International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Enhancement of Performance of Proxy Server by Reducing Web Traffic using Web Usage Mining

Nitish Patil[1], Mahesh Kumar Singh[2]

[1]M.Tech. Student VIET, G.B. Nagar

[2] Dept. of Computer Science VIET, G.B. Nagar

Abstract: ***This paper work is focused on the study of the prefetching technique applied to the World Wide Web. This technique lies in processing (e.g., downloading) a Web request before the user actually makes it. By doing so, the waiting time perceived by the user can be reduced, which is the main goal of the Web prefetching techniques. The study of the state of the art about Web prefetching showed the heterogeneity that exists in its performance evaluation.***

Keywords: Web usage mining, proxy server, caching, Data mining techniques

1. Introduction

The Web has evolved rapidly from a simple information-sharing mechanism offering only static text and images to a rich assortment of dynamic and interactive services, such as video/audio conferencing, e-commerce, and distance learning. The explosive growth of the Web has imposed a heavy demand on networking resources and Web servers. Users often experience long and Unpredictable delays when retrieving Web pages from remote sites. Hence, an obvious solution in order to improve the quality of Web services would be the increase of bandwidth, but such a choice involves increasing economic cost. However, the higher bandwidth would solve temporarily the problems since it would ease the users to create more and more resource-hungry applications, bunching again the network. Caching proved itself as an important technique to optimize the way the Web is used. Web caching is implemented by proxy server applications developed to support many users. Proxy applications act as an intermediate between Web users and servers. Users make their connection to proxy applications running on their hosts. The proxy connects the server and relays data between the user and the server. At each request, the proxy server is contacted first to find whether it has a valid copy of the requested object. If the proxy has the requested object this is considered as a cache hit, otherwise a cache miss occurs and the proxy must forward the request on behalf of the user. Upon receiving a new object, the proxy services a copy to the end-user and keeps another copy to its local storage.

2. Literature Survey

To reduce perceivable network latency, researchers focused on prefetching popular documents. In a file system, an integrated model of prefetching and caching is being explained in. Chinen *et al.* focused on referenced pages by prefetching them from hyperlinks embedded in current object. Duchamp enhanced the idea by incorporating the access frequency of hyperlinks. Pitkow *et al.* predicted the web surfer's path in pattern extraction mechanism .Worked on prediction of future requests and has built n-gram model for the same. Cooley *et al.* categorized the web mining and then presented possible research areas. A scheme for fast allocation of web pages using data mining techniques and competitive neural network is being discussed in. Garofalakis *et al.* basically provided a survey on data mining techniques and algorithms for discovering structures of web, hypertext and hyperlink. In, a generalization based clustering approach has been presented, which also incorporates attribute oriented induction. Zhang *et al.* proposed an efficient data clustering approach for very large databases, by generating hierarchical clustering of web users based on their access patterns. In order of user's web page requests, clustering technique using first-order Markov models has been provided in. To learn the mixture of first-order Markov model (which actually represent cluster), an expectation maximization model has been used.

3. Problem Definition

Firstly, this paper is aimed at providing the present state of the art with a solid framework and methodology to evaluate web prefetching techniques from the user's point of view. To do so, an open framework which permits the implementation of prefetching techniques in a flexible and easy way has been developed. Moreover, performance key metrics have been analyzed to detect the most meaningful indexes from the user's point of view. A comparison methodology that takes into accounts the costs and benefits of prefetching has also been proposed. The second main goal of the problem is, once the methodological issues are clarified, to find out how web prefetching algorithms can be improved from the user's point of view and propose a new one that outperforms those existing in the open literature. To do so, the most important algorithms in the literature are analyzed from the user's perspective. In the detailed analysis we found that their performance in current web can be improved by distinguishing two classes of objects: container objects and contained objects. We take benefit of this observation in order to design the new algorithm.As secondary objectives, we explore the performance limits of web prefetching to know the potential benefits of this technique and analyze how to adapt prefetching algorithms to the environment conditions in order to achieve the best results in each situation. As third objectives, we explore optimization of the prefetching to minimize the time

complexity of data mining algorithm or to reduce the data set for rule generation.

