Web Access Prediction Model using Clustering and Artificial Neural Network

Om Prakash Mandal
Dept. of Computer Science & Engg.
National Institute Of Technology, Patna
Patna, India

Hiteshwar Kumar Azad
Dept. of Computer Science & Engg.
National Institute Of Technology, Patna
Patna, India

***Abstract*— The number of web users is increasing rapidly day by day. With the increasing number of web access requests and explosive growth of data sources available on the web, the network traffic also increases rapidly, which results in poor user latency and difficult for user to access the web. Thus web caching and pre-fetching of web pages play an important role in this scenario. However, without having a sophisticated mechanism regarding which pages are most likely to be visited, caching is meaningless. The idea is to improving the accuracy in web access prediction. This paper proposed a next web page access prediction model. The proposed model uses Feed Forward Artificial Neural Network and the concept of web session clustering. This model can also be used in other web oriented applications. The Neural Network is trained and tested with a large data set and better prediction accuracy is achieved.**

Keywords—Artificial Neural Network; K-means Clustering; Web Caching; Web Page Prediction; Web User Sessions

I. INTRODUCTION

The domain of social networking and e-marketing is getting larger day by day. With increasing the number of users of the web, web traffic has also been increased in recent years. If somehow we can predict the user access behavior while the user interacts with web, user access latency can be improved. Web page prediction mechanism deals with the problem of forecasting the next web page which is most likely to be visited by the user. It can also be said as the problem to predict navigational pattern of a user from the current active page based on the knowledge of the previously visited pages.

Web caching and web pre-fetching plays a major role to reduce the search cost around web. To implement a good web caching, there is a need of an efficient web access prediction mechanism. User's navigation history can be obtained using server log. Using the history of user's navigation with in a period of time, a user session is created. The information contained is the session is extracted to train the prediction model. In this paper, Artificial Neural Network (ANN) is used for prediction purpose.

The objective of this paper is to develop a prediction model using Artificial Neural Network in combination with the K-means clustering algorithm. This paper is organized in five sections. Section 2 presents a brief literature survey. Section 3 explains the proposed prediction model and all its components. Section 4 discusses detailed experimental result and analysis. Section 5 concludes the paper and discusses the future research directions followed by references at the end.

II. RELATED WORK

A no. of approaches and architectural models has been proposed in recent years. Zhao et al. [1] proposed a framework to analyze the web access pattern using the historical log data.

Deshpande et al. [2] uses selective Markov model for web page prediction. To obtain highly accurate and lesser complex model, different parts from different Markov models were used.

Khalil et al. [3] also used Markov Prediction model and used simple distance based K-means clustering to improve the performance of the model.

Kim et al. [4] proposed a model by hybridizing Markov model, sequential association rule, association rule. The goal was to improve the performance by reducing the recall. But unfortunately, the overall prediction accuracy did not improved.

Chitraa et al. [5] proposed an improved web clustering technique. High quality and more accurate clusters were obtained using the technique.

Varghese et al. [6] used cluster optimization technique to improve web usage mining. They proposed the cluster as a similar data object and then used Fuzzy Logic to optimize the data set. Wang et al. [7] proposed an approach for clustering web pages by sequential alignment.

Awad et al. [8] analyzed different prediction models like Support Vector Machine (SVM), Artificial Neural Network (ANN) along with Markov Model. They also proposed an approach using Markov Model.

Poornalatha et al. [9] proposed a prediction system using clustering the web sessions and integrated distance measure.

Integrated distance measure and technique of sequential alignment is used to find the similarities between any two user sessions.

Zheng et al. [10] used Artificial Neural Network to predict the customer behaviour preference based on social media. They also predicted customers' behaviour regarding restaurant preferences using Support Vector Machine. The ANN provided 93.13% average accuracy across a no. of investigated customers.

Lin-Hong et al. [11] proposed RBF Neural Network based short-term Prediction model for Quality of Service (QoS) of web service. K-means clustering is also used to improve the prediction accuracy. Jesus et al. [12] discussed Back propagation algorithms for a broad class of Dynamic Neural Networks (DNN).

