DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.5196

ISSN No. 0976-5697

Volume 8, No. 9, November-December 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ANALYSIS OF MARKOV MODEL AND ITS APPLICATION IN WEB PAGE ANTICIPATION

Dr. Balkishan (Assistant Professor)
Department of computer science & Application
Maharishi Dayanand University (M.D.U)
Rohtak Haryana

Shivangi Sorout
Department of computer science & Application
Maharishi Dayanand University (M.D.U)
Rohtak, Haryana

Abstract: In digital era of information where accessing information in limited time frame is necessary, then Markov model helps in the anticipation of the web page. Various network obstacles and minimum bandwidth of network can be efficiently used by employing Markov model and its clustering and pre-fetching technique. Markov Model and its clustering mechanisms helps us to make clusters of similar web users having based on same specified constraints and after the application of the Anticipation framework necessary web page can be accessed.

Keywords: Markov Model, clustering, K-means clustering, Fuzzy c-means clustering and web page Anticipation

I. INTRODUCTION

Evolution of the new technologies and the application of the Markov model is turning out to be helpful for the web users who are connected through network. Markov model is used in the web page anticipation where it is used as stochastic model for the function of modeling those systems which are changing according to the user needs. Markov model works on the concept of states where it is assumed that whatever be the future state it is solely relied on the current state and not on the states which occurred before it. For the anticipation purpose it is necessary that system must possess Markov property. There is the need of determining the user's navigational behavior for the application of the Markov model because this navigational behavior is stored in web log and serve as input for the purpose of web page Anticipation. General technique for the anticipation of the web pages is the usage of the Markov model of order-k. Low order Markov model is generally preferred over higher order Markov model as low order Markov model has lesser state-space complexity.

A. Generalized Markov chain model:

Markov chain model is defined as the discrete time stochastic process where a random variable is chosen and its values over a period of time is keep tracked at discrete interval of time period. In this case a variable "V" is chosen denoting the state of the system at particular time period 't', where variable t varies from 1, 2,3,3………S. Stationary Markov chain is a unique kind of discrete time stochastic process with different following assumptions kept in mind:

- Probabilistic distribution of the state at time t+1 is depend on the state at time t not on the previous states which leads to the state at time t.
- Transition of state from time t to time t+1 is independent of the time t.

Let Pij denote the probability that the system is in a state j at time t+1 given the system is in state t at time t. If the system has infinite number of states, 1, 2, 3……..s; then stationary Markov chain can be defined as probability Transition matrix as following:

P = p11 p12 ………p1s

P21 p22………. P2s

.

And an initial probability distribution :

Q= q1 q2 ………………….qs

Where qi is the probability that system is in state I at time Q.

For implementing this web page anticipation model, navigational behavior of the current users is stored in the web log files. After the identification of the navigational behavior of current users, Clustering is performed. Clustering is the main ingredient used in the exploratory data mining and commonly used in the statistical data analysis. Main task of clustering is to segregate the group of objects into different groups so that all objects in the one particular group are of similar nature. Scope of clustering is wide and used in the many fields like machine learning, image analysis, information retrieval, bioinformatics, data compression and computer graphics. Clustering models used in the Clustering are connectivity models, centroid based model (k-means algorithm), distribution model (expectation-maximization model) and density model. When the Clusters are formed we have to use the prediction algorithm to predict the next possible states. After that Markov model is applied on the clustering sets. Markov model is a stochastic model which is used for the designing the continuously changing system in variable time period at different time slots. In this model future state is depend not only on the current state but also on the previous states covered. In the Markov model we have to train the model by estimating the transition probability which is denoted by:

Aij =P(Q(t+1)=Si | Qt=Sj)

Where, Aij is the probability of going to the new state Si at time t+1 from state Sj at time t.

The first order Markov model provides us a simple way of to accumulate sequential dependence, however it do not take the aspect of long term memory web navigational behavior. Higher order Markov model are useful for the prediction of navigational path. But with the increase of order of the Markov model subsequently there will be exponential increase in the complexity of state space. In turn we require the huge amount of training data. As the number of states increase, systems which needs to predict fast their prediction accuracy is plummeted to large extent. So the need of the hour is to have that kind of system which predict fast with enhanced accuracy.