4. Technical Approach

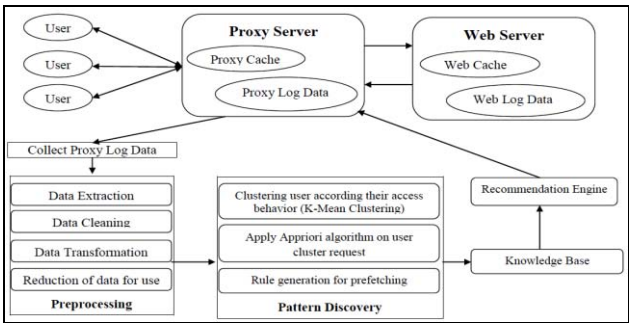


Figure1: Proposed system framework

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 work, we give a new framework for web prefetching in which we combine prefetching and caching techniques to improve the performance of proxy server.Web user visits many web sites time to time and spent random quantity of time among various visits. To deal with the user browsing behaviour, 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. Access logs provide the bulk of the Web server data, including the date, time, users IP address, and user action (e.g., whether or not the user downloaded a document or image).Preprocessing is defined as removing all the irrelevant and noise data from our actual data. 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.To cluster users we use the K-Means clustering which is used to gather different users into clusters on the basis of their usage behaviour and searching pattern. The *K-Means* is the simplest clustering algorithm widely used for web proxy server . The algorithm is used to cluster users data based on attributes into *K* clusters.

5. Result Analysis

We describe our experimental 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.sanitized-access.20070109.gz*". This file is obtained from a proxy server installation ftp://ircache.net.

```
1168300930.290 3689 128.26.236.138 TCP_MISS/304 209 GET http://www.media-indonesia.com/xml/u.gif - DIRECT/219.83.123.74 -
1168300930.292 3690 128.26.236.138 TCP_MISS/304 209 GET http://www.media-indonesia.com/xml/d.gif - DIRECT/219.83.123.74 -
1168300930.742 371 151.33.90.119 TCP_CLIENT_REFRESH_MISS/404 333 GET http://info.ddcd.jp/ddcd3_info/fujitsu/Fujitsu_JPN_CD_News(421).txt - DIRECT/210.174.185.15 text/html
1168300930.860 7 55.156.163.104 TCP_MISS/504 1611 GET http://delb.myspace.com/DartRichMedia_1_03.js - NONE/- text/html
1168300931.105 454 55.156.163.104 TCP_REFRESH_MISS/504 1683 GET http://i.a.cnn.net/money/.element/img/1.0/data/mk_snapshot/BigCharts_gradient.gif - NONE/- text/html
1168300931.243 28 55.156.163.104 TCP_REFRESH_MISS/504 1651 GET http://i.a.cnn.net/cnn/video/business/2007/01/07/debt.vs.affl.jpg - NONE/- text/html
1168300931.296 765 53.141.144.101 TCP_MISS/200 19170 GET http://pictures.match.com/pictures/12/46/51601246A.jpeg - DIRECT/63.147.175.35 image/jpeg
1168300932.118 16 55.156.163.104 TCP_REFRESH_MISS/504 1655 GET http://i.a.cnn.net/cnn/video/business/2007/01/05/foreclosure.vs.jpg - NONE/- text/html
1168300933.475 28 55.156.163.104 TCP_REFRESH_MISS/504 1673 GET http://i.a.cnn.net/money/.element/img/1.0/data/mk_snapshot/BigCharts_div.gif - NONE/- text/html
1168300933.668 5 55.156.163.104 TCP_REFRESH_MISS/504 1673 GET http://i.a.cnn.net/money/.element/img/1.0/data/mk_snapshot/BigCharts_div.gif - NONE/- text/html
```

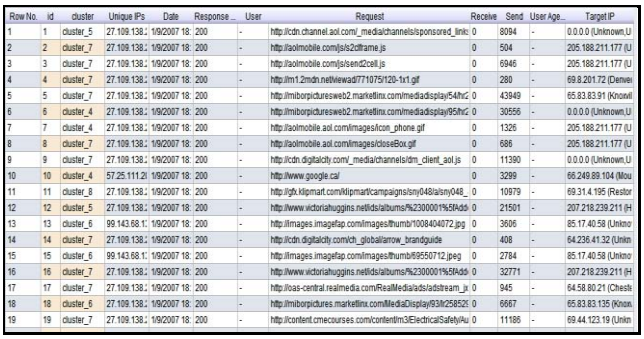
Figure 2: sample proxy server log record

After preprocess the data which is done in last section with the help of proxy explorer tool in which we filter all record rather than 200 response code. 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. Figure 3shows the data view of the cluster which shows the different IP address, request by users cluster wise. Figure 5 shows the plot view of the 10 cluster which have the ten different colors according to unique IPs. In plot view of cluster we show only two attribute at *x* axis using unique IPs and *y* axis send attribute.Next step of the proposed framework is the preprocessing. In this step, we clean the data. Figure 3 shows log records before the preprocessing operation and details of the log record which consist the data file "*pa.sanitized-access.20070109.gz*".