The overview of web page prediction mechanisms can be found in the survey presented by Kumar et al. [13]. Srivisal et al. [14] proposed a technique to predict the number of unsupervised clusters by using supervised function. Jha et al. [15] provided the very useful literature regarding Artificial Neural Network and its applications.

Azad et al. [16] proposed a novel technique to access the most relevant and accurate data sources available on the web, which integrate the idea from semantic web and synaptic web at low entropy and web pages are distributed at a hierarchical range of entropy.

III. PROPOSED MODEL AND METHODOLOGY

The process starts with obtaining user's web access data using server log files. The log data is parsed, relevant information is extracted and the sessions are created using periodic analysis. The overall operational flow is shown in the Figure 1.

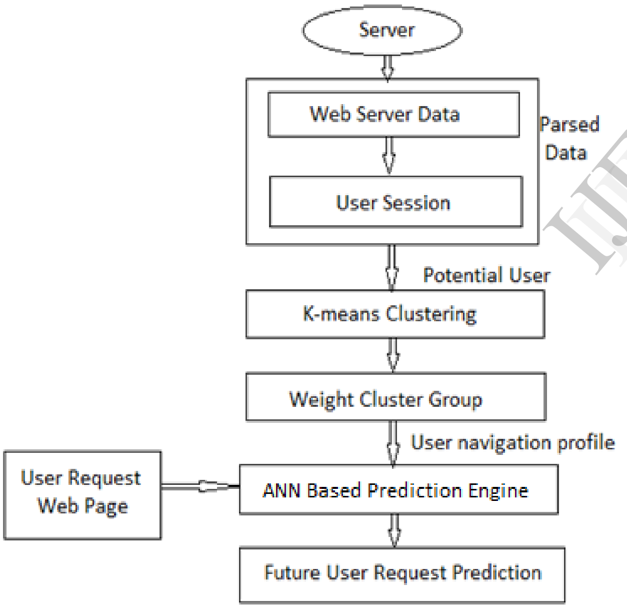


Figure 1: Operational flow of the proposed system.

In this paper, it is assumed that the web log data is parsed and the sessions have been created.

A. Components of Prediction Model

The various components which constitute the Prediction model are described as follows.

Web Server Log and User Sessions: Web server log file collects the data entry regarding web access for a particular user. Different client information such as IP address of user, the time at which the server is being accessed, HTTP method like Get or Post, URL of the requested web page, response from the server (response code) and data transferred from server to user etc. Each entry is parsed to extract above meaningful information. This information is used to create user sessions. A user session is a collection of contiguous sequence in which the client accesses web pages within a certain time bound. After the sessions are created, theses are filtered to remove irrelevant data such as image files etc. Unique id is given to each unique request identified from these filtered sessions. An example of a server log is as follows [20]:

> 122.235.126.154 - - [12/JUN/2010:00:23:48 -0600] "GET /pics/wpaper.gif HTTP/1.0" 200 6248 "http://www.google.com/asctortf/" Chrome"/.

Clusters: User sessions are divided into clusters. Each cluster contains a particular sequence of web pages. Any two clusters may have multiple similar web pages. The only difference is the sequence in which they are accessed by user.

The ANN is trained using these clusters. The clusters are formed using the K-mean clustering algorithm which is described as follows. Initially cluster centers are selected randomly. Once clusters are formed, each cluster is assigned with a unique identification number. The sequential alignment technique [7] is applied to measure the similarities between two sessions [9]. The K-mean clustering and classification algorithm is as follows:

Let X = $\{x_1, x_2, x_3 \ldots\ldots x_n\}$ be the set of data points and C = $\{c_{1,} c_2, c_3 \ldots\ldots c_k\}$ be the set of centres.

- Randomly select 'c' cluster centres. Place K points into the space represented by the data point are being clustered and these points are assigned as the initial group centroids.
- Calculate the distance between each data point and cluster has to group the closest centroid.
- Assign the data point to the cluster centre whose distance from the cluster centre is the minimum of all the cluster centres.
- Recalculate the new cluster centre (When all objects have been assigned, recalculate the positions of the K centroids). Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter," not "webers/m2." Spell units when they appear in text: "...a few henries," not "...a few H."