B. CLUSTERING:

Basically clustering is an approach where unsupervised machine learning is used. But Clustering may be done with the help of supervised method. Only attribute which distinguish supervised clustering and unsupervised clustering is that whether the pattern used for the training data are not labeled or not. In supervised clustering if new patterns are identified then they are segregated into already existing labeled groups. In unsupervised based clustering we don't have the trained labeled data only way by which they are clustered on the basis of hierarchical and non-hierarchical clustering. Very commonly used clustering we used here is distance measure k-means non-hierarchical, unsupervised clustering which segregate cluster data into even populations. Clustering can be used in different variety of applications, for instance- recommendation engine, market segmentation, social network analysis, search result grouping, medical imaging, Image segmentation and Anomaly segmentation.

K-means Clustering :K means is partition based clustering which uses iterative clustering algorithm which helps in the determination of the local maxima in each iteration of the algorithm. Clustering of this kind is performed by the following way:

1.Firstly we have to specify the required number of clusters i.e K. for understanding this algorithm we can choose k=2 for the five data points in the 2-d space.

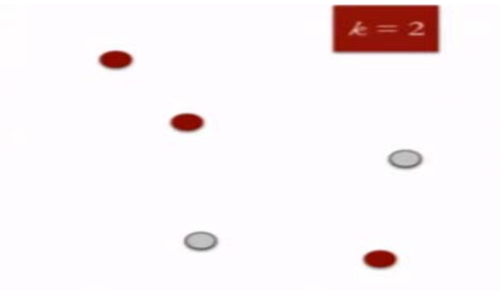
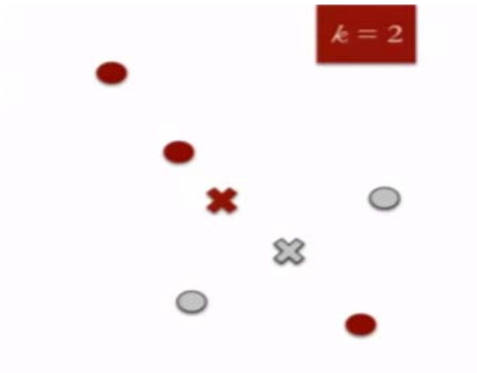


Fig I: K-means clustering

2. After the determination of the number of clusters we have to randomly assign the data points to the cluster. For example: we have assigned three data points in the cluster 1 shown by the help of red color and the two data points with the help of grey color to the cluster 2.
3. After assigning data points to the clusters we have to calculate the cluster centroid; the centroid of the data points of red cluster is shown with the help of red cross and those which are present in the grey cluster is shown with the help of grey cross.



4.

Fig II: K-means clustering

4. Then we have to Re-assign each data point to the centroid of that cluster which is close : Note that data points which are present in the bottom are allocated to the red cluster thought these data points are in proximity to the centroid of grey cluster.

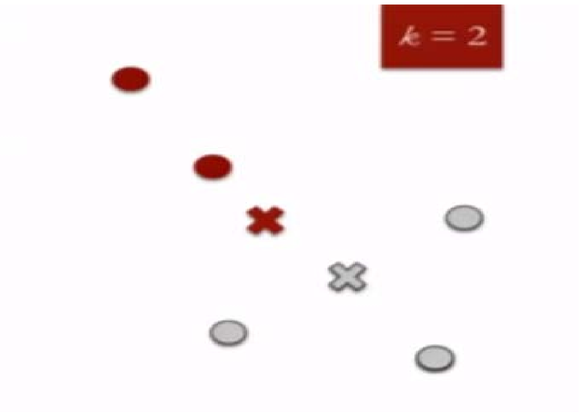


Fig III: K-means clustering

5. Centroids of the clusters are computed again keeping in mind the selection criterions of data point.

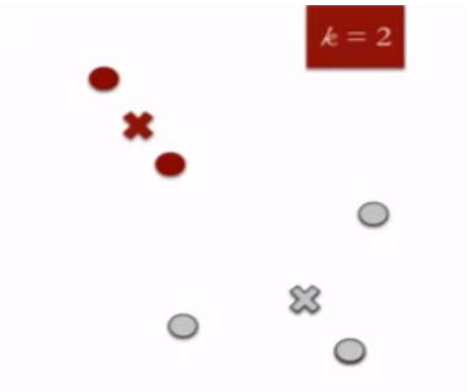


Fig IV: K-means clustering

6. After doing the above required steps we have to repeat the above steps until no improvements are possible. In the same way we have to repeat the above 4 and 5 step until we will reach the global optima. If at the end there is no possibility of further switching of data points between two clusters for two successive review of improvement and the attempt for finding the global optima.

