International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-3 Issue-2, December 2013

Proposed Framework for the Reduction of Web Congestion using Classification

Neha Kohli, Esha Dobhal, Neha Sharma

Abstract— Prefetching is the process of bringing data from the web server into the web cache before it is needed. When the client needs data, then instead of waiting for the responses from the memory, it can directly access the data from the cache. The prefetched data is stored in web cache in the form of web objects for later use. Caching is the technique of storing a copy of the data that has been requested by the client. Web caching is mainly used to reduce access latency, that is, it speeds up the process of data retrieval. It also reduces heavy load on the web server. The paper proposes a framework for reducing web traffic. The data is first extracted from the proxy server and then preprocessing is performed. The preprocessed data is then classified and the patterns to be prefetched are obtained.

Index Terms—Prefetching, classification, proxy server, cache.

I. INTRODUCTION

With the advent in technology, the web has grown as a rich collection of dynamic and interactive services. This huge growth of web has resulted into an increasing load on the web servers. Clients experience delay from remote servers whenever there is a need to retrieve data. We can overcome this problem by increasing the bandwidth but this leads to increased cost. Caching is hence used to reduce cost and to enhance the performance. A web cache is a technique used for storing the data in the form of web objects which reduces the load on the server and improves the access latency. Cache is used for storing the recent and frequently accessed data.

Proxy server is used for implementing web cache. Reverse Proxy is a service between a client and a server where the request is sent by the client to the proxy. It acts as a forwarding service which connects the server and exchanges data between the client and the server. At each request of the client, the proxy is first contacted to check whether it has a valid copy of that particular object [2]. A cache hit occurs if the requested object is found in the proxy, otherwise a cache miss occurs. In the latter case, the request of the client is forwarded to the web server. The proxy services keep a copy of a new object in the form of a temporary web object.

The patterns and the requests of a specific web user are same most of the time at a specific instant. Some web documents are also more accessed as compared to others. Some of the popular servers have most of the load as the load is not distributed uniformly.

Therefore the common object is stored in the web cache and hence the latency of the server is reduce efficiently. User web requests patterns are required to be pre-fetched to improve the web performance, for which user's next request pattern needs to be known previously. This framework uses the SVM algorithm for classification in the section 1. SVM is a supervised learning method used for classification and regression analysis. It is a classification algorithm that learns by example to classify objects into groups (two classes, relevant and non-relevant). SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable.

In this paper, a framework is proposed to support the prefetching criteria on web servers. According to the framework, there are basically five steps to predict and prefetch the user's requests, viz. data extraction from proxy web log, data preprocessing, data classification and lastly prefetching. The proposed framework has few merits over the previously used techniques.

The paper is divided into different sections for the ease. Section 2 provides an overview of primarily research done in this field. A lot of work has been done previously in this field and this part of paper explains the same. Section 3 concentrates on the proposed framework given in the paper. It's further divided into subsections viz. preprocessing, classification and prefetching. In section 4, merits and demerits of the proposed technique are explained. Finally, the conclusion of the paper has been written, explaining the work done in brief with the merits and scope of the work in the present scenario.

II. LITERATURE REVIEW

Hongiun *et.al* [1] presented various classification algorithms that have been designed to tackle the problem by researchers in different fields such as mathematical programming, machine learning, and statistics. He primarily focused on neural networks and addressed their merits and demerits.

Pallis *et.al* [2] addressed the short-term prefetching problem on a Web cache environment using an algorithm (clustWeb) for clustering inter-site web pages. The proposed scheme efficiently integrates Web caching and prefetching. According to this scheme, each time a user requests an object, the proxy fetches all the objects which are in the same cluster with the requested object.

Podlipnig *et.al* [3] provides an exhaustive survey of cache replacement strategies proposed for Web caches.

