University of Ibadan
Journal of Science and Logics in ICT Research

An Enhanced Web Page Recommendation System Using Hidden Markov Model and Page Rank Technique

Adelaja, T. T.
adelajatinuade@yahoo.com

Akinola, S. O.
solom202@yahoo.co.uk

Department of Computer Science,
University of Ibadan, Ibadan, Nigeria

Abstract

The rapid expansion of the World Wide Web (WWW) has created an opportunity to disseminate and gather information online. There is an increasing need to study the behaviour of web users to serve them better by reducing the access latency using efficient web prediction technique. Markov Models have been widely used for predicting next web page request from the users' navigational behaviour recorded in the web log. This usage-based technique can be combined with the structural properties of the web pages to achieve better prediction accuracy. This study combines both Markov Model and Page ranking, which considers the structural properties of the Web. In order to create an efficient prediction model, the original data was pre-processed in the form that can be used for unsupervised learning. The pre-processed data was then analyzed using unsupervised learning K-means clustering algorithm. To increase the efficiency of Hidden Markov Model (HMM), efficient ranking algorithm was used to identify the most relevant page in clusters. i.e. PageRank. The HMM was then used to predict users web navigation path. Briefly, results from the study shows that Hidden Markov model and Clustering can work together and provide better prediction results without compromise to accuracy though with a trade-off in time complexity, HMM is more accurate for predicting navigational paths thereby enhancing web page recommendation.

Keywords: Web Mining, Web page recommendation, Information retrieval, Page rank algorithm

I INTRODUCTION

The World Wide Web (WWW) consists of a vast resource of hyperlinked heterogeneous information such as text, audio, video and metadata. Managing the web has become more difficult due to the rapid growth of the information on the WWW. While users have access to more information and various options, it has become difficult for them to find the important information needed. This is a problem commonly referred to as information overload [1].

This has made it essential for web users to apply efficient information retrieval techniques to find and arrange the required information [2]. Large volumes of data like users' addresses, URLs requested and browsing patterns are gathered and stored automatically in access logs by web servers whenever a user visits a website. This is important because most times, users repeatedly access the same website and the records are stored in log files. These series of accessed web pages signifies web access patterns for each user, which is very helpful in finding out the users' behaviour and predicting the user's next request. This will reduce user's browsing time, thus saving user's time and decrease server load.

It is very important to study and analyze web usage by users and web traffic as a result of the increased growth rate of the WWW by using client and server side technology to extract knowledge from the available information. This process is referred to as Web Mining [3].

Web mining is of three forms

- Web content mining: extraction of information from contents of web documents.
- Web structure mining: identification of the relationship between linked web pages
- Web usage mining: extraction of required information from web log files

Web log files are maintained by the server. It contains the web pages accessed by users in each user's session, client's IP address, requested webpages, time spent on a page etc.

In this paper, a web page recommendation system is implemented using the page rank algorithm. It uses history of users' web usage patterns. The objective is to study each user's browsing patterns and accurately predict the next user's request based on the current request made by the user. This was achieved by pre-

Adelaja, T. T. and Akinola, S. O. (2017). An Enhanced Web Page Recommendation System Using Hidden Markov Model and Page Rank Technique, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 1, pp. 57 - 63

processing the data to present the data in a form that can be used for unsupervised learning, clustering the pre-processed data using K-means clustering algorithm, identifying the relevant pages in the clusters using Page Rank algorithm and developing an efficient prediction model to identify relevant pages between clusters.

The need for predicting web users' next request to improve usability and retention is obvious because it has been more difficult for users to find useful information on the WWW based on the fact that web information lacks an integrated structure. However, there are many other means by which users might have a bad web usage experience such as improper layout of web pages or web structure.

A lot of research has been done in the field of web usage mining and prediction of users' future requests. These will be reviewed in the next section.

II RELATED WORKS

Web page prediction is such an interesting field. In order to have a better understanding of recommendation systems, a brief introduction of some well-known approaches is required.

