International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

An Assessment of Web Scenario for Efficient Pre-fetching Through Big Data Analysis

MR. ANIL NAYAK, MR. SHIVENDRA DUBEY, MR. UMESH JOSHI, MR. MUKESH DIXIT

COLLEGE RADHARAMAN ENGINEERING COLLEGE, BHOPAL (M.P.)

143anil.nayak@gmail.com, shivendra.dubey5@gmail.com

ABSTRACT: Www is vast, diverse, and dynamic and having increases scalability, temporal data. The rise of Www has given rise to a large quantity of data as big data that is now available for user access. Different types of data must be managed and organized so that can be accessed by different users effectively and efficiently. Web application is most common used application all over the world in order to perform communication. There are several challenges in web applications such as security, time and space, and so on. With respect to time, the web server processes the request and then generates the response to the client. In this period of time, the Web preselects one of the best concepts to make the Web application more efficient. This paper deliberates about the web pre-fetching over numerous methods. Web caching is a famous approach for refining the performance of Web based system by retaining web objects that can be used in the near future.

Keywords: WWW, Web Mining, Web Prefetching, Web Caching, Web Response time.

I. INTRODUCTION

Rapid growth in Web application has inspire the researcher's. Everyone is surrounded by a computer network. A Web Application is a very useful application used to communicate and transfer data. An application that is retrieved over a web browser is called the Web application network. Web caching is a known approach for refining the performance of a Web system by saving web articles that can be habited in the forthcoming article. Web cache method are applied at three levels: client level, proxy level, and origin server level [1,2]. Expressively, proxy servers make users and websites useful by reducing user response time and network bandwidth. Therefore, to get a better response time, an effective caching policy must be created on a proxy server. Web caching and pre-restoration are the best approach for refining the concert of the Web by observance web objects that can be attended. Web caching can exertion autonomously or in combination with previous Web recovery. Web-based pre-restoration and caching can complement each other because Web caching abuses the temporal location to predict the requested Revisited objects, whereas preloading site uses spatial localization to anticipate web objects together in relation to The requested web objects [1] preloading is used as an attempt to delete the required cached data has the following advantages: Reduced latency, less bandwidth consumption, reduces the load web server. Preloading is a way of anticipating likely future demands and bringing the most likely documents before being actually requested. It is speculative a recovery of a resource from cache memory for expectation in the near future, thereby decreasing the loading time object [3]. Web designer cache this generally transparent to the sender and the application were then, with the exception of the potential for improved response time [4]. The web developer, when forecasting a system development, will not have sufficient material for review whether web cache is include or not. In addition, if this developer is not aware of the network protocols, then developer try to focus on the application functionality, i.e., the interface between the programming language and databases, and the aggregation of the predefined response pages [5] will be highlighted.

II. WEB PRE-FETCHING

Pre-fetching of web is highly effective technique, which is used to complement the web caching

mechanism. The web Pre-fetching forecasts the web article which is probable to be entertain in upcoming web request, but these objects are not yet requested by users as shown in figure 1. Then, the expected objects are obtained from the source server and stored in the cache. Therefore, pre-fetching of the web helps to increases the cache access time and reduces user response time [5]. When it originates to the difficult like access latencies web pre-fetching methods is used to solve such kinds of problem. Particularly, comprehensive caching methods that are distributed among users work quite well. However, the increasing trend to generate dynamic pages in response to HTTP requests for users makes them highly ineffective. Cache provides the following benefits: Reduce response time, lower bandwidth consumption, and lower web server load. The previous set is a way of anticipating potential future orders and searching for the most probable documents, before actually requesting them. It is a speculative recovery of a data in the cache memory as future expectation, thus reducing the loading time of the object [4,6].

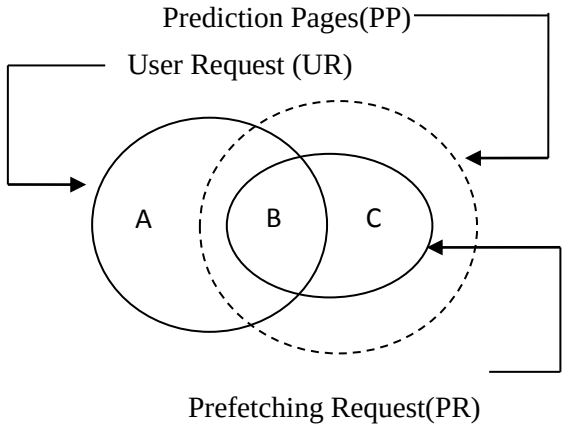

Figure 1: Web Prefetching Request Query

Pre-fetching procedures can be employed on the server, proxy, or client side. Client based pre-fetching focuses on the browsing patterns of a single user on many web servers. On the other hand, server-based pre-fetching focuses on browsing patterns for all users who access a single web site [5]. Proxy-based pre-fetching focuses on the browsing patterns of a group of users on many web servers. Therefore, this approach may reflect a shared interest of the user community. In other words, the content of pre-fetching can be shared by many users. Table 1 summarizes pre-fetch types by location [4,5].