Unique IPs	Date	Response	User	Request	Receive	Send	User Agen	Target IP
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://cdn.channel.aol.com/_media/channels/spc	0	8094	-	0.0.0.0 (Unknown,Unknown,Unknown)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://aolmobile.com/js/s2iframe.js	0	504	-	205.188.211.177 (Unknown,Unknown,United States)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://aolmobile.com/js/send2cell.js	0	6946	-	205.188.211.177 (Unknown,Unknown,United States)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://m1.2mdn.net/viewad/771075/120-1x1.gif	0	280	-	69.8.201.72 (Denver,Colorado,United States)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://miborpicturesweb2.marketlinx.com/mediac	0	43949	-	65.83.83.91 (Knoxville,Tennessee,United States)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://miborpicturesweb2.marketlinx.com/mediac	0	30556	-	0.0.0.0 (Unknown,Unknown,Unknown)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://aolmobile.aol.com/images/icon_phone.gif	0	1326	-	205.188.211.177 (Unknown,Unknown,United States)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://aolmobile.aol.com/images/closeBox.gif	0	686	-	205.188.211.177 (Unknown,Unknown,United States)
27.109.138.213 (Unknow	1/9/2007 18:21	200	-	http://cdn.digitalcity.com/_media/channels/dm_(	0	11390	-	0.0.0.0 (Unknown,Unknown,Unknown)
57.25.111.203 (Unknown	1/9/2007 18:21	200	-	http://www.google.ca/	0	3299	-	66.249.89.104 (Mountain View,California,United States)

Figure 3: Data file log records after preprocessing

After preprocess the data which is done in last section with the help of proxy explorer tool in which we filter all record rather than 200 response code. 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.



Row No.	id	cluster	Unique IPs	Date	Response ...	User	Request	Receive	Send	User Age...	Target IP
1	1	cluster_5	27.109.138.2	1/9/2007 18:	200	-	http://cdn.channel.aol.com/_media/channels/sponsored_links	0	8094	-	0.0.0.0 (Unknown,U
2	2	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://aolmobile.com/js/s2dframe.js	0	504	-	205.188.211.177 (U
3	3	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://aolmobile.com/js/send2cell.js	0	6946	-	205.188.211.177 (U
4	4	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://m1.2mdn.net/viewad/771075/120-1x1.gif	0	280	-	69.8.201.72 (Denver
5	5	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://mibonpicturesweb2.marketlinx.com/mediadisplay/54/hr2	0	43949	-	65.83.83.91 (Knowl
6	6	cluster_4	27.109.138.2	1/9/2007 18:	200	-	http://mibonpicturesweb2.marketlinx.com/mediadisplay/95/hr2	0	30556	-	0.0.0.0 (Unknown,U
7	7	cluster_4	27.109.138.2	1/9/2007 18:	200	-	http://aolmobile.aol.com/images/icon_phone.gif	0	1326	-	205.188.211.177 (U
8	8	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://aolmobile.aol.com/images/closeBox.gif	0	686	-	205.188.211.177 (U
9	9	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://cdn.digitalcity.com/_media/channels/dtm_client_aol.js	0	11390	-	0.0.0.0 (Unknown,U
10	10	cluster_4	57.25.111.2(	1/9/2007 18:	200	-	http://www.google.ca/	0	3299	-	66.249.89.104 (Mou
11	11	cluster_8	27.109.138.2	1/9/2007 18:	200	-	http://gfx.klipmart.com/klipmart/campaigns/snj048/alsnj048_	0	10979	-	69.31.4.195 (Restor
12	12	cluster_5	27.109.138.2	1/9/2007 18:	200	-	http://www.victoriahuggins.net/lds/albums/%2300001%5Fdds	0	21501	-	207.218.239.211 (H
13	13	cluster_6	99.143.68.1(	1/9/2007 18:	200	-	http://images.imagetap.com/images/thumb/1008404072.jpg	0	3606	-	85.17.40.58 (Unkno
14	14	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://cdn.digitalcity.com/ch_global/arrow_brandguide	0	408	-	64.236.41.32 (Unkn
15	15	cluster_6	99.143.68.1(	1/9/2007 18:	200	-	http://images.imagetap.com/images/thumb/69550712.jpeg	0	2784	-	85.17.40.58 (Unkno
16	16	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://www.victoriahuggins.net/lds/albums/%2300001%5Fdds	0	32771	-	207.218.239.211 (H
17	17	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://oas-central.realmedia.com/RealMedia/ads/adstream_jx	0	945	-	64.58.80.21 (Chesh
18	18	cluster_6	27.109.138.2	1/9/2007 18:	200	-	http://mibonpictures.marketlinx.com/MediaDisplay/93/r25852(	0	6667	-	65.83.83.135 (Know
19	19	cluster_7	27.109.138.2	1/9/2007 18:	200	-	http://content.cmecourses.com/content/m3ElectricalSafety/Au	0	11186	-	69.44.123.19 (Unkn