Chimphlee *et al.* [4] introduced a new model for web access prediction by incorporating clustering with Markov model. Getting an accurate prediction is a major drawback of the model due to approximations during the clustering process.

Dhyani *et al.* [5] made use of Markov process for web access prediction. The model considers all access sequences throughout the prediction process which has a high complexity.

Khalil *et al.* [6] proposed a new model for predicting next web page request using both Markov model and association rules such that if Markov model is unable to predict the next page, association rules will be used instead.

Maratea and Petrosino [7] designed a heuristic majority intelligence technique which has the ability to adapt to changing user access patterns with ease. This design is important because the web page recommendation is reliant on the nature of web logs. This technique imitates human behaviour in a new environment alongside other individuals working in the same environment and has the ability to predict in real time and with better accuracy the next page to be visited by a user.

Nigam and Jain [8] proposed a new way of structuring Dynamic Nested Markov Model (DNMM) to model users web navigational patterns. The focus is on time complexity and coverage. In DNMM, the higher order Markov model is nested inside the lower order Markov model. Through this, the second-order Markov model is nested inside the first-order Markov model and all the advantages of both models are combined and achieved in one model. High coverage is achieved and time complexity is reduced.

Anita [9] also introduced a new approach which combines a proposed model pair wise nearest neighbor clustering with Markov model. This model has high prediction accuracy and reduced state space complexity but its major drawback is that loosely connected access patterns are not considered in the mining process.

Rao and Kumari [10] introduced an approach to predict users' browsing behaviour at two stages which are the category stage and web page stage. At the category stage, unnecessary stages are excluded thus reducing the scope for calculation. Next, higher order Markov model is used to predict the user's next web page. The experiment resulted in low state complexity and prediction accuracy.

Nigam [11] evaluated and compared different models for mining the web log to predict next accessed web page. Models like Markov models and Dynamic Nested Markov Model were considered but prediction is still not accurate as irrelevant pages were recommended to users.

III METHODOLOGY

A web server log is the most important source for performing web usage mining because it stores details like browsing behaviour of the visitors of a web page, time and date of transaction, name and URL of the site, content and structure of the webpage, click streams and navigational patterns [11].

To have proper understanding of user's behaviour, the following information should be extracted from the server log.

1. Who is visiting the web site? This has to do with identifying the unique users which refers to a new user on the page so as to obtain the path that each follows.
2. The path the users take through the web pages: the click stream and navigational path of users of the web page helps to define and identify the user's interests.
3. How much time the user spends on each page: it can be concluded that a page is interesting to a user by the length of time spent on that webpage.
4. The point at which the visitors are leaving the web site: Users' session ends on the last page visited by that user.

Predicting user's next page request consists of a number of stages. Each stage has its own unique importance in determining that particular page. Data collection and Preprocessing, Clustering, Page Rank algorithm and Hidden Markov model are the major steps among them. The web Page recommendation framework is depicted in Figure 1 and the explanation follows.

