Journal of Theoretical and Applied Information Technology
31st May 2014. Vol. 63 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A COMPARATIVE ANALYSIS OF MARKOV MODEL WITH CLUSTERING AND ASSOCIATION RULE MINING FOR BETTER WEB PAGE PREDICTION

[1]P.SAMPATH, [2]Dr.AMITABH WAHI, [3]D.RAMYA

[1]Research Scholar, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam

[2] Associate Professor, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam

[3] Assistant Professor, Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore

E-mail: [1]samkala@gmail.com, [3]ramyaduraii@gmail.com

ABSTRACT

Web page prediction, that involves personalizing web users behavior and also helps the web master to improve the website structure and helps the user in navigating the site and accessing the information. World Wide Web is a huge storage place for pages and links. So that the browser can get the information through the storage place. But it takes more time to reach the users targeted page. Intermediatly browsers have to visit many unwanted links instead of targetted page. Here, different techniques has been investigate to predict the next set of webpage based on the previous action of browsers behaviour for which the log files are collected and maintained. Two different predicting techniques namely markov model along with clustering and modified markov model along with association rule mining are applied to find out web page prediction. Thus the comparison of markov model along with clustering and modified markov model along with association rule mining shows better result for page prediction.

Keywords: *Markov Model(MM), Modified Markov Model(MMM), Markov Model with Clustering(IMC), Markov Model with Association Rule Mining(MM-AR),Web Mining.*

1. INTRODUCTION

In the internet era, websites on the internet are useful for gathering information in day to day activities. So there is a development of WWW in its volumn of size and traffic of websites. There are different techniques namely SVM, Markov model, Modified markov model, Association rule mining, markov model with clustering etc .. has been used for web page prediction. web mining is the application of data mining in which the next page can be identified by tracing the users visiting behavior and then extract their interest using patterns. Because of its direct application,web mining has become one of the important areas in computer and information science. Web usage mining is the process of extracting useful information from server logs e.g. Web usage mining is the process of finding out what users are looking for on the internet. Web Usage Mining uses mining methods in log data to extract the users behavior , which is used in various applications.

Log files records information can be viewed as in the form of client IP address, URL requested etc., in different formats such as Common Log format, Extended Common Log format which is issued by Apache and IIS. The outcome of web usage mining can be utilized for target advertisement, enhancing web design, enhancing satisfaction of customer, guiding the strategy decision of the enterprise, and marketing analysis etc. Predicting the users' browsing behavior is one of web usage mining technique. To recognize the user behavior in accordance with analyzing the log data.

These instructions give guidance on layout, style, illustrations and references and serve as a model for authors to emulate. Please follow these specifications closely as papers which do not meet the standards laid down, will not be published.

2. MARKOV MODEL WITH CLUSTERING

Markov model is one of the model used for prediction technique in accordance with

Journal of Theoretical and Applied Information Technology
31st May 2014. Vol. 63 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

clustering. Log files are clustered based on their similarity or dissimilarity (link structure) or time spent on that pages or frequency . Clusters are employed to guide the prediction system. The main issue that affects the clustering accuracy is producing the selected features for partitioning.

2.1 Algorithm

The training process [1] takes place as follows:

1.Use feature selection, allocate similar Web sessions to appropriate categories.
2.Find distance measure using k -means algorithm.
3.Decide the number of clusters k and partition the Web sessions into clusters.
4.FOR each cluster
5.Return the data to its uncategorization form and extend the state.
6.Perform Markov model analysis on each of the clusters.
7.ENDFOR

The prediction process [2] or test phase involves the following:

1.FOR each coming session
2.Find its closest t cluster
3.Use the Markov model to make page prediction
4.ENDFOR

2.1.1 Feature Selection

The improved Web personalization is subject to proper preprocessing of the usage data [7], [8]. It is very important to group data according to some features before applying clustering techniques. This will reduce the state space complexity and will make the clustering task simpler.

2.1.2 Session Categorization

Consider a data set D containing N number of sessions. Let W be a user session including a sequence of pages visited by the user in a visit. D= {W1… WN}.Let P = {p1,p2… pm} be a set of pages in a Web site.

