To Prepare a Forecasting Proxy Server to Improve Hits Using SVM.

Suvarna Temgire, Poonam Gupta

Computer Engineering Department,
Pune University, Raisoni College of Engineering and Management,
Dhonkal Road, Wagholi, Pune, Maharashtra 412207, India

Abstract— **Web proxy cache plays a vital role in the performance improving of World Wide Web. Also, SVM helps to wither the internet access latency and network traffic. The prediction using SVM improves the hit ratio which is beneficial for Web caching and Web prefetching environment. The research mainly focuses to apply SVM method to the problem of user action for prediction on the web. We can predict the future web page that a handler will select through replication, we determined that our approach has computable actions such as hit rate and byte hit rate of accessed page. Web proxy caching plays an important role in improving the World Wide Web routine which is actually difficult to evaluate what exactly the user's request would be in web proxy caching practices. In this learning, we presented a new approach which depends on the expertise of SVM to learn from web proxy log data and forecast the classes of objects expected to re-visit. Hence, usage of the cache can be amplified capably. Continual experiments had revealed that support vector machine produces better and positive results.**

Keywords—**Enactment estimation, Proxy server, Web Proxies, Support vector machine, Web caching, Web prefetching.**

I. Introduction

This Support vector machines (SVMs, also support vector networks[14]) are supervised learning models that are associated with learning algorithms that analyse data and recognize patterns, used for classification and regression analysis. The elementary SVM 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. 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. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. As the number of World Wide Web (Web) users grows, Web traffic continues to increase at an exponential rate. Currently, Web traffic is one of the major components of Internet traffic. This will reduce Web access time and make more efficient use of Internet links. One of the solutions to reduce Web traffic and speed up Web access is through the use of Web caching [14]. However, Web caching is limited to reducing network bandwidth usage during peak periods [11, 12, 14]. In this paper, we focus on the use of prefetching, based on a caching server, for falling bandwidth during peak periods using off-peak period bandwidth. We have developed a statistical, batch, proxy-side prefetching scheme that improves cache hit rate, while only requiring a small amount of additional storage space. This prefetching scheme reflects Web access patterns of users. In fact, this scheme may increase total bandwidth usage slightly in comparison with standard Web caching. However, the proposed scheme can efficiently reduce Web traffic bandwidth usage during peak periods by consuming unused bandwidth during off-peak periods.

The most software based solution is web caching and prefetching techniques .The caching is introduced at three level client level, proxy level and original server level. Prefetching technique is used according to prediction on web proxy .Successful proxy servers play the key roles between users and web sites, which could reduce the response time of user request and save network bandwidth. Therefore, an efficient caching approach should be built in a proxy server for achieving better response time.

A. Web Proxy:

Web proxies provide a quick and easy way to change your IP address while surfing the Internet. Web proxies are extremely portable as they do not require the installation of additional software or modification to computer networking settings. They are used like a search engine, except that you enter a website visitors time we have searched the internet to establish a unique collection of proxy site listings. We will only accept clean, spam free and functional proxy websites to make our web proxy list the best around. When an internal user requests a Web page, the request goes through the proxy server so that it appears to the Internet to be coming from the server - from its IP address (or one of them) - and not the user's device. This anonymity provides an important measure of security by reducing the amount of information about a network and its users easily accessible to hackers on the Internet.

B. Web Caching

According to the locations where objects are cached, Web caching technology can be classified into three categories, i.e., client's browser caching, client-side proxy caching, and server-side proxy caching [1]. In client's browser caching,

Web objects are cached in the client's local disk. If the user accesses the same object more than once in a short time, the browser can fetch the object directly from the local disk, eliminating the repeated network latency. However, users are likely to access many sites, each for a short period of time. Thus, the hit ratios of per-user caches tend to be low. In client side proxy caching, objects are cached in the proxy near the clients to avoid repeated round-trip delays between the clients and the origin Web servers. To effectively utilize the limited capacity of the proxy cache, several cache replacement algorithms (e.g., [2], [3], [4], [5], [6]) are proposed to maximize the delay savings obtained from cache hits.

C. *Cache Prediction:*

Prediction caches use a history of recent cache misses to predict future misses, and to reduce the overall cache miss rate. This paper describes several prediction caches, and introduces a new kind of prediction cache, which combines the features of prefetching and victim caching. This new cache is shown to be more effective at reducing miss rate and improving performance than existing prediction caches.