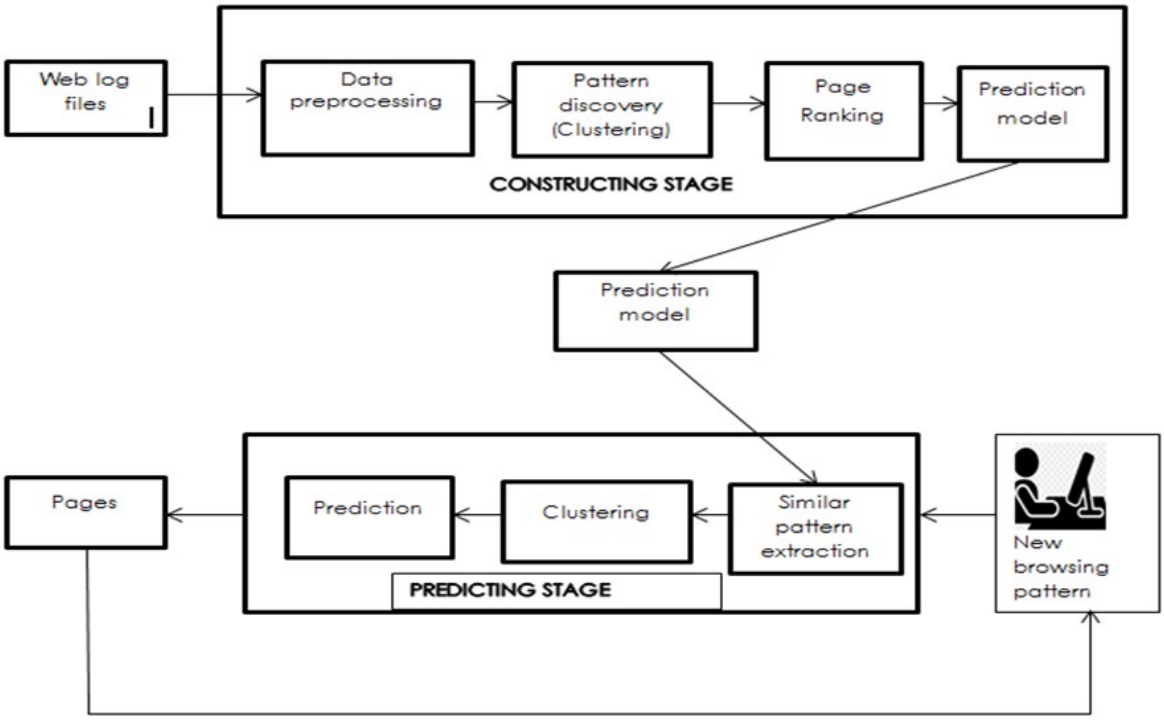

Figure 1: The Web Page Recommendation Framework

A Data Collection and Preprocessing

For the purpose of this research, a set of server logs from MSNBC server was used which describes the page visits of a total 989818 users who visited msnbc.com on September 28, 1999. The data type is a discrete sequence. Here, each sequence corresponds to page views of a user and each event in the sequence corresponds to a user's request for a page.

Data preprocessing is any type of processing carried out on raw data to prepare it for another processing procedure. Here, web usage logs were preprocessed to extract user sessions and were organized in form that can be used for unsupervised learning.

A web session can be defined as user's browsing time on a website i.e., the time between when the user first arrives at a page on the website and the time they stop using the site.

B Clustering Algorithm: K-Means

K-means clustering algorithm is one of the simplest unsupervised learning algorithms that are used for clustering large data sets. Cluster centers are arbitrarily selected, and then Euclidean Distance or any other distance measures between each data set and the cluster centers, the data points, are calculated and assigned to the most suitable cluster [12]. The aim of this algorithm is to classify a set of objects into K clusters based on their attributes such that data with the same characteristics are grouped together in a cluster and find out the centroid for each cluster. The centroid of a cluster is determined using similarity function and Euclidean distance to all objects in that cluster. The K-means algorithm then repeatedly improves within cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in previous iteration. The iteration continues until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round. In this study, web sessions were clustered using the K-means algorithm which in turn was used for prediction of the next web page.

C K-means Algorithm Steps

Input: k: the number of clusters
d: data set containing n objects
Output: a set of k clusters
Method:

```
{
    (i) Arbitrarily choose k objects from d
                as initial cluster centre
    (ii) repeat {
            reassign each object to the cluster to
            which the object is the most similar, based
            on the mean value of the objects in the
            cluster; modify the cluster by calculating
            the mean value of the objects for each
            cluster.
     } until no change.
}
```

After the web sessions have been clustered, the probability of accessing a web page in a web session must be evaluated using the rule stated in (1):

Probability (p) = $\frac{X}{Y}$ ……..(1)

Where X represents the total number of web pages in a web session and Y represents the total number of web pages in all sessions.