Pre-fetching is categorized in the following two types:

Link pre-fetching is a prefetching approach in which web page express the upcoming expectation of end user, hence the respected browser should directly respond that particular.[6]

Table 1: Comparison in Pre-Fetching Location

Pre-Fetching Location	**Data for Prediction Model**	**Advantages**	**Disadvantages**
Client	Historical and current user requests	Easy to partition user Session and realize Personalized pre-fetching.	1. Not share pre-fetching content among users. 2. Needs a lot of Network bandwidth.
Proxy	Proxy log and current user requests	1. Reflects common Interests for a group of Users. 2. Shares pre-fetching Content from different Servers among users	Not reflect common interests for a single Website from all Users
Server	Server log and current user requests	Records single website access information from all users and better reflect all users' common interests.	1. Not reflect users' real Browsing behavior. 2. Difficult to partition User session. 3. Needs additional communications between clients and servers for deciding Prefetching content

DNS pre-fetching is where the browser tries to speed up future requests by resolving the IP address of every link on webpages which is visited by the user.

III. BIG DATA

Big Data The quantity of data produced daily in the explosion of the world. The growing size of digital media and social media and the Internet of Things, nourishes going further. Data growth is amazing, this data comes quickly, with a variety (and not necessarily regulated) and contains a wealth of information which can be a key to getting an edge in competing companies. "Big Data is an assembly of very large data sets and composite that it transforms challenging to treat using traditional management databases or processing tools application data. The challenges in the areas of capturing, preservation, storage, search, sharing, transfer and analysis, and visualize these [3] data". The world has been immersed in a sea of data today. In an extensive variety of application areas, data is collected on a scale never seen before.

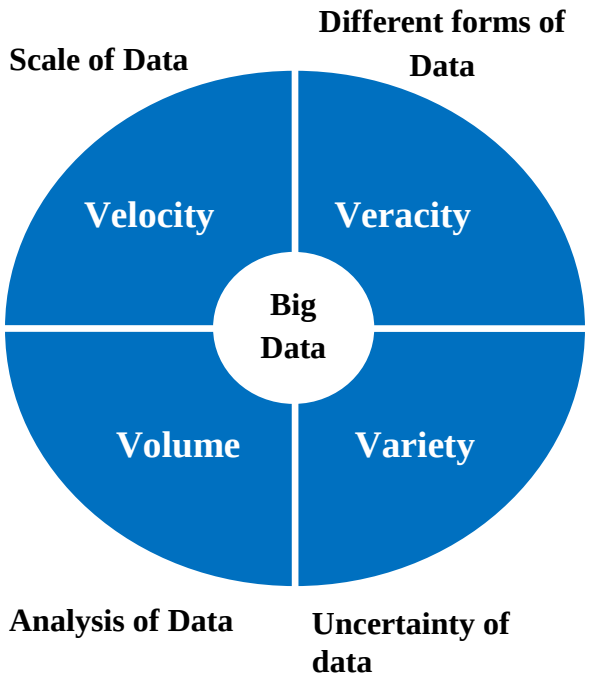

Figure 2: Big data 3V's

Decisions grounded on the above assumptions, or the carefully constructed reality models, can now be done on the foundation of the same data. The analysis of large data now covering nearly all features of current society, including mobile services, retail, manufacturing, financial services, life sciences and physical sciences. Big data is a tenure that defines an enormous amount of data, both formal and informal, used daily. But this does not mean that it is organizations: huge quantities of records can be analyzed in search of ideas that clue to enhanced judgments and planned actions [2]. Traditional data in addition to many types of new data and data management. Despite the strong demand for high analytics for data, there is currently a deficiency of scientific data and other analysts who have experience working with large volumes of data in a scattered environment, open code as shown in figure 2.

In the enterprise, suppliers have replied to this lacking by creating Hadoop features to help businesses take benefit of semi structured and unstructured data they possess. Large volumes of data can be compared with small data, moving another term often used to describe format data volume and can easily be used for self-analysis. An axiom is often quoted: "data is important for machines, data is for small people". The significance of big data does not necessarily have to change, but what we do with it. We can gross data from some source and examine it to find a reaction that allow 1) cost savings, 2) time savings, 3) development of new products and offers optimized and 4) intelligent decision-making.