ANN based Prediction Engine: In early 1950's, Neural Network and its mathematical computational models were established.

The computation of the activation level is based on the values of each input signal received from a neighbouring node. The weights on each input simulate the learning and decision making processes of the human brain [15]. ANN is the interconnection of artificial neurons working in a fashion to solve a specific problem. Neurons connected via small branches extension of nerve cells. Each cell receives signals from other cells. In an artificial neuron, the information is broadcasted throughout the network and stored in the form of weighted interconnections. Figure 2 shows the graphical representation of a neuron.

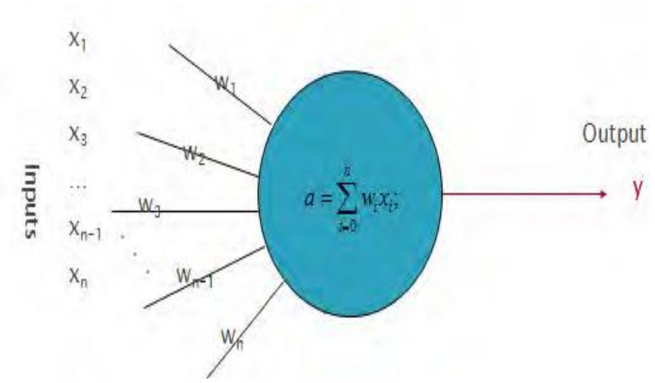


Figure 2: Graphical representation of a neuron. xi represents the inputs to the neuron and wi represents weights of the neuron. The overall input to the neuron is calculated by a=Σn i=0wix. Y represents the ouyput [21].

A layered feed forward neural network has many processing elements arranged in different layers (hidden layer, input layer, output layer). Each processing element receives input from the neurons of the previous layers and calculates weighted sum of its inputs. Figure 3 shows the general terminology of a Feed Forward Neural Network.

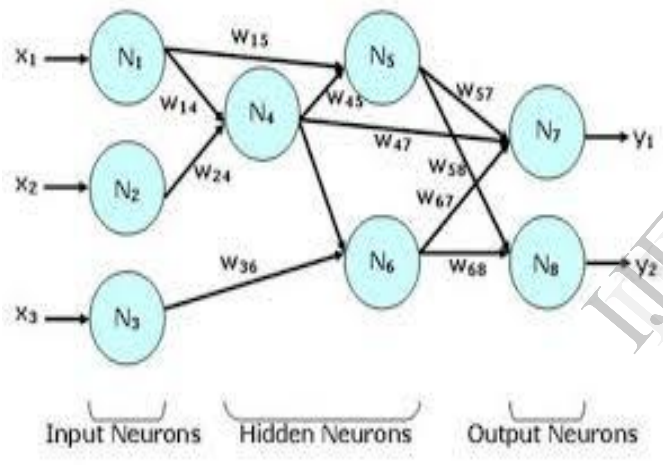


Figure 3: Layout of Feed Forward Neural Network [22]

The very first layer is the input layer which receives the user input. The output from input layer is feed forwarded to hidden layers. The hidden layers are optional and number of hidden nodes differs from application to application. There are no cycles in the entire feed forward network. The output from hidden layers is forwarded to output layers to produce final result. A threshold function may be used to test the quality of the output of a neuron in the output layer. In figure 3 the final result is shown by Y_1 and Y_2. Connection between nodes represents processing of weighted sum coming from the output of input nodes. The hidden nodes which constitute the middle layer provide internal representation for the development of ANN models.

B. *Working Mechanism*

There are four steps in the prediction model.

- Data collection from web server.
- Parsing data from server logs.
- User session and clustering of potential user.
- Prediction for user request using ANN.

The first step is to collect the log data from an active Web server. The log file contains information related to user's web access like the host (user system) users IP address, users date and time at which the web request is made, requested web page, server's HTTP reply service code and the amount of data transferred in the reply.