D Page Rank Algorithm

This algorithm was introduced by Google. Page ranking is one of the crucial aspects of any information retrieval system. Search engines usually return a lot of results for a search query; this makes it very difficult as users cannot check each page to determine which will be the most useful page. This is where page ranking comes into play.

Page rank algorithm is the link-analysis algorithm used to assign numerical weightage to webpages which determines the relative importance of web-pages [12]. This algorithm is used to rank results of a search query, usually web pages and documents. The more a web page is visited, the higher its rank and the higher its chances to be retrieved from the search engine results. The number a web page is visited is measured and it is used to compute the page rank. This signifies the importance of the web page.

Page rank was introduced in this study because it uses history of visited web pages with the aim of identifying the successive requests of users based on the current request of the user. This reduces server load and access time [13]. Page rank of a web page can be computed using equation (2):

PR= $\mu * \frac{A}{B} + (1 - \mu) * C$……(2)

A= Probability of the current web page.

B= Total number of outbound links, where outbound link is the number of outward links from a current web page to another web page.

C= Probability of the next web page.

μ = Damping factor (μ is a very small number, experimentally found to be 0.85). This is the decay factor which represents the chance that the user on a web page will stop clicking links on the current page and visits a random page by typing a new URL rather than following the link on the current page.

If the damping factor is 85%, then there is assumed to be about a 15% chance that a user will not follow any links on the web page. After the page rank has been calculated, mean value is calculated using (3):

Mean = $\frac{\sqrt{(maximum\ PR)^2 + (minimum\ PR)^2}}{2}$…..(3)

Maximum Page Rank (PR) = web pages from the current webpage with the highest page rank.

Minimum PR = web pages from the current web page with the lowest page rank.

Web pages with ranking less than the mean value were removed and not considered during the prediction process. Transition Probability was then computed for web-pages having values greater than mean value by the formula given in (4):

$AB_m^n = \frac{XY_m^n}{\sum_1^L XY_m^n}$ ……..…..(4)

XY_m^n = No. of times the access is made where i is the current web page and j is the next web page.

L = No. of outbound links.

AB_m^n = Transition Probability from nth page to mth page.

The next web-page is predicted from the highest value among all the transition probabilities

E Prediction Algorithm: Hidden Markov Model

Hidden Markov Models (HMM) are widely used in science, engineering and other areas like speech recognition, optical character recognition, etc. HMM can be used to predict and discover users' behaviour on the web and also differentiate behaviour of web users because it is good at the extraction of information and has the ability to extract information and guess hidden states in the observation symbols with high accuracy [14].

The Hidden Markov Model (HMM) is a variant of a finite state machine having a set of hidden states, **Q**, an output alphabet (observations), ***O***, transition probabilities, ***A***, output (emission) probabilities, ***B***, and initial state probabilities, **Π**. The current state is not observable. Instead, each state produces an output with a certain probability (***B***). Usually the states, ***Q***, and outputs, ***O***, are understood, so an HMM is said to be a triple, (***A***, ***B***, ***Π***) [15].

A HMM is a doubly embedded stochastic process with a hidden stochastic process that can only be observed through another set of stochastic processes that generate the sequence of observable symbols. HMM predictors are very powerful adaptive stochastic models [16].

