International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-6)

Research Article | **June 2014**

Model 2 Implementation of Forecast System for modeling user's web-browsing conduct

Ajit Patil, Prof.P.A Jadhav, Prof.S.Z.Gawali
Department of IT,
BVDUCOEP, India

Abstract—

Predicting the next page to be accessed by Web users has attracted a large amount of research work lately due to the positive impact of such prediction on different areas of Web based applications. Major techniques applied for this intention are Markov model and clustering. There are two types of Markow Model low and higher order. Markow Model of low order consist of with low accuracy, while Markow Models of high order are associated with high state of space complexity. On the other hand, clustering methods are not used for classifications as they are unsupervised methods. This paper involves incorporating clustering with low order Markov model techniques. Meaningful clusters are created by dividing pre-processed data and these are used as training data for performing 2nd order Markov model techniques. Different distance measures of k-means clustering algorithm are examined in order to find an optimal one. Experiments reveal that incorporating clustering of Web documents according to Web services with low order Markov model improves the web page prediction accuracy.

Keywords— Markov, behaviour, K-means, unsupervised k-means.

I. INTRODUCTION

Today the world is totally depending on the web. Which results in increase of digital data on the Web, due to this there overwhelming amount of research in the area of Web, and also there is research in user browsing personalization and next page access prediction. There is not a single theory or approach related to handling large and increasing amount of data with improved efficiency, performance and accuracy, since this issue is complicated. Two of the most common approaches used for Web user browsing pattern prediction are Markov model and clustering. These approaches has lots of disadvantages and limitations. Because of high accuracy in predictions Markow model is used. Low order Markov models have higher accuracy and lower coverage than clustering. In order to overcome low coverage, all-k^{th} order Markov models have Been used where the highest order is first applied to predict a next page. The order is decreased by one, If it fails to predict the page, until prediction is successful. The coverage is increased , but it is associated with higher state of space complexity On the other hand, clustering methods are not used for classifications as they are unsupervised methods. However, proper clustering groups users' sessions with similar browsing history together, and this facilitates classification. Instead of actual sessions clustering is performed on the cluster sets. Clustering accuracy is mainly depends on the proper selected features for partitioning. For example, partitioning which is based on semantic relationships or link structure or contents usually provides higher accuracy than partitioning based on frequency, time spent or bit vector. However, there is limit for even the semantic, contents and link structure accuracy is limited due to the unidirectional nature of the clusters and the multidirectional structure of Web pages. This paper involves implementation of a clustering algorithm where Web sessions are partitioned into clusters and then Markov model techniques are applied based on the clusters for accuracy and better performance of access prediction of next page. Section 2 mainly concentrates at at previous literature in the field of Markov model techniques along with combining clustering. Section 3 mainly revolves round the process which is acquired to achieve better prediction of next page. In section 4, we prove our new process experimentally and section 5 concludes our work Set.

II. LITERATURE

Predicting the next page to be accessed by the web user uses two frameworks like Markov model and clustering. Many research papers used , Markov model or a combination of both techniques to address Web page prediction by using clustering. Kim et al. combine most prediction models (Markov model, sequential association rules, association rules and clustering) in order to improve the prediction recall. Web mining techniques are use in the the proposed model . However, the new model solely depends on many essential and effective factors, like the confidence thresholds, existence of a Web site link structure and the support. These are the major factors which affect the order and the performance of the applied models and the new model. Cadez et al. on the other hand, used the different approach and combined first order Markov model with clustering. They implemented first order Markov model using the Expectation-Maximization algorithm where they partitioned site users using a model-based clustering approach. They displayed the paths for users within each cluster after partitioning the users into clusters, Our work is not a model based but distance based and we used Markov model for prediction rather than clustering. In another paper the authors construct Markov models from log files and they use co-citation and coupling similarities for measuring the conceptual relationships between Web pages that combines both Markov model and clustering techniques for Web page link prediction. To Cluster conceptually related pages Citation Cluster algorithm is then proposed. A hierarchy of the Web site is constructed from the clustering results. The

authors then combine Markov model based link prediction to the conceptual hierarchy into a prototype called ONE to assist users' navigation. The authors implement a hierarchical clustering technique which could lead to running time complexity with large Web log files. Web page prediction performance was improved by previous work, none of the papers showed an improvement in the Web page prediction accuracy. Kim et. al used a combination of models but did not improve the Web page prediction accuracy. Our work proves to outperform previous work in terms of Web page prediction accuracy using a combination of clustering and Markov model techniques. We implement a simple clustering algorithm, k-means algorithm where using different distance measures which can lead to different results. Al the results were analyzed and optimal was chosen.