Manuscript published on 30 December 2013.
* Correspondence Author (s)
Neha Kohli*, Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India
Esha Dobhal, Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India
Neha Sharma, Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India

The paper concentrated on proposals for proxy caches that manage the cache replacement process at one specific proxy. A simple classification scheme for these replacement strategies was given and used for the description and general critique of the described replacement strategies. Although cache replacement is considered as a solved problem, we showed that there are still numerous areas for interesting research.

Steven W. Norton *et.al* [4] described the IDX algorithm for classifier problems. This algorithm generated better decision trees. Decision trees are a popular representation for classifiers. The interior nodes of a decision tree are tests applied to instances during classification. Branches from an interior node correspond to the possible test outcomes. Classification begins with the application of the root node test, its outcome determining the branch to a succeeding node.

Deborah R, Carvahlo *et.al* [5] addressed the classification task of data mining whereby the discovered knowledge was expressed in the form of high-level, easy-to-interpret classification rules. In order to discover classification rules a hybrid decision tree/genetic algorithm method was proposed. The central idea of this hybrid method involved the concept of small disjuncts in data mining where a set of classification rules was regarded as a logical disjunction of rules, so that each rule represented a disjunct. In this hybrid approach, two genetic algorithms (GA) specifically designed for discovering rules covering examples belonging to small disjuncts, whereas a conventional decision tree algorithm was used to produce rules covering examples belonging to large disjuncts.

Sharma and Dubey *et.al* [6] provided the literature survey in the area of web mining. The paper basically focuses on the methodologies, techniques and tools of the web mining. The basic emphasis is given on the three categories of the web mining and different techniques incorporated in web mining.

Sharma and Dubey *et.al* [7] used the concept of web caching and prefetching to reduce web congestion. They introduced Fuzzy C means clustering and Prediction by Partial Matching (PPM) algorithm to refine the process of classification. Lou *et.al* [8] investigated the problem of user transaction identification in proxy logs. In a proxy logs, a single user transaction may include pages references from one site as well as from multiple sites. Moreover, different types of transactions are not clearly bounded and are sometimes interleaved with each other as well as with noise. Thus an effective transaction identifier has to identify interleaved transactions and transactions with noise, and capture both the intra-site transactions and the inter-site transactions. It presented a cut-and pick method for extracting all these transactions, by cutting on more reasonable transaction boundaries and by picking the right page sequences in each transaction.

Sathiyamoorthi *et.al* [9], authors discusses various data preprocessing techniques that are carried out at proxy server access log which generate Web access pattern and can also be used for further applications. Web access patterns are basically used for prediction and prefetching.

Greeshma *et.al* [10], web prefetching techniques and other directions of web prefetching are analyzed and discussed. These techniques are applied to reduce the network traffic and improve the user satisfaction. Web prefetching and caching can also be integrated to get better performance.

III. THE PROPOSED FRAMEWORK

Web caching is used to reduce network traffic and congestion by caching the web at proxy. The paper presents a framework for a prefetching scheme so as to improve the web performance. The prefetching scheme interprets the user's request depending upon their previous access behavior [7]. The previous access records of the user are extracted from reverse proxy log data and this extracted data is then preprocessed. Thereafter, the preprocessed data is classified using the Support Vector Machine (SVM) algorithm [11, 12]. SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. One disadvantage of many classification techniques is that the classification process is difficult to understand so the SVM algorithm has been devised as replacement for them. Hence, depending upon this prediction, web documents will be pre-fetched in the proxy server cache, so as to improve the cache hit ratio.

Improvement of cache hit ratio will lead to decrease in the latency and would also decrease web traffic. The architecture of the proposed framework has been shown in figure 1 and figure 2 shows the pseudocode for the framework steps.

A. Data Extraction

Firstly, data from reverse proxy web log is extracted, as the web log keeps the user records for their previous requests. Data extraction is the process of retrieving data from data sources for preprocessing.