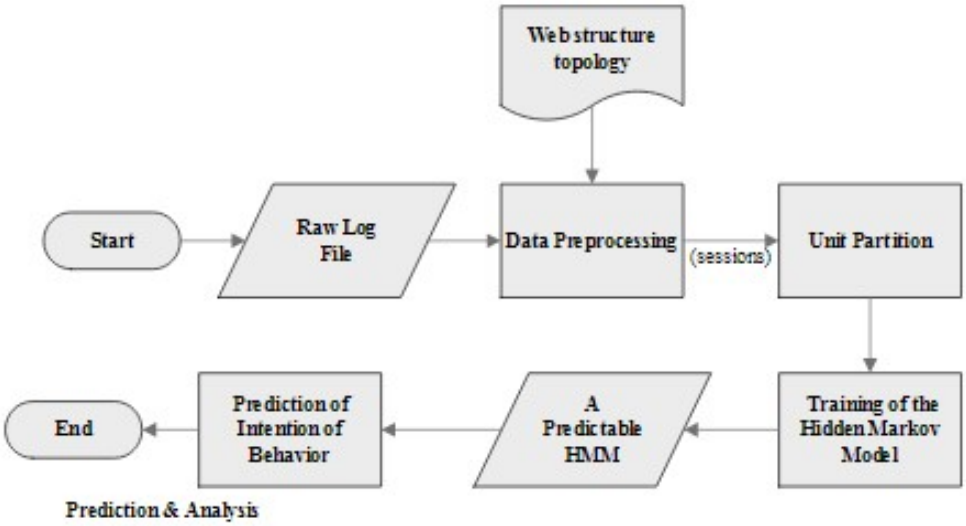

Figure 2: The flowchart of using HMM to predict the intention of consumer [14]

The Hidden Markov Model is characterized by the following:

- N, the number of hidden states in the model. The states are connected such that any state can be reached from any other state. The states are denoted as S= {S_1, S_2, S_3,…..,S_n} and the state at time t is q_t.
- M, the number of distinct observation symbols correspond to the physical output of the system being modeled. The individual symbols are denoted as V= {V_1, V_2, V_3,…..,V_m}.
- The state transition probability distribution A= {A_{ij}} where A_{ij}= P[q_{t+1} =S_j| q_t = S_i] , 1≤ I, j≤ N [17]. For the special case where any state can reach any number of states in a single step, we have a_{ij}>0 for all i, j. For other types of HMM, we would have a_{ij} = 0 for one or more (i, j) pairs.
- The observation symbol probability distribution in state j, B ={b, {k}}, where $B_i(k) = P[V_k + t|q_t = S_j]$, $1 \le j \le N$, $1 \le k \le M$
- The initial state distribution $\pi = \{ \pi_i \}$ where $\pi_1 = P[q_1 = S_j]$, $1 \le i \le N$

This allows to generate an observation seq. O = O1O2...OT

- Set t = 1, choose an initial state q1 = Si according to the initial state distribution π
- Choose O_t = V_k according to the symbol probability distribution in state S_i, i.e., b_{ik}
- Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i, i.e. a_{ij}
- Set t = t + 1, if t < T then return to step 2

F Evaluation Parameters

(i) Prediction Accuracy: Prediction accuracy is a very important parameter for the prediction model. It measures the accuracy of the prediction model applied. Prediction accuracy can be derived using (5):

$$\text{Prediction Accuracy} = \frac{\text{(Number of correct prediction)}}{\text{(Number of test sessions)}} \quad \ldots\ldots(5)$$

Number of correct prediction = the number of test user navigation sessions which are correctly predicted.
Number of test sessions = the total number of test sessions on which prediction is performed

(ii) **Coverage**: The coverage of the model is defined as the ratio of number of times a model is able to predict number of requests in test set.

IV RESULTS AND DISCUSSION

This section is mainly concerned with the implementation, results and analysis of the model. The research focuses on developing a recommendation model to improve the efficiency of web caching in real time that has a more efficient way of making suggestions based on previous web usage in order to optimize user's experience. The result presentation is in two phases; the first phase focuses mainly on the results of the algorithms used while the second phase focuses on the evaluation and benchmarking of the results with that of existing models.

A K-Means Clustering Result

The algorithm partitioned web sessions into clusters based on the attributes and similarities of the webpages. The objective was to find the cluster centroid for each cluster using similarity function and Euclidean distance to all objects in that cluster. Figure 3 shows the result of the clustering of a user's web session on the webpage.

