ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

Dimensionality Reduction using Markov Model for Web Personalisation

Varun Hooda[1], Mamta Kathuria[2]

Department of Computer Engineering, YMCA University of Science and Technology, Faridabad, Haryana, India[1,2]

Abstract: One of the important applications of web mining is web personalization. Markov model is quite helpful technique in this regard. But this very model has limitation of large size data requirement and high dimensionality. This paper has efficiently dealt with the problem of high dimensionality in Markov model. Dimensionality is reduced by analyzing the users' access behaviour and clustering of web pages based on similar navigational pattern.

Keywords : Web mining, Markov model, Dimensionality, Personalization, Clustering, Analysis.

I. INTRODUCTION

Web mining is the technique of extracting information from WWW (World Wide Web). WWW is one of the most comprehensive resources of information. Information required by almost every user is contained in it. Usefulness of data retrieved by user through data mining is measured by the relevancy of information obtained. Various techniques of obtaining the useful data from WWW are included in web mining. Web mining can be broadly classified into three types [1,2]:

a) Web Structure Mining
b) Web Content Mining
c) Web Usage Mining

Web usage mining applications include web personalization. Web personalization analyses and modifies the users' interface according to him. This helps in users' convenience for visiting the web site. One of the techniques helpful for web personalisation is Markov model.

II. RELATED WORK

The main data source in the web usage mining and personalization process is the information residing on the web site's logs. Web logs record every visit to a page of the web server hosting it. The entries of a web log file consist of several fields which represent the date and the time of the request, the IP address of the visitor's computer (client), the URL requested, the HTTP status code returned to the client, and so on. The web logs' file format is based on the so called "extended" log format, proposed by W3C [4]. Web usage mining includes the procedure of identification of representative trends and browsing patterns describing the web site activity by analysing the users' behaviour. Various methods are available for analysing the web log data. Researchers have been using well known data mining techniques like association rules discovery [5], sequential pattern analysis [6, 7], clustering [8], probabilistic models [9, 10], or a combination of them [11,12].

III. WEB PERSONALISATION USING MARKOV MODEL

Web personalisation is achieved using Markov model on the basis of past history of access of user. Markov Model enables us to study and analyze the large amount of web log and use the huge amount of big data for web mining and personalization. Markov Model is used to predict the users' next behaviour based on his history of actions.

A. Markov Model

Markov Models [3,13] can be considered as an application of Markov chain in the navigation of a user. Markov model is useful in predicting the next action of the user based on the previous actions. In context of web page accesses, this model can be used for predicting the possibility of access of a web page based on history of web page accesses. But Markov model suffers from some limitations, those are:

B. Limitations of Markov Models

(i) Large size input data requirement

Basis of Markov Models is statistical processes, so the prediction is dependent on quantity of the available web log. Hence need for very large size of input data is a shortcoming of the Markov Models.

(ii) High Dimensionality Problem

Transition matrix created from transition graph is usually of very big size due to more number of web pages. Problem of dimensionality can be reduced by clustering of similar web pages.

C. The Transition Graph

The transition graph of a user or a group of users gives the past access to the web pages and also the frequency of visit from one web page to another. Transition graph can be considered as a weighted graph, nodes representing the web pages and weights of edges represent the frequency of page visit between those two pages connected by a hyperlink.

The transition graph in Fig.1 gives the exact view of a users' web pages history in a course of time[3]. Here nodes of the graph are web pages, numbers indicated 1 through 12. Node 'S' shows start page and node 'E' shows end page of users' access. Weight on the edges shows the number of visit from one web page to another.

The transition probability from page *i* to page *j* in only one step may be computed using a transition graph. Using the weights of the graph links, we can create a transition

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

probability matrix that includes the one-step transition probabilities in the Markov model.

D. Probability Transition Matrix

Transition matrix showing probability of page visit from one page to another can be calculated using the information from transition graph. Probability matrix, A can be calculated using the formula:

$$A(s,s') = \frac{C(s,s')}{\sum_{s''} C(s,s'')} \qquad (1)$$

Here C(**s,s'**) is the count of the number of times **s'** follows **s** in the training data. So to calculate probability from page **s** to **s'**, count from page **s** to **s'** is divided by the total number of counts from page **s** to all other pages.