The preprocessed data is then classified and certain access patterns are hence obtained. Depending upon these access patterns of users, prediction for next requests of users will be done. Hence, data is extracted very carefully from the proxy web log so that accurate information about the user access behavior is taken [7].

B. Data Preprocessing

Preprocessing means removal of noise and irrelevant information from the data set present. Preprocessing is important before applying clustering on the extracted data. Pre-processed data improves the efficiency and facilitates ease of mining process [13].

Data preprocessing is not a simple task, it consists of basically 4 important steps as explained below:

First step, data cleaning, is the process of cleaning the data entries by filling the missing values, removing inconsistencies and smoothing the noisy data.

Second step is data integration. Data might be taken from multiple resources for analysis or prediction, as used in the framework scenario. So, data taken from multiple resources needs to be clumped together into a single file and this process of clumping data from all the sources is known as data integration.

Third step is data transformation of pre-processed data. Data transformation is the process of transforming the data into the forms relevant for the mining process.

Fourth Step, data reduction is done. In data reduction, data set is reduced in such a manner that after and before reduction, the data set produces the same analytical results [7].

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

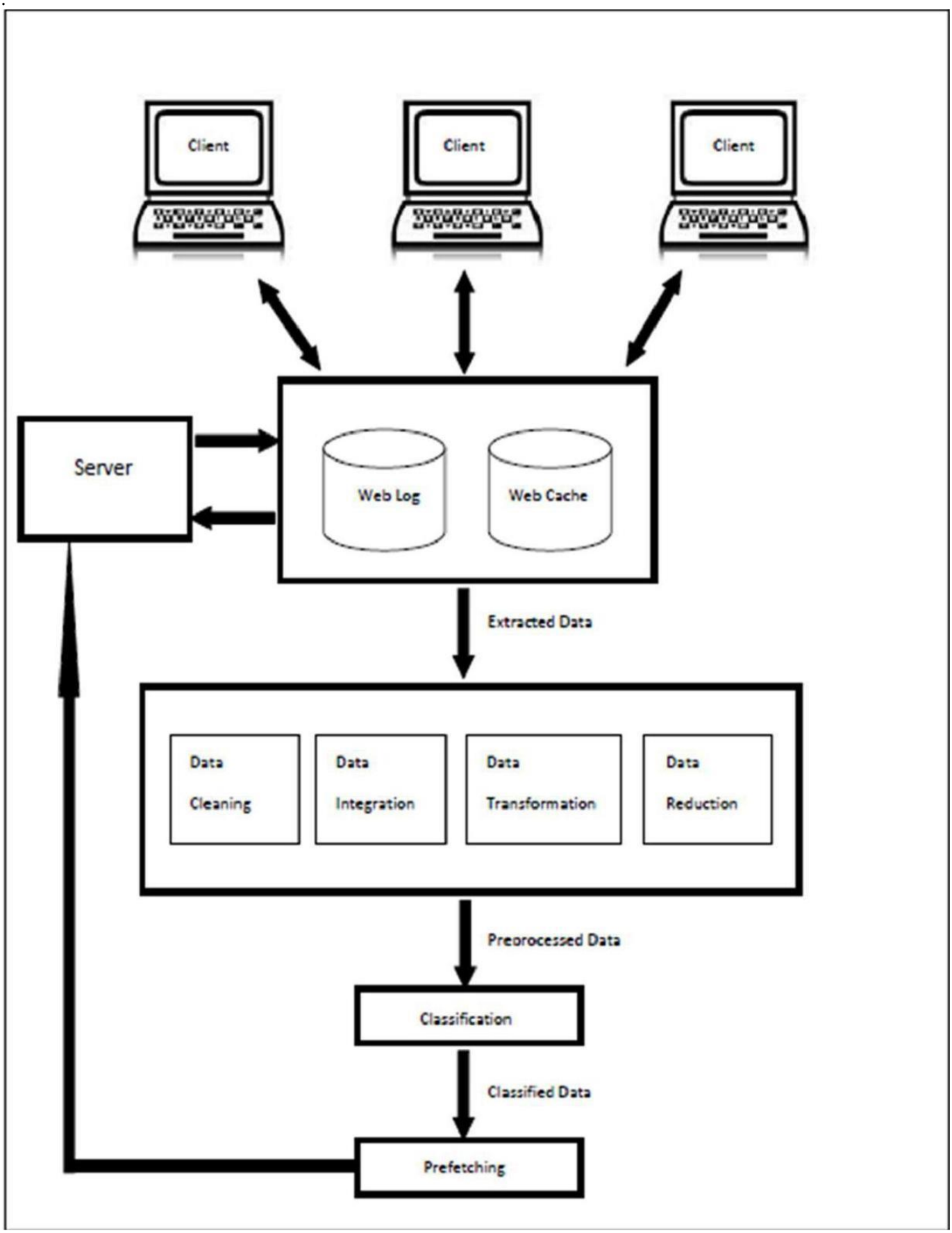

Figure 1: The proposed framework

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

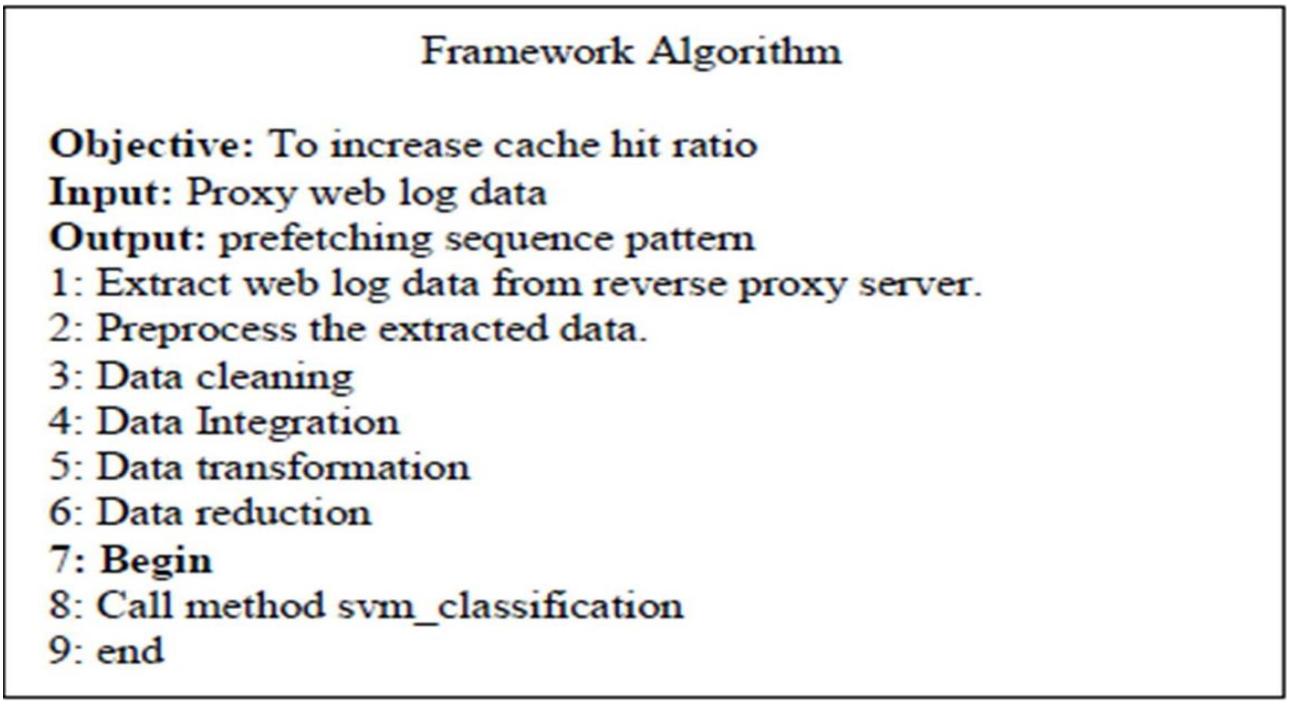

Figure 2: Pseudo code of svm based classification Algorithm

C. Support Vector machine for classification

The framework uses support vector machine algorithm for classification as the next step. It takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other.