The data collected from the raw web log file is parsed to extract relevant information. The sessions are created by clustering all HTTP requests that are originated from a single IP address and applying a timeout approach to break these clusters into unique sessions. Each unique request in these sessions is assigned a unique identity. The sessions are then filtered to remove unnecessary information like images file etc. Each unique session contains sequence of requests coming from the same IP address.

The sessions are created based on number of pages. Each session contains 10 web pages. This assumption is made to obtain better result while measuring similarity between two sessions. Each session is assigned a unique number starting from S_1 to S_n as shown below.

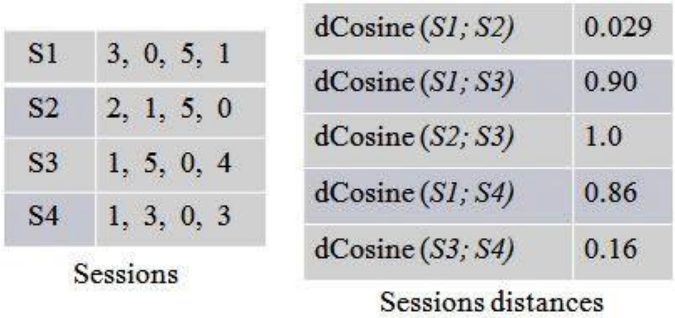
S_1: pp_1, pp_2, pp_3, pp_4, pp_5, pp_6, pp_7, pp_8, pp_{14}, pp_{15}
S_2: pp_1, pp_2, pp_3, pp_4, pp_6, pp_9, pp_{10}, pp_{11}, pp_{12}, pp_{13}
S_3: pp_1, pp_2, pp_6, pp_7, pp_8, pp_{11}, pp_{12}, pp_3, pp_4, pp_{15}
S_4: pp_1, pp_6, pp_7, pp_4, pp_6, pp_1, pp_2, pp_{13}, pp_9, pp_5
$S_{5:}$ pp_2, pp_6, pp_5, pp_8, pp_3, pp_1, pp_2, pp_3, pp_{14}, pp_{15}
S_6: pp_1, pp_4, pp_1, pp_5, pp_4, pp_3, pp_7, pp_6, pp_{12}, pp_5

Where pp_1, pp_2, pp_3 etc. are the unique pages.

After the sessions are formed, Cosine Distance measure is used to find the distance between two sessions.
The Cosine distance measure is calculated as:

$$dCosine\ (S_i, S_j) = \sum (Si_k, Sj_k) / \sqrt{} \sum (Si_k)^2 \sqrt{} \sum (Sj_k)^2 \qquad (1)$$

The distances calculated using equation (1)



S1	3, 0, 5, 1
S2	2, 1, 5, 0
S3	1, 5, 0, 4
S4	1, 3, 0, 3

Sessions

dCosine (*S1; S2*)	0.029
dCosine (*S1; S3*)	0.90
dCosine (*S2; S3*)	1.0
dCosine (*S1; S4*)	0.86
dCosine (*S3; S4*)	0.16

Sessions distances

Figure 4: Measument of distance between Sessions

Clusters are formed according to the least distances between sessions, or the closest distances between sessions. Therefore, (S1; S2) will form a cluster and (S3; S4) will form another cluster.

In the next step, k-means algorithm is applied and the relevant clusters of training sessions are generated using Cosine distance between sessions. Initially cluster centers are selected randomly. Once clusters are formed, each cluster is assigned with a unique identification number.

Figure 5 represents two clusters with their cetroids.