IV. RELATED WORK

Many ways have been developed to enhance the efficiency of web servers, which include hardware optimization (speed, bandwidth) and software explanations (more convenient models, protocols, better algorithms) [10,11]. A normally procedure and operative method is the preloading of certain records to the resident cache previously the user is expected to request this data in the forthcoming request so that it is voluntarily presented locally rather than from remote sites. Of course, the preload process is recovery from distant sources, but this can be done without the superficial interruption from the user's opinion, simply because there is always a time interval between consecutive requests from the same user in the web environment and server The web can use this time frame to fetch pages Previously anticipated [1,2,7]. Successful pre-participation will not only limit delays in demands for web objects by users, but will also reduce overall web steaming and load on web servers.

Saldhi et.al [11] dealed with important data that may come from different sources in different formats to run simultaneously on a Hadoop cluster, and use of the technical proposal and carry results effectively. The author has applied the methodology proposed in the preliminary data for the industrial society to the intelligence business, with the final objective of finding the profits generated by the company and

trends throughout the year. The author did data analysis, which lasted a full year trends are repeats year after year. W. Premchaiswadi et.al [12] provided a framework for the exploitation of new and effective web log for users of online groups. In general, our framework contains three main steps. 1) Calculating similarity measure in a track in a Web page, 2) identify as a approach to assemble a client group 3) generate a report grounded on the Hadoop framework Map Reduce .S. Narkhede [13] shows that the Map Reduce has been extensively useful in numerous computing and data intensive applications and fields is also a significant programming for the cloud computing model. Hadoop is an open source application of Map Reduce working on terabytes of data using basic equipment. This model has a flat Map Reduce programming Hadoop to analyze log files on the Internet so the author can be hit by a number of specific Web applications. This system uses Hadoop to accommodate the log file and the outcomes are estimated using the card and cut jobs. Tentative outcomes demonstration a number of successes for each field in the log file.

Table 2: Summery of Related Work

Paper Title	**Approached used**	**Merits**	**Demerit**
An Improved Web Cache Replace-ment Algorithm Based on Weight-ing and Cost[8]	• Propose a novel, high-performance cache replacement algorithm for the web cache, named weighting size and cost replacement policy (WSCRP) bases on the weighting replacement policy. • When the cache space cannot satisfy the new request object, the replacement policy WSCRP replaces the largest weighting and cost object.	Work over influence of various factors on the Web object as frequency, time, and cost value are considered.	Cache inefficiency degrades the web pre fetching performance.
Toward Rend-erring Latency Reduction for Compostable Web Services via Pri-ority-based Object Caching[14]	• Use the stochastic optimization framework, and decompose the problem into a set of one-shot optimization problems, proved to be NP-hard. • Finally, they integrate the resulted approximation algorithms into an online algorithm.	This approach use two greedy algorithms, with different computation complexity and the same performance bound. .	High latency, non-priority-based Caching strategy, non-spectrum allocation method that enhance the service latency.
Semantic-rich Markov Models for Web Pre-fetching[9]	Use semantic information as a criteria for pruning states in higher order selective markov models and compare the accuracy and model size of this idea with semantic-rich markov models and with traditional markov models .	Use semantic-rich infor-mation with markov models that reduce latency	Lower priority based caching strategy.
Analyzing web application log files to find hit count through the utilization of Hadoop Map Reduce in cloud computing environment[13]	This model has a flat MapRedece programming Hadoop to analyze log files on the Internet so the author can be hit by a number of specific Web applications.	This system uses the Hadoop file system to store the log file and the results are evaluated using the card and cut jobs.	Use unstructured log file having redundancy, noise and high dimension data.

Tinghuai Ma et.al [15] presented several models to designate the critical-object aware caching scheme, and formulated it as a constrained optimization problem. Using the stochastic optimization framework, we decomposed the problem into a set of the one-shot optimization problems. After deriving two estimate method for the one-shot optimization problems, author's developed an online algorithm with performance bound. Through real-trace driven simulations, we verified that their procedure could moderate the initial rendering time of web pages, improve the cache hit ratio and reduce the network traffic. Han Hu [14] proposes a new high recital cache replacement algorithm for the web cache, called WSCRP, grounded on the weight replacement policy. The algorithm recalculates the weight of the objects by adding the cost attribute in the cache and then orders the weight. In addition, the influence of several factors on the Web object is measured as a frequency, a duration and a cost value. When the cache space cannot satisfy the new request object, the replacement policy is WSCRP replacing the most significant weighting and cost object.