A training set of patterns is separable if there exists at least one linear classifier defined by the pair (w, b) which correctly classifies all training patterns. SVM makes use of kernel functions in order to transform the problem. In this way we can apply linear classification technique to non-linear classification data. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible [12]. The pseudocode for the algorithm is also being provided in figure 3.

D. Web Prefetching

According to the framework, prediction is done using classification and no other algorithm is required for prediction. Web prediction will let cache know, what might be the requests of the users in near future, then accordingly data will be maintained in the cache and as a result number of cache hits will be improved. After predicting the user's request patterns, the prefetching is done. The web cache prefetches the patterns and provides it to the user [9].

These patterns hence result in ease of availability of the data that has already been accessed.

This paper shows how prefetching combined with classification increases the speed of data retrieval and thus reduces web congestion.

IV. MERITS OF THE FRAMEWORK

The base of the framework is web cache and support vector machine based classification algorithm. There are a number of advantages of using these in this framework, which actually are the merits of the proposed approach:

- It reduces bandwidth consumption resulting in reduction in network traffic and congestion.
- Access latency is also reduced.
- It disseminates the load of web server among proxies and leads to load reduction of the remote server.
- If the remote server is not available due to remote server's crash or network partitioning, it can access a cached copy at proxy. Thus, the robustness of the web service is enhanced [11].
- It helps to analyse the access patterns of the organisation.
- Support vector algorithm is effective in high dimensional space and it uses a subset of training points in the decision function, so it is also memory efficient.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

SVM_CLASSIFICATION

Given some training data $\mathcal{D}$, a set of n points of the form

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n}$$

where the y_i is either 1 or −1, indicating the class to which the point $\mathbf{x}_i$ belongs. Each $\mathbf{x}_i$ is a p-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of points $\mathbf{x}$ satisfying

$$\mathbf{w} \cdot \mathbf{x} - b = 0,$$

where $\cdot$ denotes the dot product and $\mathbf{w}$ the (not necessarily normalized) normal vector the hyperplane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$. If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations

$$\mathbf{w} \cdot \mathbf{x} - b = 1$$

and

$$\mathbf{w} \cdot \mathbf{x} - b = -1.$$

By using geometry, we find the distance between these two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$, so we want to minimize $\|\mathbf{w}\|$. As we also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$$\mathbf{w} \cdot \mathbf{x}_i - b \geq 1 \qquad \text{for } \mathbf{x}_i \text{ of the first class}$$

or

$$\mathbf{w} \cdot \mathbf{x}_i - b \leq -1 \qquad \text{for } \mathbf{x}_i \text{ of the second}$$

This can be rewritten as:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \quad \text{for all } 1 \leq i \leq n.$$

We can put this together to get the optimization problem:

Minimize (in $\mathbf{w}, b$)

$$\|\mathbf{w}\|$$

subject to (for any $i = 1, \ldots, n$)

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1.$$

For x, y on S, certain functions $K(x, y)$ can be expressed as an inner product (usually in a different space). K is often referred to as a kernel or a kernel function. If one is insightful regarding a particular machine learning problem, one may manually construct $\varphi : S \to V$ such that

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_V$$

and verify that $\langle \cdot, \cdot \rangle_V$ is indeed an inner product.