Idea of prefetching on a history of cache miss, cache hit, predictive caches, improves the performance by reducing the overall cache miss rate. Cache is a small high-speed memory. Stores data from some frequently used addresses (of main memory). Cache hit Data found in cache. Results in data transfer at maximum speed. Cache miss data is not found in cache. Processor loads data from M and copies into cache. This results in extra delay, called miss penalty. The Cache Hit Ratio is the ratio of the number of cache hits to the number of misses, usually expressed as a percentage.

D. *Web Prefetching*

Web pre-fetching has been recognized as one of the major techniques for improving the performance of the web by latency reduction. Prefetching technique takes advantage of idle time of the network to pre-fetch the anticipated web pages [1-4]. Pre-fetching technique only induces a little burden on client and network for finding the anticipated web pages and fetching it in advance. Pre-fetching increases web traffic but reduces latency in accessing the web pages. Efficient pre-fetching techniques control the web traffic and achieve maximum pre-fetch hits. In dynamic pre-fetching technique, web traffic is controlled by intelligent agents and effective utilization of bandwidth of the existing link is achieved.

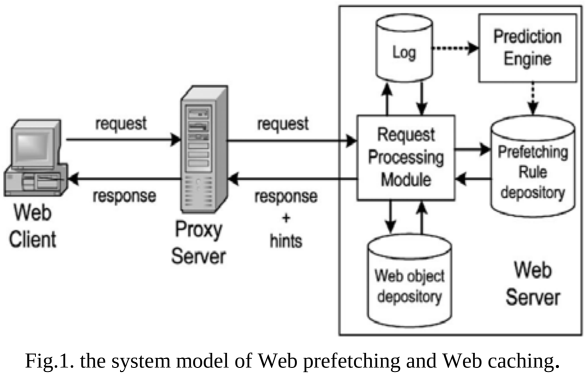

Fig.1. the system model of Web prefetching and Web caching.

E. *Categories of Prefetching Approaches*:

Web prefetching has been extensively studied. Roughly speaking, major approaches fall in the following three categories: probability based, clustering based and using weight-functions.

1) *Probability Based Prefetching*

The central problem for Web prefetching is the prediction algorithm. When a request comes, a decision needs to be made on which page would mostly likely to be requested next time. Probability based prediction is a natural approach. Probabilities are calculated using the history access data. This method assumes that the request sequence follows a pattern (is not random) and the probabilities are trying to follow this pattern. One of the advantages for this approach is the number of pages prefetched can be controlled.

2) *Weight-Function Based Prefetching*

The cost of the network traffic and server workload as the overhead of the programs was not considered. To consider factors other than just the probabilities (such as size, priority), a function that involves multiple factors is needed.

3) *Clustering Based Prefetching*

Clustering based prefetching methods make decisions using the information about the clusters containing pages that have been previous fetched, anticipating that pages that are “close” to those previously fetched pages are more likely to be requested in the near future. We shall discuss four methods in this category:

(1) Support vector machine (SVM)
(2) Graph-based clustering
(3) Prefetching candidate mining (PCM)
(4) Neural network

4) *Support vector machine (SVM):*

In machine learning, support vector machines (SVMs, also support vector networks [7]) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM 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.

The Support Vector Machine (SVM) has achieved a lot of attention since it is developed. It is widely used in many areas because of its powerful ability of classification and regression, such as textual classification, face recognition, image processing, hand-written recognition and so forth.