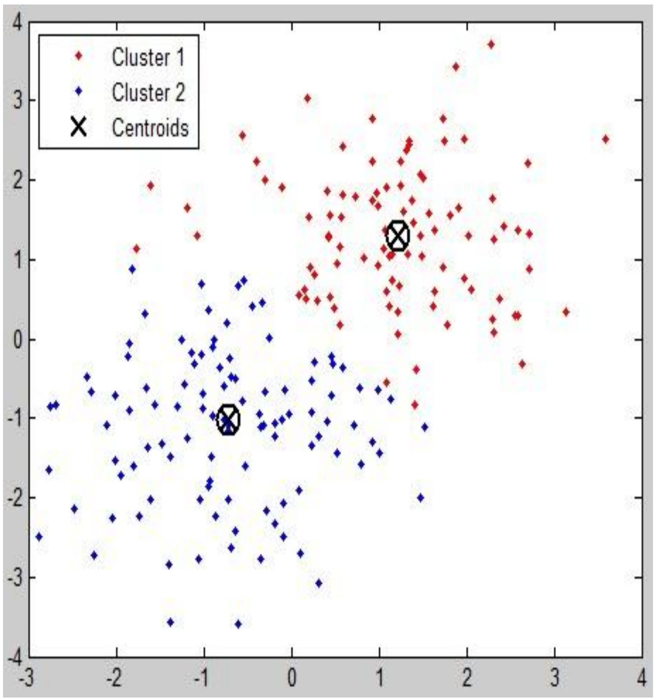


Figure 5: Clusters and their centers

It is worth mentioning that the prediction model works in both modes: Online and Offline.

In online mode, the cluster which has the highest probability to find the web page is found.

In offline mode pre fetching of web pages is performed. While the user is not active, the server refines the web cache by fetching the pages which are most likely to be visited when the user will become active.

For each cluster, a numerical value is obtained by calculating the weighted sum of the total Score of a session.

The final step involves the application of Feed Forward Neural Network. The input layer is feed with the cluster value and using 10 levels of hidden layers the most relevant (highest probability) cluster is found.

IV. Result and Analysis

The 80% clusters are used to train the ANN. Rest 20% clusters are used to test the prediction engine. Self Organizing Map (SOM) distribution of weights of training data used in neural network is shown in figure 6.

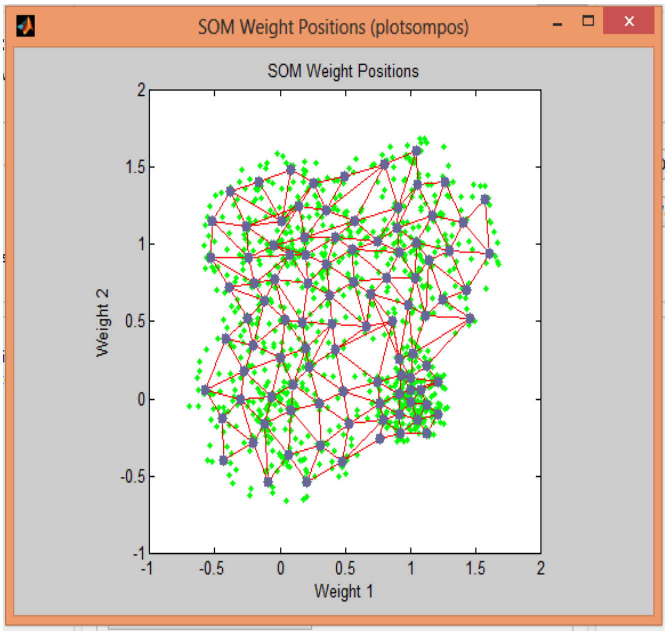


Figure 6: Weight Positions after clustering the data

Figure 7 shows the performance of the system validating the input data.

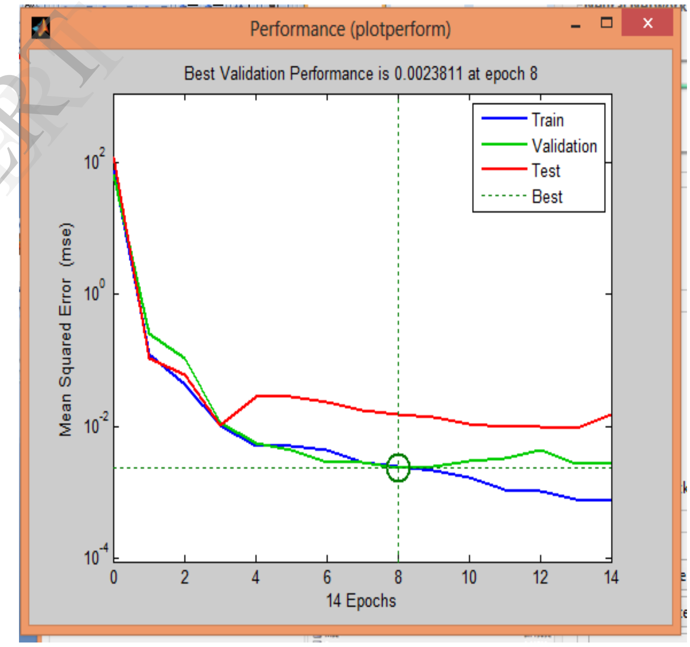


Figure 7: Validating input data using nftool.