Table 1: Number of Categories

	Dataset 1
#session	2,520
#categories	196

After identifying all categories for each data set, it was necessary to run the session categorization algorithm below. The categories are formed as follows:

Input: Dataset(D) containing N number of sessions WN.
(1) FOR each page pi in session Wi
(2) IF pi Ci
(3) wi.count++
(4) ELSE,
(5) wi = 0
(6) ENDIF
(7) ENDFOR
Output: Ds (Dataset) containing N number of Sessions SN

Table 2: Session Categorization

1	2	3	4	6	8	9	10	15	19	23
0	0	0	0	3	0	1	0	0	2	1
0	0	0	1	0	0	0	0	5	0	1
1	0	0	0	0	4	0	0	0	0	0

Therefore, the number of clusters used for each data set was a result of applying k-means algorithm to the data set and, then applying ISODATA algorithm to the resulting clusters. We achieved best results for D1 when k=7.All clusters were attained using Cosine distance measure. To help find an appropriate k-means clustering distance measure we can apply to data sets D1, we examine the work presented by [10], [11]. By finding the average of distance values between points within clusters and their neighboring clusters. The mean value should be higher, then we can get the better cluster. For that in the dataset1,cosine distance measure has been used. The Markov model accuracy increases if the Web sessions are well clustered due to the fact that more functionally related sessions are grouped together. This grouping of Web sessions into meaningful clusters helps increase the Markov model accuracy.

Table 3: Comparison Table Of MM With IMC

	D1
MM	72,524
IMC	11,682

Figure 1 compare the accuracy of MM and the integration of Markov model and clustering (IMC) for the data sets using Cosine distance measures for the clusters with k=7 for dataset D1.

Journal of Theoretical and Applied Information Technology
31st May 2014. Vol. 63 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

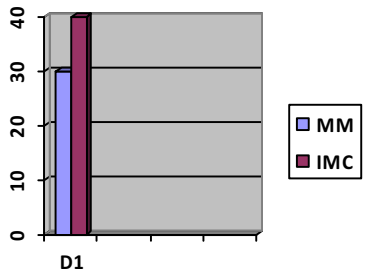


Figure 1: Prediction Accuracy Of MM With IMC

3. MARKOV MODEL WITH ASSOCIATION RULE MINING

In markov model, prediction technique is used to predict the set of pages in building prediction model to reduce its size. Considering all the sessions (P1, P2, P3), (P1, P3, P2), (P2, P1, P3), (P2, P3, P1), (P3, P1, P2), and (P3, P2, P1) as one set P1, P2, P3. Main task for the browser on the Web can be done using different paths regardless of the ordering that the users choose. In addition, we have to reduce the size of prediction model by eliminate the sessions that have repeated pages. It might result when the user accidentally clicks on a link and hits the back button. Two-Tier Framework Training Process as:

Input: M is the set of prediction model of size N: T is a set of training examples
Output: A set of trained classifiers and an example Classifier EC.
1.For each model(m) in M train m on T
2.For each training example e in T and a classifier model m in M Do if m predicts the target of e correctly then map e to m and record the confidence of m in prediction.
3.For each example e in T, if e is mapped to more than one model then filter the labels so that only one label is kept.
4.Train EC on the training set T', where T' is a training set that has all examples in t and each example is mapped to one model in M

Note that the last page of the session is assumed to be the final destination and it is separated from the sessions.

ARM is a data mining technique that has been applied successfully to discover related transactions. Specifically, ARM focuses on associations among frequent itemsets. Consider an example as supermarket store, ARM helps uncover items purchased together which can be utilized for shelving and ordering processes. In the following, we briefly present how we apply ARM in WPP.