Fuzzy C-means Clustering : Fuzzy c-means clustering is a kind of clustering which have the scope of the data points to be able to assign to two clusters but this can be done with the help of some previously decided membership criterion. In the fuzzy c-means clustering we have to determine the membership function for the assignments of data points to the two clusters up to the some extent.

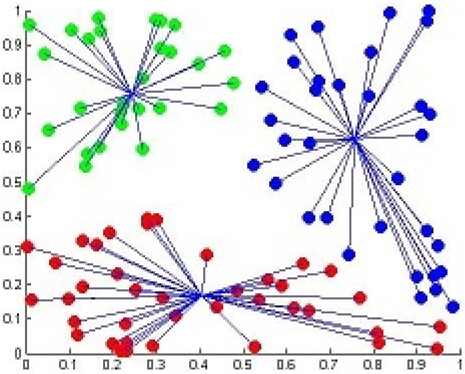


Fig V: Fuzzy c-means clustering

Advantages

1) Best result are achieved for overlapped data set and is computed efficiently and easily than complicated process involving K-means algorithm.

.

2) In k-means clustering every data point is uniquely allocated to the particular cluster but in fuzzy c-means clustering every data point is also allocated membership so that data point can be allocated to two or more clusters simultaneously.

Disadvantages

1) A prior specification of the number of clusters.

2) With lower value of β we get the better result but at the expense of more number of iteration.

3) Euclidean distance measures can unequally weight underlying factors.

II. RELATED WORK

Analysis and review of the previous work is indispensible so that redundant work do not take place and for guiding the work in the specific direction for achieving the desired results. Different authors used Markov model for different kind of Anticipation and other related problems for improving the user experience. Yang, Li and Wang in their research work gives the concepts of the Association rule based method for the prediction of the web request of the user. In his work they used two different kinds of the dimensions for the purpose of making the framework of anticipation model which solve the user needs. The first kind of dimension uses the different kinds of representation methods such as: subset rules, subsequence rules, latest subsequence rules, substring rules and latest substring rules. The second dimension consists of the methods and mechanism of rule selection and mainly selection methods like longest-match, most-confident and pessimistic-selection are used. [1].

Yang et al. (2004) gives overview about the various association rules for web page Anticipation. Different representations of association rules which are: Subset rules, Subsequence rules, Latest subsequence rules, Substring rules and Latest substring rules. The author concerned the precision of these five association rules representations using different selection methods, the latest substring rules were proven to have the highest precision with decreased number of rules [2].

Liu et al. (1998) introduced a customized marketing based on the web approach using a combination of clustering and association rules. The author collected information about customers using forms, Web server log files and cookies. It categorized customers according to the information collected. K-means clustering algorithm uses only the numerical data so due to this reason author used the concept of Partitioning around mediods (PAM) algorithm which works on the process of making the clusters using the categorical scales. Then perform association rule techniques on each cluster [3].

Kim et al. (2004) introduced combination of all three models together. It improves the performance of Markov model, sequential association rules, association rules and clustering by combining all these models together. For instance, Markov model is used first. When the Markov model does not involve or consider the active state then in turn sequential association rules are used to cover those states. If sequential association rules cannot cover the state then after that association rules are used and also if the association rules doesn't cover the remaining state then after that clustering algorithm is applied. The author's work improved recall and it did not improve the Web page prediction accuracy [4].

Vakali et al. (2003) categorized web data clustering into two classes (I) users' sessions-based and (II) link-based. Clustering based on user's session uses web log data and then based on criterion of similar characteristics, web log data are integrated into one particular group. Web log data used gives the information about the transaction performed by the user from that moment user become that part of the network to that moment when he leave the network. Records of all transaction performed by user is stored into the web log file. Each record in the log file involves a unique entity which gives the information about the client's IP address, date and time when the transaction is performed and when the final demand of the user is meet and some other information like size of the object requested, type of protocol is also stored. [5].

A. *Application in web page Anticipation*

Web Server HTTP Request Anticipation: Important application of the anticipation model is the usage of the probabilistic link Anticipation for anticipation of HTTP request. For the enhanced server performance and throughput enormous work has been done on the large scale. The work performed so far mainly consists of the probabilistic analysis of file sizes of requested information, needed patterns and paths of requests files and mechanism used for caching so that all task can be performed efficiently in the prescribed time frame. In past various methods like with the help path profiles we can build a sequence of prefix tree paths for the anticipation of the next request. But most importantly probabilistic sequence generation models like Markov chain can tackle all problems in relation with HTTP request anticipation. [6].