III. EXISTING METHODOLOGY

Web page prediction means in short is anticipating the next page to be accessed by the user or the link the Web user will click at next when browsing a Web site. For example, what may be chance that a web browser visiting a site that sells computers will buy an extra mouse while buying a laptop? Or, maybe there is a greater chance the user will buy an external usb optical drive instead. Users' past browsing experience is very fundamental in extracting such information. This is when modeling techniques come at hand. For instance, using clustering algorithms, we are able to personalize users according to their browsing experience. Different users with different browsing behavior are grouped together and then prediction is performed based on the data mining and also based on the users' link path in the appropriate cluster. Similar kind of prediction can be in effect using Markov models conditional probability. For instance, if 50% of the users access page D after accessing pages ABC, then there is a 50/50 chance that a new user that accesses pages ABC predicting user intent on web.

IV. MARKOV MODELS

Markov models are becoming very commonly used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages Let P = {p1, p2, …, pm} be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited *l* pages, then prob (pi|W) is the probability that the user visits pages pi next. Page p*l*+1 the user will visit next is estimated We combine clustering and markov model for project implementation Will access page D next. Our work improves the Web page access prediction accuracy by combining both Markov model and clustering techniques. It is based on dividing Web sessions into groups according to Web services and performing Markov model analysis using clusters of sessions instead of the whole data set. This process involves the following steps:

I. Preprocess the Web server log files in a manner where similar Web sessions are allocated to appropriate categories.
II. Analyze and calculate using data mining different distance measures and determine the most effective and suitable distance measure
III. According to the chosen distance measure, Decide on the number of clusters (k) and partition the Web sessions into clusters
IV. Return the data to its uncategorized and expanded state for each cluster .
V. Perform Markov model analysis using whole data set.
VI Find the appropriate cluster the item belongs to for each item in the test data set,.
VII. Calculate 2-Markov model accuracy using the cluster data as the training data set.
VIII. Calculate the total prediction accuracy based on clusters.
IX. Compare the Markov model accuracy of the clusters to that of the whole data set.

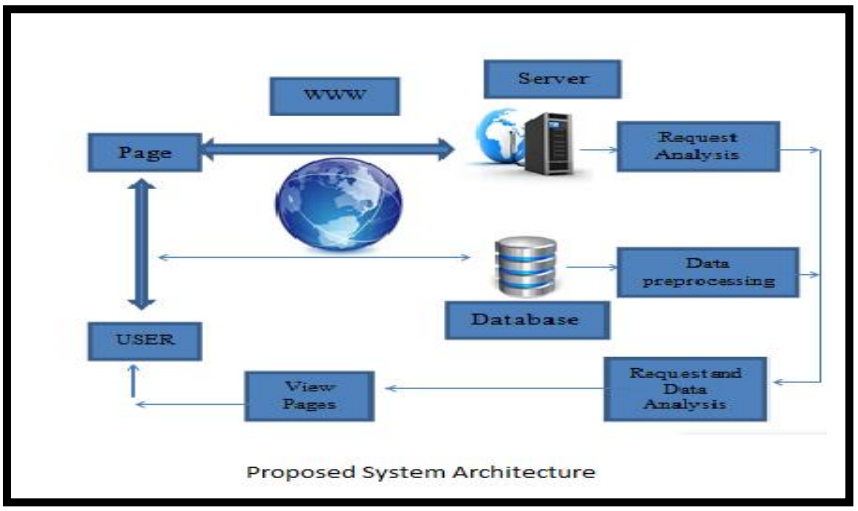
V. FEATURE SELECTION

Before applying clustering techniques and Because of the overwhelming amount of web data, it is very important to group data according to some features. This feature will help us to reduce the state space and will make the clustering task simpler. If the features are not selected appropriately, there is no way we can get good clusters no matter what type of clustering algorithm is used. wang et al. presented different feature selections and metrics that form the base of e-commerce customer groupings for clustering purposes. They examined many features like resource usage, services request, pattern of navigation pattern . The result of their experimentations proved that all features yield similar results and thus, grouping customers according to one of the features selected should do the job. For our purposes, we are grouping the pages, and not users, according to services requested since it is applicable to our log data and is very simple to implement. This yields best results if we group the pages according to services requested.