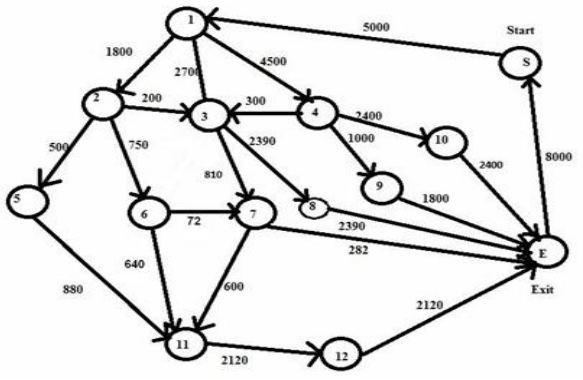

Fig. 1 The Transition Graph of user Web Page access

IV. PROPOSED WORK

A. Dimensionality Reduction Process Architecture

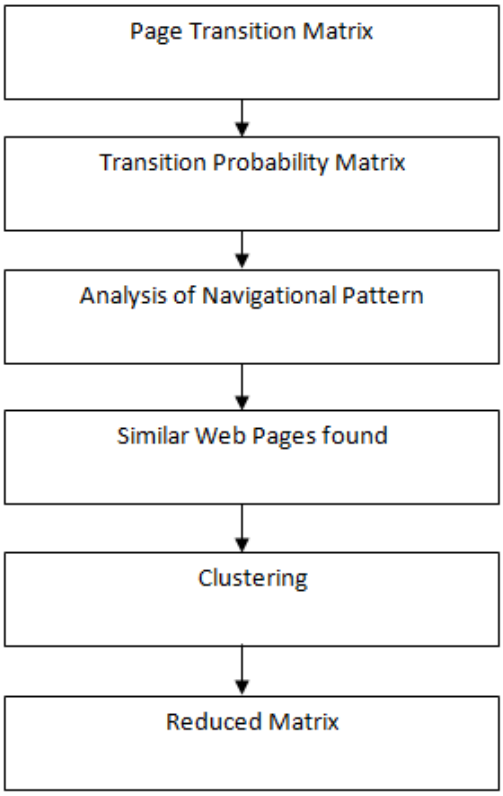

Fig. 2 Architecture of Proposed Dimensionality Reduction Process

Fig. 2 explains the proposed work for the reduction of dimensionality of Markov Model. It can be considered as a six stage process in which output of one stage is used as an input for next stage. These stages are individually as

i. Page Transition Matrix

This is the very first step in which a transition matrix is created which contains the number of transitions user has made from one web page to another. This matrix is created using the web history of user from the transition graph.

ii. Transition Probability Matrix

Transition probability matrix will be created by applying formula of equation (1) to the Page Transition Matrix. Thus the formed Transition Probability Matrix contains the probability of visiting one page from another based on past accesses of user.

iii. Analysis of Navigational Pattern

Navigational pattern of user is studied by analysing the Transition Probability Matrix.

iv. Similar Web Pages found

Analysis in previous stage resulted in a number of web pages found with similar navigational pattern.

v. Clustering

In this stage appropriate clustering technique is applied to the web pages found with similar pattern in users' web navigation.

vi. Reduced Matrix

After clustering, the new matrix is created having lesser size than previous matrix. Hence the reduced matrix is obtained in final stage of dimensionality reduction process.

B. Algorithm: Dimensionality Reduction

1. Create page transition matrix A where A(i,j) is the number of visits from page i to page j.
2. If (A(i,j) = 0)
   A'(i,j)= 0;
   for i=0 to n-1

$$A'(i,j) = \frac{A(i,j)}{\sum_{j} A(i,j)}$$

   //here A'(i,j) is the probability of visiting from page i to page j.
3. Transpose the matrix A' to A''
4. int p=1;
   for i=0 to n-1{
   int count=0;
   for j=0 to n-1{
   if A''(i,j)== 1.0{
   count= count+1
   if count==1
   temp=j;
   if count>1{ //then more than one pages point to a single page
   X[0]=temp;
   x[p++]= j }}}} //X contains the pages to be clubbed
5. Transpose matrix A'' to A' again
6. for i=0 to n-1