Training, validation and testing of prediction system are shown in color Red, Green and Blue respectively. The best performance is obtained at 8th epoch which is represented by circle. Calculated Mean Square Error (MSE) is 3.71909e-3. Input data of 2*800 matrix size has been used for training the system. The target data used has a matrix size of 4*800. After training the system, it is validated with input data. The best validation points achieved is 0.0023811 at epoch 8.

As the training data increases prediction accuracy also increases.10 or 20 neurons are used in the neural network.

Figure 8 shows the relationship between Prediction accuracy and the no. of sessions (size of training data).Which show the performance of the neural networks.

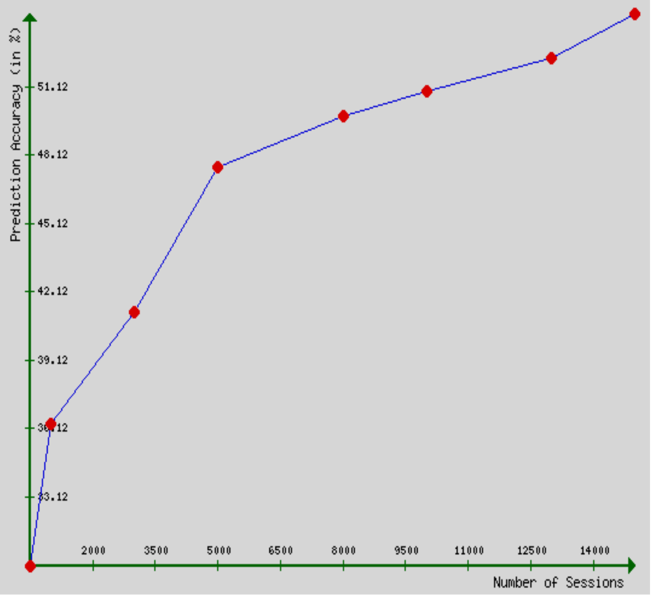


Figure 8: Relation between no. of session and accuracy of the model.

V. CONCLUSION AND FUTURE SCOPE

Web caching is a very important aspect in web user latency. Many approaches have been proposed in this area. Prediction of next web page to be accessed by user is an adequate solution to this problem. Many other web oriented areas can also utilize this prediction. Domains like e-commerce, web personalization, page importance index etc. would be benefitted using the approach. In this paper, one such prediction model is presented and significant result is obtained.

The motivation behind using neural network comes from the fact that a neural network can perform more complex task as compared to a linear program cannot. Also, neural network exhibits better reliability because of their mesh like structure and parallel nature [20].

The prediction accuracy majorly depends upon the clustering scheme. Efficient and relevant clustering of web sessions depends on similarity measure used. The parameter and metric used for similarity measure can be tested by applying clustering mechanism on a pre-used labelled data, where the clustering result is known a-priori. However, the pre-used data and result must be hidden from the algorithm.

How to further improve the clustering scheme (keeping web structure in mind) and to optimize the prediction algorithm will be objective for the future research works. Sliding window can be integrated with the proposed approach to get better efficiency in prediction.

REFERENCES

[1] G. Zhao, Q., Bhowmick, S. S. and L. Gruenwald, “Wam:miner: in the search of web access motifs from historical web log data,” *in ‘CIKM05 conference’, Germany*, pp. 421-428.

[2] M. Deshpande, G. Karypis, “Selective markov models for predicting web page Accesses,” *[J] ACM Transaction on InternetTechnology*,2004, vol. 4(2), pp. 163- 184.