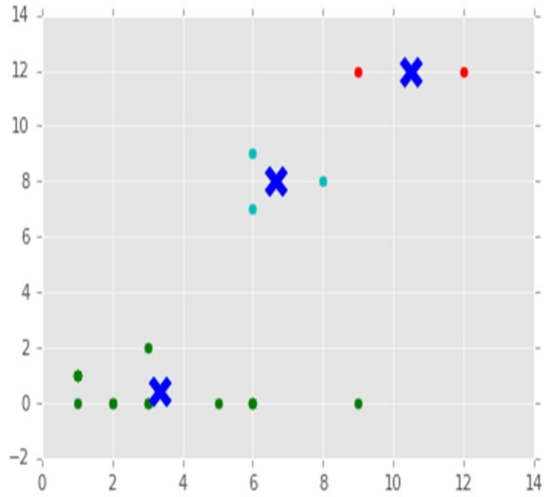

Figure 3: K-means clustering result.

B Page Rank Algorithm

The ranking for the webpages in the first cluster were calculated as;
Page 4: 0.2732
Page 10: 0.2522
Page 8: 0.2495
Page 5: 0.1278
Page 6: 0.0474
Page 7: 0.0250
Page 9: 0.0250

The mean value was derived to be 0.1372.
Figure 4 is a graph representing the page rank result.

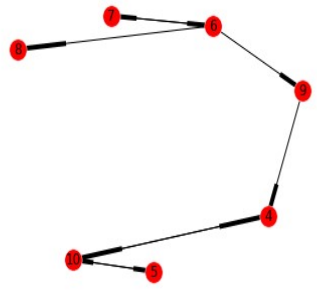

Figure 4: Page rank Graph

C Hidden Markov Model

The model treats each page as a node and the transition/edge between the nodes represents the probability that each page followed from the previous one. This was calculated from the frequency of page transition from our dataset. For example, the probability that page 4 follows page 1 in the sequence is quite high. To improve the model, we kept adjusting the scores for the states and transition probabilities by increasing the frequency with the model using an HMM adaptation of the expectation maximization algorithm. In the expectation step, we calculated all of the page transition from a particular page, summed the scores, and then calculated the probability of each page. Each state and transition probability was then updated by the maximization step of the algorithm to make the model better predict the next page.

D Model Evaluation

Table 1 shows data representation of transition probability and prediction accuracy. There were two sets of data. In order to evaluate the model for prediction accuracy, the test data was used, then an attempt was made to predict the next page by choosing any page as the current page.

Table 1: Model Evaluation

Current web page	Predicted web page	Transition probability	Prediction accuracy
1	**4**	**0.66**	**0.75**
5	**10**	**0.51**	**0.60**
6	**8**	**0.50**	**0.80**
2	**8**	**0.33**	**0.40**
9	**5**	**0.30**	**0.50**

This study has shown that Hidden Markov model and Clustering can work together and provide better prediction results without compromise to accuracy. Markov model is the most commonly used prediction model because of its high accuracy. In comparison to the recommendation model discussed in the article of Nigam [11], the author combined Dynamic Nested Markov Model with PageRank, prediction is still not accurate as irrelevant pages were recommended to users. Meera & Shaikh, [18] compared the accuracy of HMM, Markov Model and Hybrid Model, it can be concluded that HMM has high prediction accuracy.

K-means algorithm was chosen to resolve the problem of worse prediction accuracy of higher Markov model. It is a powerful method for arranging users' session into clusters according to their

similarity. Also to streamline the number of predicted page and enhance the accuracy of HMM, PageRank was used to remove the irrelevant pages.

V CONCLUSION

This study has shown that Markov model and Clustering can work together and provide better prediction results without compromise with accuracy. Hidden Markov Model (HMM) technique can be used to overcome the issues of web page prediction. This powerful combination of algorithms in this study has its drawback in terms of computational complexity but also leads to higher prediction accuracy. With a trade-off in time complexity, the study has shown that HMM is more accurate for predicting navigational paths. The finding in this study is a good guide for efficient web page prediction.