F. *WWW:*

The WWW can be considered as a large distributed information system where users can access to shared data objects. Its usage is inexpensive, and accessing information is faster using the WWW than using any other means [9]. The main problem of these may result to extreme congestion on the network and load on the servers, all resulting in unacceptable dilapidation of the Quality of Service (QOS) at the user end. The web caching is required to alleviate the situation. But the cache management has many challenges in balancing the process of meeting the demands of the users on the one hand and ensuring optimal consumption of system resources on the other hand [10]. Web prefetching is the process of deducing client's future request for web document and getting that document in to the cache, in the background, before an explicit request is made for them. Prefetching capitalizes on the spatial locality present in request streams that is correlated reference for different document and exploits the client's idle time, i.e., the time between successive requests [11]. Web caching is a widely deployed technique in the web architecture that takes advantage of the web object's temporal locality to reduce the user perceived latency. Web caching stores the web objects requested by users. The client side avoid requesting again the objects to the original web servers [12]. An important benefit of the WWW is that many web servers keep a server access log of its users. These logs can be used to train a prediction model for future document accesses. Based on these models, it can obtain frequent access patterns in web logs and mine association rules for path prediction [13]. Our proposed SVM method predicts a user access request before it is actually demanded. The key issue of SVM is to establish an effective user prediction model that extracts useful knowledge from user request sequence and make a prediction of the web pages that the user is likely to request in the near future.

This technique takes advantage of the spatial locality shown by the web objects [14].The prefetching technique has two main workings is that the prediction engine and the prefetching engine.

The prediction engine runs a prediction algorithm and to predict the next user's request. The prefetching engine handles decide to prefetch [12]. As show in figure 2, the Predictions (PD) are the number of objects predicted by the prediction engine [15]. Prefetch Request (PR) represents the number of objects prefetched. The number of objects prefetched that are requested later by the user is the Prefetch Hit (PH). The opposite of the prefetch hit is the Prefetch Miss (PM), which represents the number of prefetched objects that were never demanded by the user. Finally, User Request (UR) refers to the total amount of objects requested by the user (prefetched or not), and the user request not prefetched represents the number of objects demanded by the user that were not prefetched [15].

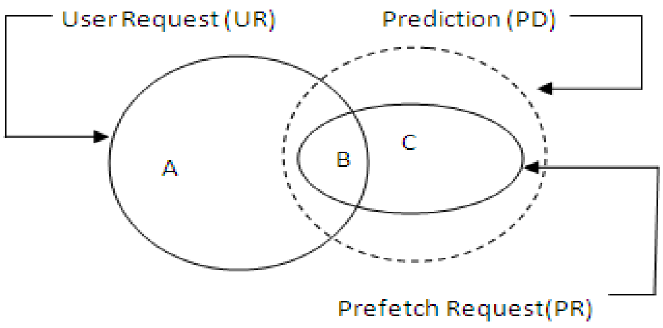

Figure 2: Web prefetching type of requests.

The set of prefetch request is a subset of the prediction set. The result of the intersection between the user request set and prefetch request set is the prefetch hit subset. This subset is the main factor to reduce the perceived latency. In figure.2, A represents a user request not prefetched, which is a user request neither predicted nor prefetched. B is a prefetch request made by the prefetching engine that is requested later by the user, thus becoming a prefetch hit. C is a prefetch miss resulting from an unsuccessful prediction that was prefetched but never demanded by the user. This request becomes extra traffic and extra server load [14].

II. Literature Survey

Because analog circuits such as abnormal noise contained in the information, to the support vector machine to build up the optimal classification brings difficulties, in [16] Jing Tang et. al. propose a new method for analog circuit fault diagnosis. First of all, time-domain signal extraction circuit statistical parameters, a set of fault characteristics and then use kernel density estimation method, proposed a form of fuzzy membership function construction, to eliminate the impact of noise characteristics. State monitoring and fault diagnosing of rolling bearing by analysing vibrating signal is one of the major problem which need to be solved in mechanical engineering. In the paper [17] Lu Shuang et. al. , a new method of fault diagnosis based on principal components analysis and support vector machine is presented on the basis of statistical learning theory and the feature analysis of vibrating signal of rolling bearing.

In this paper [14] Shuang Lu et. al., a new method of fault diagnosis based on K-L transform and support vector machine (SVM) is presented on the basis of statistical learning theory and the feature analysis of vibrating signal of ball bearing. The key to the fault bearings diagnosis is feature extracting and feature classifying. Multidimensional correlated variable is converted into low dimensional independent eigenvector by means of K-L transform.

The pattern recognition and the nonlinear regression are achieved by the method of support vector machine. In the light of the feature of vibrating signals, eigenvector is obtained using K-L transform, fault diagnosis of ball bearing is recognized correspondingly using support vector machine multiple fault classifier.