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 6, June 2014

{
for j=0 to n-1
{

$$A'(i,x[0]) = \sum_{p=0}^{p} A'(i,x[p]);$$

$$A'(x[0],j) = \sum_{p=0}^{p} A'(x[p],j);$$

}}

7. for i=0 to n-1
int r,s=0;
if (i==x[1])||i==x[2]||….||i==x[p])
continue;
for j=0 to n-1
if (j==x[1])||j==x[2]||….||j==x[p])
continue;
B[r][s] = A'[i][j];
//B is the reduced matrix

V. IMPLEMENTATION

Taking the information from transition graph as an input and using the number of visits between pages in above formula given, probability of visiting each page from one another is calculated. In this way probability matrix is formed.

A. *Probability Matrix*

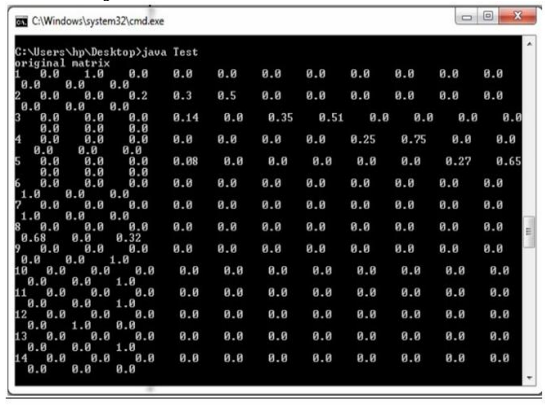

Fig.3. Original Matrix of Probabilities of Transition

Fig.3 is generated probability transition matrix showing probability of visiting of one web page from another. These probabilities are calculated on the basis of access history of user taken from transition graph of web navigation of user.

Obtained probability transition matrix is analysed and found that matrix for large transition graphs will have very high dimensionality. Hence it is required to reduce the high dimensionality of Markov model to some extent.

B. *Dimensionality Reduction in Probability Transition Matrix*

Navigational pattern of user can be analysed and used for reduction in size of transition matrix. Web pages having similar navigational pattern can be clubbed to make clusters. Clustering of similar web pages based on navigational pattern will help in decreasing the size of probability matrix. Hence similar web pages are clubbed and hence finally reduction in dimensionality of Markov model is achieved. Size of reduced matrix is lesser than that of original probability matrix.

C. *Reduced Dimensionality Matrix*

Fig.4 is reduced matrix of probability of transition of web pages. Size of matrix is reduced by clustering of web pages based on similar navigational pattern.

It has been found that the size of reduced matrix has decreased finally reducing the problem of dimensionality of Markov model to some extent. Size of the matrix has decreased from 14 to 11. As dimensionality is of second order, there is considerable decrease in space complexity of Matrix.

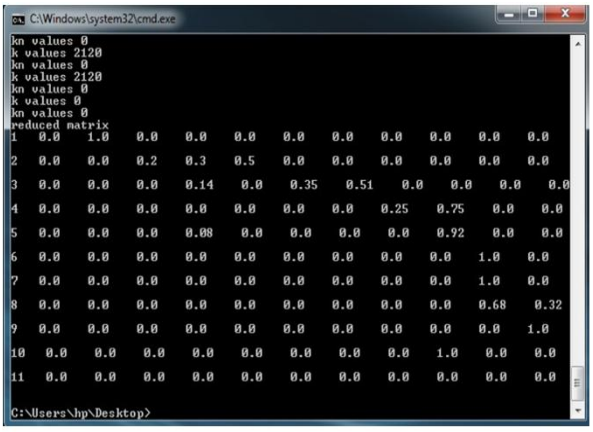

Fig.4. Reduced Transition Probability Matrix

VI. RESULTS EVALUATION

Clustering of Web pages on the basis of navigational pattern has resulted in reduction of dimensionality of Markov model which is depicted with the help of a graph as shown below in fig. 5.