Figure 4: Data view of different cluster

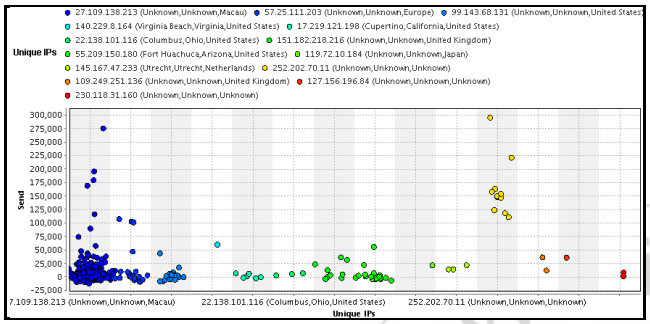


Figure 5: Plot view of data in different clusters

6. Conclusion

Web prefetching has been researched for years with the aim of reducing the user perceived latency; however, studies mainly focus on prediction performance rather than on the user's point of view. This paper work has shown that a fast and accurate prediction is crucial for prefetching performance, but there are more factors involving the user's benefit. We have described and solved the main limitations when evaluating the performance from the user's perspective. To fairly evaluate web prefetching techniques, an experimental framework has been developed. By simulating the whole web architecture, all performance issues were measured.

7. Future Work

Despite the large amount of topics about web prefetching with which this paper deals, there still exist several scenarios to be explored. In this subsection, we describe some of them, showing new challenges and expected interesting results. Through this paper, we explored the prefetching algorithms for relative small traffic and object traffic increases. For those servers in which the increase of requests is not problem a more aggressive policy should be adopted. In such a case, the prediction algorithms should be probably redesigned to consider a wider range of relation between object requests.

References

[1] Khalil, Faten (2008) Combining web data mining techniques for web page access prediction. [Thesis (_PhD/Research).
[2] Charu C. Aggarwal, Joel L. Wolf and Philip S. Yu. Caching on the World Wide Web. IEEE Transactions on Knowledge and Data Engineering, vol. 11, no. 1, pages 95–107, 1999.
[3] 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.
[4] 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.
[5] Cooley R., Mobasher B., and Srivatsava J., "Web Mining: Information and Pattern Discovery on the World Wide Web." ICTAI'97, 1997.
[6] Podlipnig S, Boszormenyi L. A survey of Web cache replacement strategies. ACM Comput Surveys 2003;35(4):374–98.
[7] W.-G. Teng, C.-Y.Chang, and M.-S. Chen. Integrating web caching and web prefetching in client-side proxies. IEEE Transactions on Parallel and Distributed Systems, 16(5):444– 455, May 2005.
[8] B. Wu and A. D. Kshemkalyani.Objective-optimal algorithms for long-term Web prefetching.IEEE Transactions on Computers, 55(1):2–17, 2006.
[9] Farhan, Intelligent Web Caching Architecture, Faculty of Computer Science and Information System, UTM University, Johor, Malaysia, 2007.
[10] E. Markatos and C. Chronaki. A top-10 approach to prefetching on the web. In Proceedings of the INET Conference, 1998.

Author Profile

Nitish Patil received B.E. degree from Nagpur University in 2009.Now He is doing M. Tech from VGI G. B. Nagar, AKTU Lucknow. He is doing his Dissertation in Web Mining.

Mahesh Kumar Singh received his B.Tech from UPTU Lucknow and he did his M.Tech. from Jamia Hamdard University. He is pursuing his Ph.D. from Kota University and doing research on study of web mining and its application to web business intelligence.