The use of induction motors is widespread in industry. Many researchers have studied the condition monitoring and detecting the faults of induction motors at an early stage. Early detection of motor faults results in fast unscheduled maintenance. In study [12] Aydin, I et.al presented a new artificial immune based support vector machine algorithm which is proposed for fault diagnosis of induction motors. Support vector machines (SVMs) have become one of the most popular classification methods in soft computing, recently. However, classification accuracy depends on kernel and penalty parameters. Artificial immune system has abilities of learning, memory and self-adaptive control.

The kernel and penalizes parameters of support vector machine are tuned using artificial immune system.

The training data of support vector machine are extracted from three phase motor current. The new feature vector is constructed based on park's vector approach. The phase space of this feature vector is constructed using nonlinear time series analysis. Broken rotor bar and stator short circuit faults are classified in combined phase space using support vector machines.

It is well-known that the correct diagnosis for wireless sensor network can avoid the paralysis of entire systems. In [13] Wang Zhi et. al., fault diagnosis for wireless sensor network based on genetic-support vector machine is presented in the paper. In SVM, inappropriate training parameters can lead to over-fitting or under-fitting. Thus, genetic algorithm is used to select the appropriate training parameters of support vector machine.

In [11] Shuang Lu presented the method of fault diagnosis of rolling bearings based on wavelet packet transform and support vector machine. The key to fault bearings diagnosis is feature extracting and feature classifying. Wavelet packet transform, as a new technique of signal processing, possesses excellent characteristic of time-frequency localization and is suitable for analysing the time-varying or transient signals. Support vector machine is capable of pattern recognition and nonlinear regression.

According to the frequency domain feature of rolling bearing vibration signal, energy eigenvector of frequency domain is extracted using wavelet packet transform method. Fault pattern of rolling bearing is recognized using support vector machine multiple fault classifier.

In [12] Hong Song; et. al. a new method of identifying bus faults based on support vector machine is proposed. First PSCAD/EMTDC is used to simulate the bus fault state, and then a support vector machine model is established after extracting Simulation Data, carrying out data pre-treatments. Different kernel functions is used to train respectively for determining internal fault, external fault of bus and correct identification of fault type.

Network fault knowledge acquisition is a necessary part of intelligent network management. In the paper [13] Jing Wu et. al. , knowledge acquisition of two hierarchies is designed for modern network of large scale and some performance parameters instead of management information base are used to model the network faults so that the evaluations of network fault knowledge acquisition can easily be uniformed.

In view of the non-stationary features of vibration signals of gear and the difficulty to obtain a large number of fault samples in practice, a fault diagnosis scheme based on empirical mode decomposition (EMD) entropy of singular values and support vector machine is put forward in the paper [9] Zhang Chao; et. al. . Firstly, original acceleration vibration signals are decomposed into a finite number of stationary intrinsic mode functions (IMFs); the initial feature vector matrixes are formed by the intrinsic mode functions. Secondly, using the singular value decomposition technique to the vector matrixes, the singular values are obtained. Finally, the singular values serve as the fault characteristic vectors to be inputted to the support vector machine classifier and the work conditions and fault patterns are identified by the output of the classifier.

Research on turbopump fault detection is significance in engine health monitoring. Support Vector Machine (SVM) is a novel machine learning method, and we can use it in turbopump fault detection to solve the problems such as small sample and nonlinear problems. In this paper, [14] Tao Hong et. al. established a kind of adaptive two-class C-Support Vector Machine (C-SVM) algorithm for fault detection based on original C-SVM algorithm, and described the main parts of the algorithm such as fault feature extraction, kernel function choosing and classifier real-time updating in detail.

SVMs can be used to solve various real world problems:

- SVMs are helpful in text and hypertext categorization as their application can significantly reduce the need for labelled training instances in both the standard inductive and transductive settings.
- Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback.
- SVMs are also useful in medical science to classify proteins with up to 90% of the compounds classified correctly.
- Hand-written characters can be recognized using SVMs.

III. MATHEMATICAL MODEL

- Let U= {u1, u2……u_N} be total N number of users of web proxy.
- Let V (u_1) = {V_{11}^{t1}, V_{21}^{t2} ,…,V_{m1}^{tp}} be the total number of pages visited by User u1 in one month at different time t = {t_1, t2….t_p}
- Let C be the cache used for storing pages
- Let Cand = {Cp_1, Cp_2….Cp_k} be the set of candidate link which can be prefetched.
- We formulate our problem as given C and set of links to find the pages which can be prefetched so that cache hit ratio will be increased.