VI. PROJECT METHODOLOGY

Proposed Methodology Web users are facing the problems of information overload and drowning due to the significant and rapid growth in the amount of information and the number of users. As a result, how to provide Web users with more exactly needed information is becoming a critical issue in web-based information retrieval and Web applications. Though the many websites provides Internet shopping Facilities for convenience, and to save the time of the customers, only 10% customers are using online shopping facility due to some problems like finding relevant information, producing new knowledge, poor performance of the web pages etc. Based on suggestions from literature we developed a methodology shown in above architecture, which can be used for general applications of web usage mining. It is a behavioral adaptation process to the analysis of web data. Web data are a real source to analyze the user behavior in the web. An important step is cleaning and preprocessing of the web data. Personalization of web sites is a very

challenging field of both, current research as well as applications that have as goals e.g. individualized marketing for e-shopping or dynamic recommendations to a web visitor based on his/her profile and usage behavior. Analyzing web data can also be used for system improvements providing the key to understanding web traffic behavior. Advanced load balancing, data distribution or policies for web caching as well as higher security standards are potential benefits of such improvements. Similar analyses could be used for modification of web sites. Understanding visitors' behavior in a web site provides hints for adequate design and update decisions. Business intelligence covers the application of intelligent techniques in order to improve certain businesses, mainly in marketing.



Proposed System Architecture

VII. PROPOSED MODEL

In order to study web user navigational behavior it will be important to clarify the system first. Web users are considered human entities that, by means of a web browser, access information resources in a hypermedia space called the World Wide Web (WWW). Common web users' objectives are information foraging (looking for information about something), social networking activities (e.g. Facebook), e-commerce transactions (e.g. Amazon Shopping), bank operations, etc. On the other hand, the hypermedia space is organized into web pages that can be described as perceived compact subunits called "web objects." The design of web pages is created by "web masters" that are in charge of a group of pages called a "web site." Therefore, the WWW consists of a vast repository of interconnected web sites for different purpose. While current approaches for studying the web user's browsing behavior are based on generic machine learning approaches, a rather different point of view is developed in this thesis. A model based on the neurophysiology theory of decision making is applied to the link selection process. This model has two stages, the training stage and the simulation stage. In the first, the model's parameters are adjusted to the user's data. In the second, the configured agents are simulated within a web structure for recovering the expected behavior. The main difference with the machine learning approach consists in the model being independent of the structure and content of the web site. Furthermore, agents can be confronted with any page and decide which link to follow (or leave the web site). This important characteristic makes this model appropriate for heavily dynamic web sites. Another important difference is that the model has a strong theoretical basis built upon physical phenomenon. Traditional approaches are generic, but this proposal is based on a state-of-the-art theory of brain decision making. The proposal is based on the Markov's Model. The Markov's model simulates the artificial web user's session by estimating the user's page Sequences and furthermore by determining the time taken in selecting an action, such as leaving the site or proceeding to another web page. Experiments performed using artificial agents that behave in this way highlight the similarities between artificial results and a real web user mode of behavior. Furthermore, the performance of the artificial agents is reported to have similar statistical behavior to humans. If the web site semantic does not change, the set of visitors remains the same. This principle enables the predicting of changes in the access pattern to web pages related to small changes in the web site that preserve the semantic. The web user's behavior could be predicted by simulation and then services could be optimized.

VIII. CONCLUSION

The paper gives a brief literature survey of research field in web user browsing prediction.the higher order markov model is studed and found to be best for methodology to implement.

REFERENCES

[1] A. Banerjee and J. Ghosh. Clickstream clustering using weighted longest common subsequences. SIAM Conference on Data Mining, Chicago, pages 33–40, 2001R. Caves, Multinational Enterprise and Economic Analysis, Cambridge University Press, Cambridge, 1982.

[2] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-based clustering and visualization of navigation patterns on a web site.Data Mining and Knowledge Discovery, 7, 2003.

[3] M. Deshpande and G. Karypis. Selective models for predicting web page accesses. Transactions on Internet Technology, 4(2):163–184,2004.