References

[1] El-Yazeed, N. M. (2012). A Survey on Web Recommendation Systems Based on Web Usage Mining. *Institute for Management and Computer, Port Said University, Egypt*, 1-6.

[2] Selvan, M. P. (2012, march). Survey on Web Page Ranking Algorithms. *International Journal of Computer Applications, 41*, 0975 – 8887.

[3] Maurya, J., Sandeep, S., Harish, P., & Pinki, J. (2014). A Survey on: Methods of Web Behavior Prediction by utilizing Different features. *International Journal of Advanced Research in Computer Science and Software Engineering, 4*(3), 207-210.

[4] Chimphlee, S., Salim, N., Ngadiman, M. S., Chimphlee, W., & Srinoy, S. (2006). Rough Sets Clustering and Markov Model for Web Access. Prediction. *Proceedings of post graduate annual seminar*, 470-475.

[5] Dhyani, D., Bhowmick, S. S., & Ng, W. (2006). Modelling and predicting web page accesses using Markov Processes. *IEEE, Computer Society*, 332-336.

[6] Khalil, F., Li, J., & Wang, H. (2006). A framework of combining Markov model with association rules for predicting web page accesses.

[7] Maratea, A., & Petrosino, A. (2009). An Heuristic Approach to Page Recommendation in Web Usage Mining. *Ninth International Conference on Intelligent Systems Design and Applications*, (pp. 1043-1048). Pisa, Italy.

[8] Nigam, B., & Jain, S. (2010, November). Generating a New Model for Predicting the Next Accessed Web Page in Web Usage Mining. *Third International Conference on Emerging Trends in Engineering and Technology, 1*(1), 485-490.

[9] Anitha, A. (2010). A New Web Usage Mining Approach for Next Page Access Prediction. *International Journal of Computer Applications, 8*(11), 7-10.

[10]Rao, V. V., and Kumari, V. V. (2011). An Efficient Hybrid Successive Markov Model for Predicting Web User Usage Behavior using Web Usage Mining. *International Journal of Data Engineering (IJDE), 1*(5), 43-62.

[11] Nigam, B. (2015, February). Evaluation of Models for Predicting User's Next Request in Web Usage Mining. *International Journal on Cybernetics & Informatics (IJCI), 4*(1), 1-9.

[12] Soumen, S., Anjali, T., & Debopriya, P. (2016, April). Enhanced Model of Web Page Prediction using Page Rank and Markov Model. *International Journal of Computer Applications (0975 – 8887), 140.*

[13] Phyu, T. (2013, March). Proposed Approach For Web Page Access Prediction Using Popularity And SimilarityBased Page Rank Algorithm. *International Journal of Scientific & Technology Research, 2*(3), 240-245.

[14] Chun-Jung, L., Fan, W., & I-Han, C. (2005). Using Hidden Markov Model to Predict the Surfing User Intention of Cyber Purchase on the Web. *Proceedings of ICSSSM '05. 1.* IEEE Xplore Conference Services Systems ans Services Management

[15] Nikolai, S. (2010, February 15). *Hidden Markov Models*. Retrieved 2016, from Shokhirev.com: http://www.shokhirev.com/nikolai/abc/alg/hmm/hmm.html

[16] Arpad, G., & Adrian, F. (2014). Web Page Prediction Enhanced With Confidence Mechanism. *Journal of Web Engineering, 13*, 507 - 524.

[17] Jelinek, F. (1969, November). A Fast Sequential Decoding Algorithm Using a Stack. *IBM J. Res. Develop, 13*(6), 675 - 685.

[18] Meera, N., & Shaikh, S. (2015). Predicting User's Web Navigation Behaviour Using Hybrid Approach. *Internation Conference on Advanced Computing Technologies and Application* , 3-7.