Figure 3: Pseudo code of svm based classification Algorithm [11, 12]

V.CONCLUSION

In this paper, the problem of network congestion and web latency has been addressed. The paper proposes an efficient approach using prefetching for the same. Cache stores the copies of documents passing through it, so that if next time request comes for the same document, it can be fulfilled directly from the cache. The paper proposes a new approach to improve cache performance. After analyzing the user access behaviour, proxy server can prefetch the next requests of the users. The paper basically changes the conventional technique by applying support vector machine algorithm for classification.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The framework proposes that caching the data and pre-fetching the upcoming requests of the user's will improve the web server performance. The technique explains that proxy web log data must be pre-processed and then classified according to the user access behaviour. At last, prefetching rules are generated. It will result in performance improvement of the web via web latency reduction; cache hit ratio improvement; faster response of user requests and web traffic reduction. Application
of this framework in the real world will actually improve the cache hit ratio and congestion problems in the network.

VI. FUTURE WORK

For the future, our plan is to investigate the performance of web after applying the proposed framework and its implementation on proxy servers. Further, comparing the proposed approach with the present scenario and trying out other classification algorithms for further improvement of the web, are in the future scope. Finally, extending the use of classification and prefetching in other applications such as in mobile environment, content distribution network etc. is another aim.

REFERENCES

[1] Hongjun Lu Rudy Setiono Huan Liu "Effective Data Mining Using Neural Networks**"** December 1996 Vol. 8, No. 6, pp. 957-961

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

[3] Podlipnig S. and L. Boszormenyi, "A Survey of Web Cache Replacement Strategies", ACM Computing Surveys, Vol. 35, No. 4, December 2003, pp. 374–398

[4] Steven W. Norton "Generating Better Decision Trees" Siemens Corporate Research, Inc. 755 College Road East, Princeton, NJ 08540 swn@demon.siemens.com.

[5] Deborah R. Carvalho "hybrid decision tree/genetic algorithm method for data mining" Universidad Tuiti do Parana (UTP Computer Science Dept, Av. Comendador Franco, 1860. Curitiba-PR 80215-090 Brazil, Alex A. Freitas2 Computing Laborator University of Kent at Canterbury Canterbury, Kent, CT2 7NF ,U.K.

[6] Sharma N. and S. K. Dubey, (2012) "A Hand to Hand Taxonomical Survey on Web Mining", International Journal of Computer Applications (0975 – 8887), Vol. 60, No.3.

[7] Neha Sharma and Sanjay Kumar Dubey "Fuzzy c-means clustering based prefetching to reduce web traffic", Amity University, Noida (U.P.), 201303, India International Journal of Advances in Engineering & Technology, Mar. 2013. ©IJAET ISSN: 2231-1963

[8] Lou W., G. Liu, H. Lu, and Q. Yang, (2002) "Cut-and-Pick Transactions for Proxy Log Mining", C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 88–105, Springer-Verlag Berlin Heidelberg.

[9] Sathiyamoorthi V. and Dr. M. Bhaskaran, (2011) " Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data through Proxy Server", IJCSNS International Journal of Computer Science and Network Security, Vol.11, No.11.

[10] Greeshma G. Vijayan and J. S. Jayasudha, (2012) "A survey on web pre-fetching and web caching techniques in a mobile environment" Natarajan Meghanathan, et al. (Eds): ITCS, SIP, JSE-2012, CS & IT 04, pp. 119–136.

[11] S.V.N. Vishwanathan, M. Narasimha Murty " SSVM : A Simple SVM Algorithm" , Dept. of Comp. Sci. and Automation ,Indian Institute of Science, Bangalore 560 012,INDIA

[12] Svm-http://en.wikipedia.org/wiki/Support_vector_machine.

[13] Han J. and Kamber M., (2006) "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, Elsevier.

Neha Kohli is currently pursuing Bachelor of Technology in Information Technology from Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India.

Esha Dobhal is currently pursuing Bachelor of Technology in Information Technology from Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India.

Neha Sharma is currently working as Assistant Professor in Northern India Engineering College, Guru Gobind Singh Indraprastha University, New Delhi, India. Her area of research is data warehousing and data mining.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)