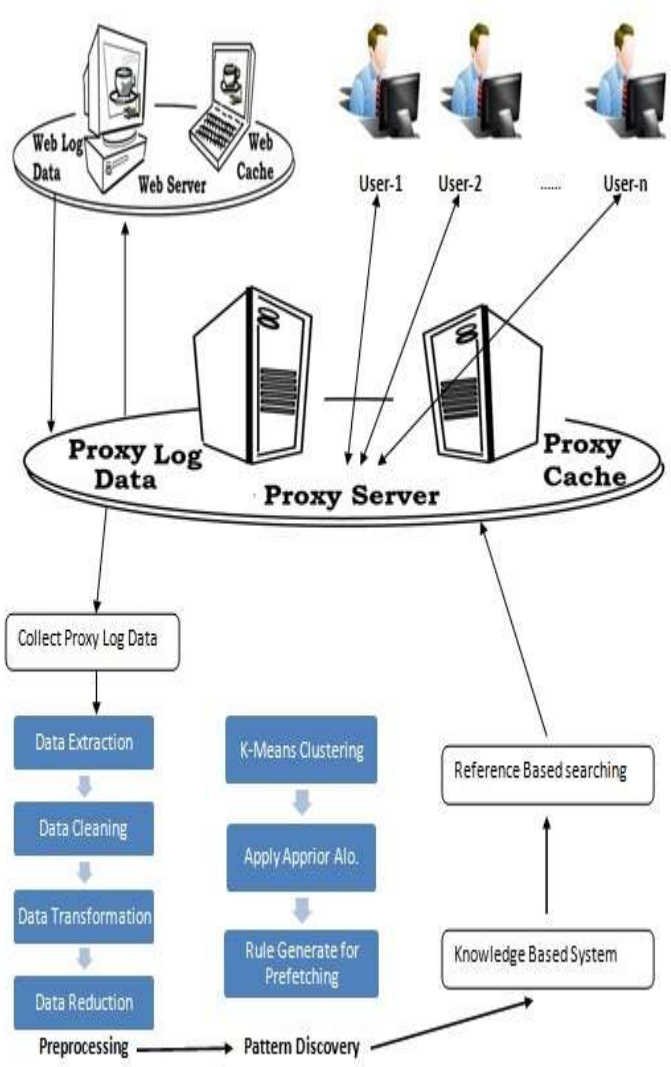

Fig 2 Proposed System Frameworks

4.1 Proxy log data

Web user visits many web sites time to time and spent random quantity of time among various visits. To deal with the user browsing behavior, we should analyze the proxy server log file. In fussy, the web proxy access log is an in order file with one user access data per line. Web proxy log files make available information about actions performed by a user from the moment the user logs.

4.2 Preprocessing

Preprocessing is defined as removing all the irrelevant and noise data from our actual data. Preprocessing stage is

divided into four phases i.e., Data Extraction, Data Cleaning, Data Transformation, Data Reduction. In our proposed approach during preprocessing phase we carried out the cleaning task to filter out all the unwanted entries from the proxy log data. We use the proxy log explorer tool to preprocess the log record of the proxy server. For our work we have filtered out all the log entries which have status code other than '200', therefore only the requests that are fulfilled are analyzed to make further rules.

4.3 Pattern Discovery

In our proposed framework, we have used pattern discovery technique of the Data Mining Process (DMP). DMP is the most important stage of our proposed approach in which we are applying clustering to create users clusters based on their browsing behavior. After completing the clustering process, we are applying Association Rule Mining (ARM) to discover rules among the clusters to predict the future request of web resource. Clustering on the dataset and ARM is described in the following subsection.

a. **Clustering:** To cluster users we use the K-Means clustering which is used to gather different users into clusters on the basis of their usage behavior and searching pattern. The *K-Means* is the simplest clustering algorithm widely used for web proxy server [14]. The Algorithm is used to cluster user's data base on attributes into *K* clusters. Each Cluster has its center (known as centroid) at point *C* .The centroid is calculated from mean distance of all records in the cluster [15, 16].
In this study, we make an assumption that users in the same cluster should have same surfing habits and patterns. Users surfing habits can be determined by several factors such as the time of day of their access, and their most frequently visited websites. After forming the cluster of the similar user on the basics of their access behavior, next step is applying the association to generate the rule.

b. **Association Rule Mining (ARM):** ARM is defined as the process of finding frequent patterns, associations, correlations among set of items or objects in transaction database, relational databases, and other information repositories. Application areas of ARM are basket data analysis, catalog design etc. Using ARM we can easily predict the occurrence of an item based on the occurrences of other items in the transaction. Two main parameters that are very important in ARM and play a very important role in predicting items using rules are Support (S) and Confidence (C). In our proposed framework, we use the Apriori algorithm to find the patterns and generate the rules for the prefetching. Apriori algorithm is very useful to predict the future request of the user according to their past access. Apriori algorithm is the best algorithm used to find association rules[7].