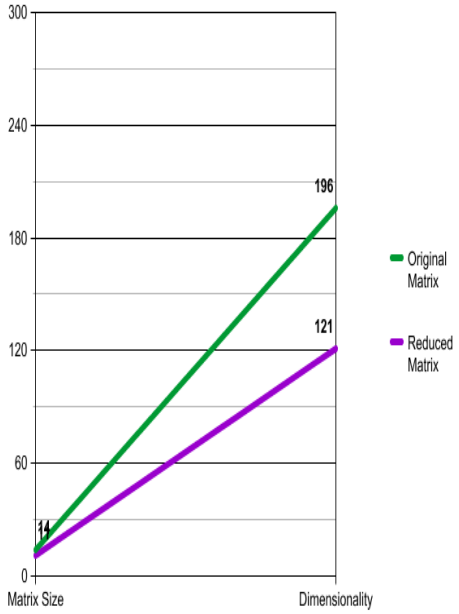

Fig.5. Difference in Dimensionality before and after clustering of Web pages

VII. CONCLUSION

In this paper, we have discussed the Web Personalisation as an important application of Web Mining. Markov model is discussed as its usage in personalisation and its limitations are analysed. High dimensionality problem of Markov model is reduced to some extent by carefully studying the users' access behaviour of WWW. Clustering of web pages visited by user based on similar navigational pattern has resulted in decrease in space complexity of probability transition matrix.

REFERENCES

[1] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.
[2] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from WebData, ACM SIGKDD Explorations Newsletter, January 2000, Volume 1 Issue 2.
[3] J. Zhu, J. Hong, J.G.Hughes, Using Markov Chains for Link Prediction in Adaptive Web sites, in Proceedings of the First International Conference on Computing in an Imperfect World, 2002.
[4] Extended Log File Format, http://www.w3.org/TR/WD-logfile.html
[5] M.S. Chen, J.S. Park, P.S. Yu, Data Mining for Path Traversal Patterns in a Web Environment, in Proc. of the 16th Intl. Conference on Distributed Computing Systems (1996)
[6] B. Berendt, Using site semantics to analyze, visualize and support navigation, in Data Mining and Knowledge Discovery Journal, 6: 37-59 (2002)
[7] A.G. Buchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, J.G. Hughes, Navigation pattern discovery from Internet data, in Proc. of the 1st WEBKDD Workshop, San Diego (1999)
[8] R. Krishnapuram, Anupam Joshi, Olfa Nasraoui, Liyu Yi, Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining, in IEEE Transactions of Fuzzy Systems, (2001)
[9] Jose Borges, Mark Levene, Data Mining of User Navigation Patterns, in Web Usage Analysis and User Profiling, published by Springer-Verlag as Lecture Notes in Computer Science, 1836: 92-111
[10] M. Deshpande, G. Karypis, Selective Markov Models for Predicting Web-Page Accesses, in ACM Transactions on Internet Technology, 4(2):163-184, (2004)
[11] I.Cadez, D.Heckerman, C.Meek, P. Smyth, S. White, Visualization of Navigation Patterns on a Web Site Using Model Based Clustering, in Proc. of ACM KDD2000 Conference, Boston MA (2000)
[12] A. Ypma, T. Heskes, Categorization of web pages and user clustering with mixtures of Hidden Markov Models, in Proc. of 4th WEBKDD Workshop, Canada(2002)
[13] "Study of Basics of Web Mining and Markov Models for Personalization" Varun Hooda and Mamta Kathuria, IJRIT International Journal of Research in Information Technology, Volume 2, Issue 2, February 2014,Pg: 159-164.

BIOGRAPHIES

Varun Hooda, received his M.Tech. (Computer Engineering- Networking) from YMCA University of Science and Technology, Faridabad in the year 2014. His research interests include Web Mining. E-mail: hooda.vaarun@gmail.com

Mamta Kathuria, is working as an Assistant Professor in the Department. of Computer Engineering, YMCA University of Science and Technology, Haryana, India. She holds a Masters Degree in Computer Engineering and is currently pursuing her PhD in Computer Engineering from YMCA University, Faridabad. Her research interests include Web Mining. E-mail: mamtakathuria7@rediffmail.com