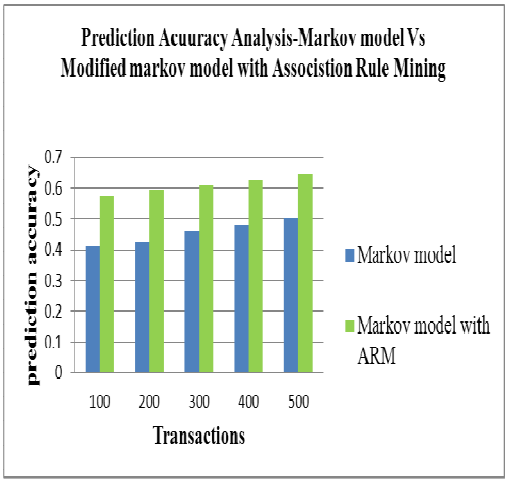
In WPP, prediction is conducted according to the association rules that satisfy certain support and confidence as follows. For each rule, R=X→Y , of the implication, that is X denotes as user session and Y denotes the target destination page. Prediction is resolved as follows:

By setting the cardinality value greater than one, then only prediction can resolve to more than one page. Moreover, setting the minimum support plays an important role in deciding a prediction. In order to mitigate the problem of no support for X $\cup$ Y, we can compute prediction (X' → Y),

Table 3: Prediction Accuracy Analysis-Markov Model Vs Modified Markov Model With Association Rule Mining

Transactions	Markov model	Markov model with ARM
100	0.415	0.573
200	0.423	0.594
300	0.457	0.608
400	0.481	0.626
500	0.502	0.647

Figure 2: Prediction Accuracy Analysis-Markov Model Vs Modified Markov Model With Association Rule Mining



4. CONCLUSION

Thus the paper has been investigated with two different technique for finding out the better web page prediction. The web pages in the user sessions are allocated into categories according to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

web services. Then k-means clustering algorithm is used in the most appropriate number of clusters and distance measure. Markov model techniques are applied to each cluster as well as to the whole data set 1. The web access pattern mining and prediction scheme is analyzed with different log files, which are collected from the web servers. The system is tested with markov model and modified markov models with association rule mining. The prediction accuracy is used as the performance metric to evaluate the quality of the system. Finally thus the paper analysed as markov model with clustering provides a better web page prediction in accordance with modified markov model with association rule mining based on their prediction accuracy. In future we extend our work with the comparision modified markov model, Hierarchical Agglomerative Clustering, Fuzzy Possibilistic Clustering and boosting and bagging model for finding better prediction.

REFRENCES:

[1] Awad.M,Khan.L and Thuraisingham.B, "*Predicting WWW surfing using multiple evidence combination*," VLDB J., vol.17, no.3, May 2008, pp.401–417.

[2] Hassan.M.T, Junejo.K.N and Karim.A, "*Learning and predicting key Web navigation patterns using Bayesian models,*" in Proc.Int.Conf.Comput.Sci.Appl.II, Seoul, Korea, 2009, pp. 877–887.

[3] Mamoun Awad.A and Issa Khalil, "*Prediction of User's Web-Browsing Behavior: Application of Markov Model*" IEEE Transactions on Systems, Man and Cybernetics-Part b:Cybernetics, vol.42, no.4, August 2012.

[5] Trilok Nath Pandey, Ranjita Kumari Dash , Alaka Nanda Tripathy ,Barnali Sahu, "*Merging Data Mining Techniques for Web Page Access Prediction: Integrating Markov Model with Clustering*" , IJCSI , Vol. 9, no 1, November 2012.

[6] Eirinaki,M.,Vazirgiannis,M. & Kapogiannis, D.(2005),'*Web path recommendations based on page ranking and markov models*',WIDM'05 pp. 2–9.

[7] Banerjee,A. & Ghosh,J.,'*Clickstream clustering using weighted longest common subsequences'*, SIAM Conference on Data Mining, Chicago pp. 33– 40, 2001

[8] Eirinaki,M.,Lampos,C.,Paulakis,S. & Vazirgiannis,M ,'*Web personalization integrating content semantics and navigational patterns*', WIDM'04 pp. 2–9,2004.

[9] Nasraoui.O and Petenes.C,"*Combining Web usage mining and fuzzy inference for Website personalization*," in Proc.WebKDD, 2003, pp.37–46.

[10] Halkidi,M., Nguyen,B., Varlamis,I. & Vazirgiannis,M,'*Thesus: Organizing web document collections based on link semantics'*, The VLDB Journal 2003(12), 320–332.

[11] Strehl,A.,Ghosh,J. & Mooney,R.J,'*Impact of similarity measures on web-page clustering'*, AI for Web Search,2000,pp. 58–64.