4.4 Knowledge based System

The knowledge base is a warehouse of extracted rules which have been derived by applying the data mining [7] on the web log data of proxy server [12, 13, 14]. These rules are used in the prefetching and sending recommendation by the recommendation engine to the proxy server using generated rule.

5. RESULT ANALYSIS

In this section, we describe our proposed work which is done on the proxy server log data. Dataset is collected from the ircache.net website to carry out our experimental work. We have done the experimental work on the dataset "pa.sanitizedaccess.20070109.gz".This file is obtained from a proxy server installation ftp://ircache.net in Fig. 3.

Fig. 3 Sample proxy server log record of ircache.net

Now we cluster users with the help of *K-Means* algorithm. During *K-Means* clustering on the preprocess data with RapidMiner5 tool create 10 different clusters on the dataset in Fig. 4.

Fig. 4 Data view of the different cluster.

In this section, result analysis has discussed on the basis of experimental work in which we have tested out the hit and byte ratio for two pre defined schemes i.e., LRU and LFU. Simulations have been carried out by implementing the proposed framework in C using MAT LAB. In the experiment we have checked normal scenario without prefetching and then implemented our prefetching approach and tested out the hit and byte hit ratio. We have compared (in Fig.5) our approach using LFU algorithm for page replacement.

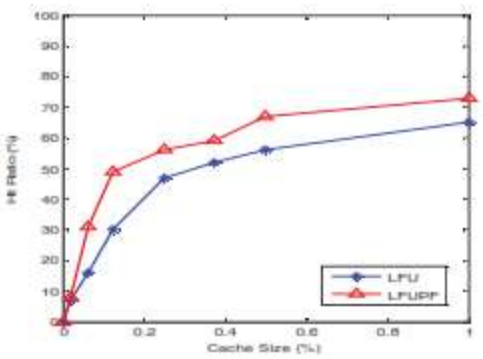

Fig 5 Hit ratio comparison between LFU & LFU with prefetching

6. CONCLUSION

In this paper, we have proposed a framework for prediction of web requests of users and accordingly, prefetching the content from the server. The dataset "*pa.sanitized-access.20070109.gz"* is used for the experimental work which has collected from ftp://ircache.net. The proposed framework improves performance of web proxy server using web usage mining and prefetching scheme which is clear in the result section. The overall performance of Cache(both Page Replacement namely LRU and LFU) by using this proposed framework hit ratio is improved up to 5% to 7% as shown in result section.

REFERENCES

[1] P. Kolari and A. Joshi, "Web mining: Research and practice", Computer Science Engineering .July/August (2004) 42–53.

[2] Jyoti, Sharma & Goel, "A Novel Approach for clustering web user sessions using RST", Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT, 2(1), 2009, 656661.

[3] Abhay Singh & Anil Kumar Singh, "Web Pre-fetching at Proxy Server Using Sequential Data Mining", 2012 Third International Conference on Computer and Communication Technology.

[4] Nanhay Singh, Arvind Panwar and Ram Shringar Raw, "Enhancing the Performance of Web Proxy Server through Cluster Based Prefetching Techniques", 978-1-4673-6217-7/13/$31.00 c2013 IEEE

[5] Cooley R., Mobasher B., and Srivatsava J., "Web Mining:Information and Pattern Discovery on the World Wide Web." ICTAI'97, 1997.

[6] 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.

[7] Garofalakis M. N., Rastogi R., Sheshadri S., and Shim K., "Data mining and the Web: past, present and future." In Proceedings of the second international workshopon Web information and data management, ACM, 1999.

[8] 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.

[9] Podlipnig S., Boszormenyi L., "A survey of web cache replacement strategies, ACM Computer Surveys", 374-398, 2003.

[10] Chen Z., Fu AW., & Tong FC., "Optimal algorithms for finding user access sessions from very large web logs", World Wide Web, 259-279, 2003.

[11] Cheng-Yue, Chang Ming, Syan Chen, "On exploring aggregate effect for efficient cache replacement in transcoding proxies",Parallel and Distributed Systems, IEEE Transactions, 2003.

[12] Sumi Choi, Shavitt Y., "Proxy location problems and their generalizations, Distributed Computing Systems Workshops", 2003.

[13] Lou, Liu Lu & Yang, "Cut-and-Pick for Proxy Log Mining", In Proc. 8th international conference extending database technology, EDBT 2002, Prague, Czech Republic, 2002, 88–105.

[14] Chen, W, & Zhang X, "Popularity-based PPM: An Effective web pre-fetching technique for high accuracy and low storage", International Conference on Parallel Processing, 2002.

[15] Xiao J, & Zhang Y., "Clustering of web users using sessionbased Similarity measures",International Conf.Computer Networks and Mobile Computing, 223-228, 2001.

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

[17] Nanhay Singh, Arvind Panwar and Ram Shringar Raw, "Enhancing the Performance of Web Proxy Server through Cluster Based Prefetching Techniques", 978-1-4673-6217-7/13/$31.00 c2013 IEEE