IV. PROPOSED ALGORITHM

For presenting a new supervised machine learning algorithm, SVM model is used to implement web prefetching algorithm. Firstly, the request is sent by client to the web server and waits for response. At the beginning the cache is empty. There are total numbers of requests of users processed in web proxy server. It also a check for the total number of pages visited by different users in one month at different time and as per prediction cache is used for storing pages. It set the links as per cache which can be prefetched. We formulate our problem as given cache and set of links to find the pages which can be prefetched so that cache hit ratio will be increased. SVM prediction is applied with request page on server and then fetch next page. With this request the web server sends the requested web page such as predicted page to the cache. If the prediction is correct, sends a request, the page in the cache and then calculate hit rate and byte hit rate. For next new session of request then the cache will again send the request to the server, which applies the SVM prediction model to send the next future page.

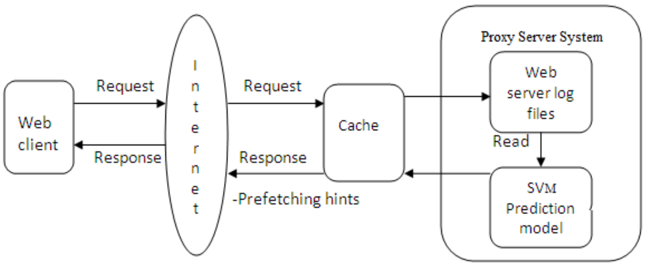

Figure 3: Proposed web prediction model.

Step of projected algorithm is following:

1. Let k be the number of clusters. Select random users and assign them to each cluster as a centroid.

2. Using Cosine similarity assigns the remaining users to appropriate clusters. Cosine similarity between centroid of cluster and user be classified using V(u) vector by using following formula.

$$\mathrm{Sim}(u1, u2) = \frac{\sum_{m=0}^{m-1} V(u1).V(u2)}{\sqrt{\sum_{i=0}^{m-1} V(u_1)^2 . \sum_{i=0}^{m-1} V(u_2)^2}}$$

3. Calculate new centroids and assign users to new clusters until centroid doesn’t move.

4. Using SVM functions

$$\mathrm{Max} \left\{ \sum_{i=1}^{n} r_i - ½ \sum_{i,j=1}^{n} r_i r_j y_i y_j (x_i x_j) \right\}$$

$$\text{Subject to } \sum_{i=1}^{k} r_i y_i = 0$$

$$0 <= r_i <= c \quad \text{for } i = \{1, 2, 3 \ldots n\}$$

Extract the feature from clusters achieved in previous sections 5. When user visits any URL that stores the pages in cache and use least frequently used policy for replacing pages.

6. Periodically use of clustering in previous section is used to keep updated clusters.

7. When users visit pages it obtains the candidate links.

8. Using SVM function for every page it predicts whether user likes it or not.

9. Prefetch only those pages from candidate links which users like.

V. Result

The result achieved will improve the performance of the proposed search method using SVM. It will process the total number of users of web proxy and total number of pages visited by each user in one month at different time and according to SVM, cache stores the expected browsing link pages which set of candidate link can be prefetched. The proposed system would be implemented under Microsoft Windows operating system. This web log contains all the HTTP requests collected from Web. We have showed that predictive system performance in term of hit rate and byte hit rate. Both hit rate and byte rate are growing in our experiment. SVM Regression used to predict the user’s next request in web prefetching by extracting useful knowledge from historical user requests. The number of web pages on the Internet has grown explosively in the past and such growth is expected to be more acute in the future.

Requests	Hit Rate	Cache Size	Byte Hit Rate
1000	66.6481	0.0136	66.5100
2000	74.5864	0.0239	88.8722
4000	81.1205	0.0459	93.0043
7000	84.9020	0.0760	95.7791
80000	85.7083	0.0870	96.1029
130000	88.5166	0.1412	98.1029

Table 1. Performance Analysis using SVM

The above table request shows number of request of dataset that are applied for prediction. We examine each set of requests on prediction based proxy server using SVM to improve the hit ratio. Completion of prediction process then calculates hit rate, cache size and byte hit rate. These calculations are based on predicted requests and when SVM prediction predicts more requests then it increases hit rate and byte rate.