[4] C. F. Eick, N. Zeidat, and Z. Zhao. Supervised clustering -algorithms and benefits. IEEE ICTAI'04, pages 774–776, 2004.
[5] S. Gunduz and M. T. OZsu. A web page prediction model based on clickstream tree representation of user behavior. SIGKDD'03, USA, pages 535– 540, 2003.
[6] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis. Thesus: Organizing web document collections based on link semantics. The VLDB Journal, 2003(12):320–332, 2003
[7] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. ACM Computing Surveys, 31(3):264–323, 1999
[8] D. Kim, N. Adam, V. Alturi, M. Bieber, and Y. Yesha. A clickstreambased collaborative filtering personalization model: Towards a better performance. WIDM '04, pages 88–95, 2004
[9] J. Pitkow and P. Pirolli. Mining longest repeating subsequences to predict www surfing. USENIX Annual Technical Conference, pages 139–150, 1999
[10] R. Sarukkai. Link prediction and path analysis using markov chains. 9th International WWW Conference, Amsterdam, pages 377–386, 2000.
[11] Lars Backstrom, Ravi Kumar, Cameron Marlow, Jasmine Novak, and Andrew Tomkins. Preferential behavior in online groups. In Proc. Int. Conf. on Web Search and Web Data Mining (WSDM) ,NewYork,NY,USA,2008
[12] Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z.: Demographic prediction based on user's browsing behavior, WWW '07: Proceedings of the 16th international conference on World Wide Web , ACM, New York, NY, USA, 2007, ISBN 978- 1- 59593- 654-7
[13] Lee, T.-Y.: Predicting User's Behavior by the Frequent Items, 2007
[14] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining World Wide Web browsing patterns," J. Knowl. Inf. Syst., vol. 1, no. 1, pp. 5–32, 1999.
[15] M. T. Hassan, K. N. Junejo, and A. Karim, "Learning and predicting key Web navigation patterns using Bayesian models," in Proc. Int. Conf. Comput. Sci. Appl. II, Seoul, Korea, 2009, pp. 877–887.
[16] M. Awad, L. Khan, and B. Thuraisingham, "Predicting WWW surfing using multiple evidence combination," VLDB J., vol. 17, no. 3, pp. 401–417, May 2008.
[17] M. Awad and L. Khan, "Web navigation prediction using multiple evidence combination and domain knowledge," IEEE Trans. Syst., Man,Cybern. A, Syst., Humans, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.
[18] I. Zukerman, W. Albrecht, and A. Nicholson, "Predicting user's request on the WWW," in Proc. 17th Int. Conf. UM, 1999, p. 393.
[19] M. Levene and G. Loizou, "Computing the entropy of user navigation in theWeb," Int. J. Inf. Technol. Decision Making, vol. 2, no. 3, pp. 459–476, 2003.
[20] Dembczynski, K., Kotłowski, W., Sydow, M.: Effective Prediction of Web User Behaviour with User-Level Models, 2007.
[21] E. Adar, D. S. Weld, B. N. Bershad, and S. D. Gribble. Why we search: visualizing and predicting user behavior. In WWW , 2007.
[22] Marcelo Maia, Jussara Almeida, and Virg´ılio Almeida . Identifying user behavior in online social networks. In Proceedings of the 1st Workshop on Social Network Systems ,Social Nets'08,pages1–6, NewYork, NY, USA, 2008. ACM.
[23] Richardson, M., Dominowska, E., Ragno, R.: Predicting clicks: estimating the click-through rate for new ads, WWW '07: Proceedings of the 16th International Conference on World Wide Web , ACM, New York, NY, USA, 2007.
[24] LEIVA , L. A. 2011. Mining the browsing context: Discovering interaction profiles via behavioral clustering. In Adjunct Proceedings of the 19th Conference on User Modeling, Adaptation, and Personalization (UMAP) . 31–33
[25] M UELLER ,F. AND LOCKERD , A. 2001. Cheese: Tracking mouse movement activity on websites, a tool for user modeling. In Proceedings of Extended Abstracts on Human Factors in Computing Systems (CHI EA). 279–280.
[26] GUO,Q. AND A GICHTEIN , E. 2012. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In Proceedings of the 21st International Conference on World Wide Web (WWW). 569–578.
[27] HAUGER,D.,PARAMYTHIS, A., AND WEIBELZAHL, S. 2011. Using browser interaction data to determine page reading behavior. In Proceedings of the 19th International Conference on User Modeling, adaption, and Personalization (UMAP).147–158
[28] HUANG,J.,WHITE,R.W., AND DUMAIS , S. 2011. No clicks, no problem: Using cursor movements to understand and improve search. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI). 1225–1234.

AUTHOR

1.Ajit .R. Patil is Research scholar with M.tech information technology from Bharati vidyapeeth university college of engineering pune-46.
2.Pramod Jadhav is Research scholar pursuing Phd in computer science from baharati vidyapeeth university college of engineering and assistant professor at university.