Markov chains models can be easily incorporated and easily extendable into server in a very simple way. In this model of Anticipation of the requests on the basis of the previous web navigation history of the client, firstly client sends its information retrieval request to the web server. Web server uses the probabilistic link prediction module for the anticipation of what probability the same user will request further. In this whole process of the Anticipation Markov model is used as the adaptive model which changes its data structure and mechanism according to the user demands. [6].

Adaptive Web Navigation: The second usage of the Markov chain Anticipation model is the usage of probabilistic link prediction for adaptive navigation which fulfills the user demands in minimum time frame. In this method navigation agent is build up which gives us the information of suggested sites which can be the target for the client, (which is already accessing the information quite frequently) with the help of previous navigation behavior of the user. Framework which we used in this current case anticipated link might not be the same link which is currently being accessed. [7].

On the basis of users actual navigation sequences whatever be the anticipated link are given it will include explicit jumps of user between the disjoint web sites. In Use of user specific modeling link, link Anticipation will reside over the client side rather than the server side. If we use framework and architecture implemented link anticiaptor then it will be treated as servlet which will reside in server. Whenever the client click on the needed URL for information access then this information will be transferred to the servlet on the server side which in turn processed the link. After processing of the servlet is done it will give the whole list of possible links that client may request in future. [8].

Tour Generation: Markov chains is also used as guide for the Tour generation module. In Tour generation module firstly input is given in the form of URL which is accessed by the client first time while starting for the information access. In this process by usage of the Markov chains states sequence is generated which is basically a path for the information retrieval. Whatever the states are generated, are presented to the user (client) as a Tour. [9].

Personalized Hub/Authority: Idea of Hubs and authorities are basically applied and implemented on the web graph structure. Hubs are those web sites which are quite popular and easily targetable, easily accessed and give efficient result to the user. Authorities are basically those web sites whose main focus is toward one relevant area and are generally huge repository of information about one particular topic which can be accessed by the clients. Here the word personalized used is related with the scope of certain set of prescribed users and clients and also targetable towards specific websites. [9].

Personalized Hubs/Authorities enhances the whole idea of these specified terms whose main concentration is on particular kind of users and sites with the help of path of traversal patterns. For representation of all these needed data for future access and for analysis of the transition Markov chain transition matrix is build up which his treated as traversal connectivity matrix .

The idea of iterative estimation of Hub and Authority weights using this Markovian transition matrix can be applied to extract the prominent personalized hubs/authorities. The algorithm is similar to the one described in with the important difference of initialization of the Hub and Authority weights using the transition probabilities specified by the Markov chain transition matrix .

III. CONCLUSION

The Markov Model provides the framework which is used for the Anticipation of the web page by using the user's currently accessed web page. Most of the researches used it to improve the user navigational behavior by anticipating the web page in minimum time slot and by efficiently using the minimum resources of the network. So to mitigate the challenges in web page Anticipation, Markov model in association with the clustering and other pre-fetching mechanisms.

IV. REFERENCES

[1] Q. Yang, T. Li and K. Wang, "Building Association -Rules Based Sequential Classifiers for Web-Document Prediction,"Journal of Data Mining and Knowledge Discovery,Vol. 8, 2004.

[2] Jia Yang, J. Zhang, and K. Beach, "A Survey of Web Caching Schemes for the Internet", ACM SIGCOMM, 2004.

[3] M. Liu, M. Junchang, and G. Zhimin, "Finding Shared Fragments in Large Collection of Web Pages for Fragment-based Web Caching", Second IEEE International Symposium on Network Computing and Applications (NCA'06), 1998.

[4] D. Kim, N. Adam, V. Alturi, M. Bieber, and Y. Yesha, "A Click Stream based Collaborative Filtering Personalization Model: Towards a better performance". WIDM '04, pages 88–95, 2004.

[5] W. Vakali, S. Yu, and D. Cai, "Improving pseudo- relevance Feedback in Web Information Retrieval using Web Page Segmentation", In Proceedings of the Twelfth International World Wide Web Conference, WWW2003, pp. 11-18, Budapest, Hungary, May 20-24, 2003.

[6] J. Zhu, J. Hong, and J. G. Hughes, "Using markov models for web site link prediction", HT'02, USA, pages 169–170, 2002.

[7] Chen M. S. Park J. S.,and Yu P.S., "Data mining for path traversal patterns in a web environment", In ICDCS, pages 385-392, 1996.

[8] M. Eirinaki and M. Vazirgiannis, "Usage-based PageRank for Web Personalization", In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05), Lousiana, 2005.

[9] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques", TOIS, 20(4):422-446, Oct. 2002.