V. PROBLEM IDENTIFICATION

Web is a key resource in order to share the information along the world. It has large number of news, advertisements, global connectivity between people and lots of knowledge for the students. This massive use of Web or WWW makes it more important in the world of research. Researcher has the challenge to make the web applications more efficient. Many researchers work on it and give new idea in order to give the better results from the previous one. There is a huge need to improve the response time of server for web applications as shown in table 2. Current Web has a massive repository due to increase its use suddenly. It has to focus on both the quality and quantity of web contents. Even, when the speed of Internet has improved with the reduced costs, the traffic is getting heavier. The enormous information makes it difficult to find the relevant information quickly. This led to the effort to improve the speed, by reducing the latency, make the web more relevant and meaningfully connected.The Cache prefetching plays an important role in order to enhance the response time and make the application well-organized. The web prefetching is a technique in order to preprocess the user requests, before they are actually demanded. Therefore, the time that the user must wait for the requested documents can be reduced by hiding the request latencies. Prefetching is the method for reducing Latencies. The user always expects an interactive response, better satisfaction and quality of output.

VI. CONCLUSION

Investigation of Web log is an inventive and exclusive territory that continuously designed and revised by the merging of numerous emerging web technique. Because of its interdisciplinary nature, the diversity of issues addressed, the variety and the number of Web applications, is subject to many different methodologies and different research. Log file is a crucial party of web application. In this manner the log analysis is also plays an important role in the various applications and treats as big data.

VII. REFERENCE

[1] K.Ramu, Dr.R.Sugumar and B.Shanmugasundaram "A Study on Web Prefetching Techniques" Journal of Advances in Computational Research: An International Journal Vol. 1 No. 1-2, January, 2012

[2] Waleed Ali, Siti Mariyam Shamsuddin, and Abdul Samad Ismail "A Survey of Web Caching and Prefetching", Int. J. Advance. Soft Comput. Appl., Vol. 3, No. 1, March 2011

[3] Daesung Lee and Kuinam J. Kim, "A Study on Improving Web Cache Server Performance Using Delayed Caching", IEEE 2010, pp 1-5.

[4] B. Nigam and S. Jain, "Analysis of Markov model on different web Prefetching and caching schemes," Computational Intelligence and Computing Research (ICCIC), 2010 IEEE International Conference on, Coimbatore, 2010, pp. 1-6.

[5] Junchang Ma and Zhimin Gu, "Finding Shared Fragments in Large Collections of Web Pages for Fragment-Based Web Caching," Fifth IEEE International Symposium on Network Computing and Applications (NCA'06), Cambridge, MA, 2006, pp. 251-254..

[6] B. D. Davison, "A Web caching primer," in IEEE Internet Computing, vol. 5, no. 4, pp. 38-45, Jul/Aug 2001.

[7] G. Barish and K. Obraczke, "World Wide Web caching: trends and techniques,"

in IEEE Communications Magazine, vol. 38, no. 5, pp. 178-184, May 2000.

[8] T. Ma, Y. Hao, W. Shen, Y. Tian and M. Al-Rodhaan, "An Improved Web Cache Replacement Algorithm Based on Weighting and Cost," in IEEE Access, vol. 6, pp. 27010-27017, 2018.

[9] N. R. Mabroukeh and C. I. Ezeife, "Semantic-Rich Markov Models for Web Prefetching," 2009 IEEE International Conference on Data Mining Workshops, Miami, FL, 2009, pp. 465-470.

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

[11] Saldhi, D. Yadav, D. Saksena, A. Goel, A. Saldhi and S. Indu, "Big data analysis using Hadoop cluster," Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on, Coimbatore, 2014, pp. 1-6.

[12] W. Premchaiswadi and W. Romsaiyud, "Extracting weblog of Siam University for learning user behavior on MapReduce," Intelligent and Advanced Systems (ICIAS), 2012 4th International Conference on, Kuala Lumpur, 2012, pp. 149-154.

[13] S. Narkhede, T. Baraskar and D. Mukhopadhyay, "Analyzing web application log files to find hit count through the utilization of Hadoop MapReduce in cloud computing environment," IT in Business, Industry and Government (CSIBIG), 2014 Conference on, Indore, 2014, pp. 1-7.

[14] Han Hu, Yuanlong Li, and Yonggang Wen"Toward Rendering-Latency Reduction for Composable Web Services via Priority-based Object Caching" IEEE Transactions on Multimedia, 2017

[15] Tinghuai Ma et.al "An Improved Web Cache Replacement Algorithm Based on Weighting and Cost" IEEE Access , 201