[3] F. Khalil, J. Li, H. Wang,”Integrating recommendation models for improved web page prediction accuracy,” *Proc. 31st Australasian Computer Science Conference (ACSC 2008,* University of Southern Queensland, Toowoomba, Australia, 4350.

[4] D. Kim, l. lm, N. Adam, V. Atluri, M. Bieber, Y. Yesha, “A clickstream-based collaborative filtering personalization model: towards a better performance,” *Proceedings of the 6th annual international workshop on web information and data management ACM*, 2004, pp.88-95 DOI:10.1145/1031453.1031470.

[5] V.Chitraa, Dr.A S Thanamani, “An enhanced clustering technique for web usage mining”, *International Journal of Engineering Research & Technology (IJERT,* June – 2012, Vol. 1 Issue 4, ISSN: 2278-0181.

[6] N. M. Varghese and J. John, “Cluster optimization for enhanced web usage mining using fuzzy logic”, *IEEE ISSN978-1-4673-4805-8/12*, 2012.

[7] W. Wang, O. R. Za¨ıane, “Clustering web sessions by sequence alignment “*University of Alberta Edmonton, Alberta, Canada*.

[8] M. A. Awad and I. Khalil,” Prediction of user’s web-browsing behavior: application of markov model”, *IEEE Transaction on Systems, Man and Cybernetics—Part B: Cybernetics*, 2012, Vol. 42, NO. 4.

[9] G. Poornalatha, R. S. Prakash, “Web page prediction by clustering and integrate distance measures” *IEEE/ ACM Trans. Syst., Man, Cybern. A Syst., Humans*, Sep 2012, vol. 44, no. 2.

[10] B. Zheng, K. Thompson, S. S. Lam, S. W. Yoon “Customers’ behavior prediction using artificial neural network” *Poceedings of the 2013 Industrial and System Engineering Research conference*.

[11] Z. lin-hong “A short-term prediction for QoS of web service based on rbf neural networks including an improved k-means algorithm,” 2010, *International Conference on Computer Application and System Modelling (ICCASM)*.

[12] O. De Jesús and Martin T. Hagan, “Back propagation algorithms for a broad class of dynamic networks” *IEEE Transactions on Neural Networks*,January 2007, Vol. 18, No. 1.

[13] Sunil Kumar, Ms. Mala Kalra, “Web page prediction techniques: a review”, *International Journal of Computer Trends and Technology (IJCTT)*, July 2013, vol. 4 Issue 7.

[14] C. Srivisal, C. Lursinsap “Predicting number of unsupervised clusters by supervised function,” *International Joint Conference on Computational Sciences and Optimization*, 2009.

[15] G. K. Jha “Artificial neural network and its application,” *I.A.R.I New Delhi*.

[16] H. K. Azad, Kumar Abhishek, “Semantic-Synaptic web mining: A novel model for improving the web mining”, IEEE International Conference on Communication Systems and Network Technologies (CSNT-2014), pp.454-457, 2014.

[17] R. L. Carter, R. Morris, R. K. Blashfield,”On the partitioning of squared Euclidean distance and its applications in cluster analysis,” *Journal of Psychometrika, Springer New York*, March 1989, Vol. 54, no. 1, pp 9-23.

[18] S. H.Ghwanmeh,”Applying clustering of hierarchical k-means-like algorithm on arabic language”, *International Journal of Technology*, November 2005, Vol. 3, No. 3, pp. 168-172.

[19] J. Roman and J. Akhta, “Back propagation and recurrent neural networks in financial analysis of multiple stock market returns” *Proceedings of the 29th Annual Hawaii International Conference on System Sciences*, 1956.

[20] A web server log file sample, available on 25 April 2014[online]: *www.jafsoft.com/searchengines/**log_sample**.html.*

[21] N. Budhnai, C. K, JHa, S. K. jha, “Application o neural network in analysis of stock market prediction,” *International Journal of Computer Science & Engineering Technology (IJCSET)*, April 2012, Vol. 2 no. 4, pp 61-68.

[22] An Example of Feed Forward Neural Network, available on 25 April 2014 [online]: http://www.emilstefanov.net/Projects/NeuralNetworks.