The SVM search will do with performance and design parameter which is useful for prediction based web proxy server to improve hit ratio.

VI. Conclusion and Future Scope

Increasing attraction of WWW over the earlier period of few years has imposed a significant load upon the internet and World Wide Web is huge distributed information system and users can access shared data object.so results of internet service are slow down such as retrieve page from server and decrease the performance of system. To solve this problem, we have applied machine learning technique SVM for prediction based proxy server in web prefetching.it improve hit and byte rate. This paper presents our research work on Support vector machine through machine learning approaches. The main contributions of this paper is search directions will process monthly request as per time to increase hit ratio, Prediction is periodical. Also every page is predicated whether user likes it or not. It prefetches only those pages from candidate links which users like.

In future there can be some other ways to get better performance of system, one of which can be as Incremental Support Vector Machine (ISVM).

ACKNOWLEDGMENT

I would like to express our sincere gratitude and thanks to Mrs. Poonam Gupta, Asst. Professor, Department of Computer Engineering in G.H. RAISONI College, for her help, Inspiration and intellectual impact, which made this paper as desired. Her valuable guidance, reviews and suggestions have been instrumental in the successful finishing of this work.

REFERENCES

[1] G. Barish and K. Obraczka, “World Wide Web Caching: Trends and Techniques,” IEEE Comm. Magazine, Internet Technology Series, pp. 178-185, 2000.

[2] K. Chinen and S. Yamaguchi, “An Interactive Prefetching Proxy Server for Improvement of WWW Latency,” Proc. Seventh Ann. Conf. Internet Soc., June 1997.

[3] P. Cao and S. Irani, “Cost-Aware WWW Proxy Caching Algorithms,” Proc. 1997 USENIX Symp. Internet Technology and Systems, 1997.

[4] C. Aggarwal, J.L. Wolf, and P.-S. Yu, “Caching on the World Wide

[5] Web,” IEEE Trans. Knowledge and Data Eng., vol. 11, no. 1, pp. 94-107, Jan./Feb. 1999.

[6] R.P. Wooster and M. Abrams, “Proxy Caching That Estimates Page Load Delays,” Proc. Sixth Int’l World Wide Web Conf., 1997.

[7] Y.-H. Wu and A.L. Chen, “Prediction of Web Page Accesses by Proxy Server Log,” World Wide Web, vol. 5, no. 1, pp. 67-88, 2002.

[8] Wang Zhi, Fault diagnosis for wireless sensor network based on genetic-support vector machine, 2691- 2694, 2012.

[9] Shuang Lu; Weizeng Chen; Meng Li, Fault Pattern Recognition of Rolling Bearing Based on Wavelet Packet and Support Vector Machine 5516- 5520, 2011.

[10] Hong Song; Hao Wu, The applied research of support vector machine in bus fault identification.1326-1329,2013.

[11] Jing Wu; Jian-Guo Zhou; Pu-Liu Yan; Ming Wu, A study on network fault knowledge acquisition based on support vector machine,3893- 3898 Vol. 6. 2013

[12] Zhang Chao; Yang Li-dong; Chen Jian-jun, Fault diagnosis of gear cases based on entropy of singular values and support vector machine 4207- 4211, 2013

[13] Shengchun Wang; Qing Zhang; Tonghong Jin; Shijun Song, Study on the fault diagnosis based on wavelet packet and support vector machine 3457-3461, 2013

[14] Vyas, B.; Maheshwari, R.P.; Das, B., Fault analysis of controllable series compensated transmission line with Wavelet Transform and Support Vector Machine 15, 2012

[15] Wenying Feng, Karan Vij .2008. Web cache prefetching by multi - dimensional matrix. IEEE, pp. 265-270.

[16] Jing Tang; Yun'an Hu; Tao Lin; Yu Chen, Analog circuit fault diagnosis based on fuzzy support vector machine and kernel density estimation, V4-544 - V4-548, 2012.

[17] Lu Shuang; Yu Fujin, Fault Pattern Recognition of Bearing Based on Principal Components Analysis and Support Vector Machine, 533-536, 2011.