Dimensionality Reduction in Markov Model for Web Personalization

Monika Ahlawat
Student of Computer Science and Engineering Department, Sat Priya Group of Institutions, Rohtak, Haryana, India.

Abstract – It wouldn't be wrong to say extracting information through web mining has dramatically increased due to personalized web pages. Traffic over the networking sites have significantly intensified due to amount of web pages on particular websites. Markov model is pretty sound system in such regard. This model does have its drawbacks as it has restraints over large dimensions of data requirement and its high dimensionality. Markov model has proficiently tackled the issue of high dimensionality in this paper. Web pages with similar navigational pattern have been grouped by examining user's access pattern and hence Dimensionality is condensed.

Index Terms –Web Mining, Markov model, Dimensionality, Personalized, Grouping.

1. INTRODUCTION

Web mining is a method through which information is derived out of WWW (World Wide Web). World Wide Web is considered to have captured widespread information, which incorporates every aspect for any desired user. Effectiveness of the data derived by any user through web mining is always paramount to the significance of the information retrieved. There has been different methods for obtaining important and appropriate data from World Wide Web through web mining which is further categorized in three dimensions:

- Web Structure Mining
- Web Content Mining
- Web Usage Mining

Web usage mining applications include web personalization through which web user requirements are dealt more effectively and information required is evaluated and modified according to them. In doing so it helps the interface in visiting the web site. The most prominent practices for web personalization is Markov model.

2. RELATED WORK

Data source of web mining and personalization process is the information existing in its website logs. Each and every surfing activity on a particular page is being recorded by web logs on the server hosting it. Whenever someone visits any page certain important information is captured by web log i.e. URI requested, IP address of the computer used by the user, times and activity requested by user, HTTP status code which was returned to the client and so on. W3C (4) has suggested a format under which web logs should be based on and that's called "extended" log format. User behavior and patterns in terms of activity during web mining have been analyzed and it provides an indication towards those patterns followed by users. Log data has been analyzed in different manners and few of the techniques which have been more prominent among researchers for data mining are association rules discovery [5], sequential pattern analysis [6, 7], clustering [8], probabilistic models [9, 10], or a combination of them [11,12].

3. PORPOSED MODELLING

Web personalization is attained through Markov model which is predominantly sourced around user's history, this model is also used to determine the user's forth coming behavior using and analyzing his past history. Under Markov model we are enabled to use large quantity of data from the web log and use that information for web mining and personalization.

3.1 Markov Model

Markov chain in the navigation of user is also considered as an application of Markov model (3). The model is tremendous in predicting users next movement based on its past navigational history. In relation to web page accesses it could be used for forecasting likelihood of user's next web page access based on history, but on contrary Markov model has its own limitations:

3.2 Limitations of Markov Models

- Large size input data requirement

Model is based purely on statistical processes hence the prediction is reliant on quantity of web log availability. We can establish that Markov models need of a very large sized input data shows inadequacy in the model.

- High Dimensionality Problem

Due to huge amount of web pages the transition matrix is created from transition graph is usually of big size. Dimensionality can be reduced by grouping of similar web pages.

3.3 The Transition Graph

Transition graph provides us the past access to web pages and also highlights the tendency about how frequently are those visits done by a user or group of users. Transition graph can be deemed as weighted graph, nodes demonstrating the web pages and weights of edges epitomize the frequency of page visit among those pages connected by a hyperlink.

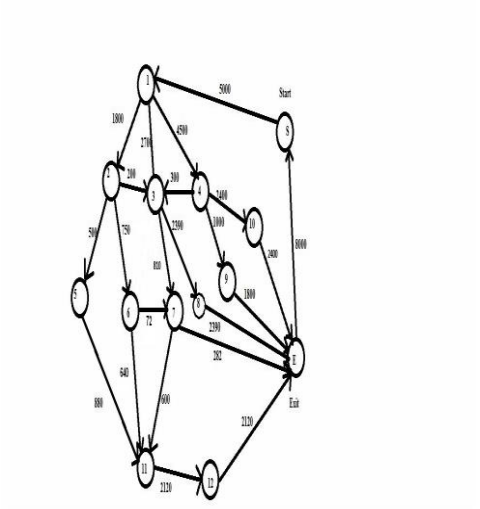

Fig.1 : The Transition Graph of Users Web page access

The transition graph in Fig. 1 provides precise understanding of user's web pages history in a span of time. In graph nodes are web pages, numbers indicated 1 through 12. Node 'S' illustrates start page and node 'E' illustrates end page of user's access. Weight on the edges reflects the number of visits in that span from one web page to another.

Transition graph has computed transition likelihood from page i to page j in only one step. Using the weights of graph links, we can create a transition likelihood matrix which incorporates one step transition probabilities in Markov model.

3.4 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

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

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 i to all other pages.

4. RESULTS AND DISCUSSIONS

Probability visiting has been calculated from one page to another by grabbing information from transition graph as an input and using number of visits among pages in the formula provided above. Hence probability matrix has formed.

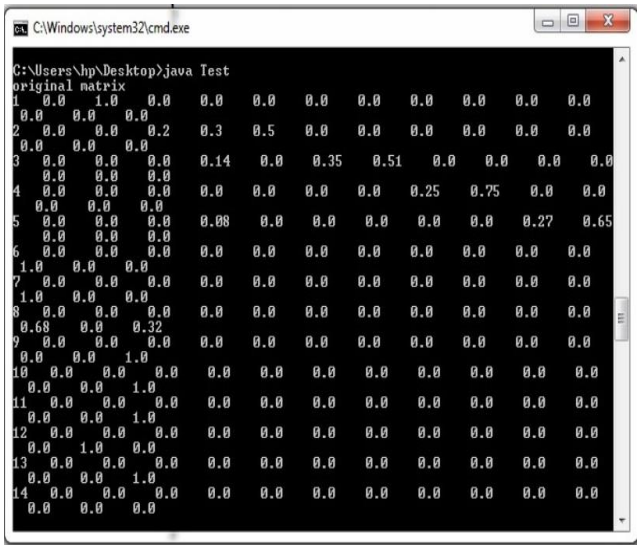

Fig.2: Original Matrix of Probabilities of Transition

Fig.2 is produced probability transition matrix illustrating probability visits 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.

It is important to lessen the dimensionality of Markov model to some degree because attained probability transition matrix is analyzed and found that matrix from large transition graphs would have very high dimensionality.

4.1 Dimensionality Reduction in Probability Transition Matrix

Grouping of similar web pages based on navigational pattern will significantly assist in reducing the size of probability matrix. Web pages with similar navigational pattern can be clubbed to make clusters. Transition matrix size can be reduced once the navigational pattern of a user analyzed. Web pages which are clubbed collectively would result in reduction in Markov model dimensionality which is hence achieved. Size of reduced matrix is lesser than that of original probability matrix.

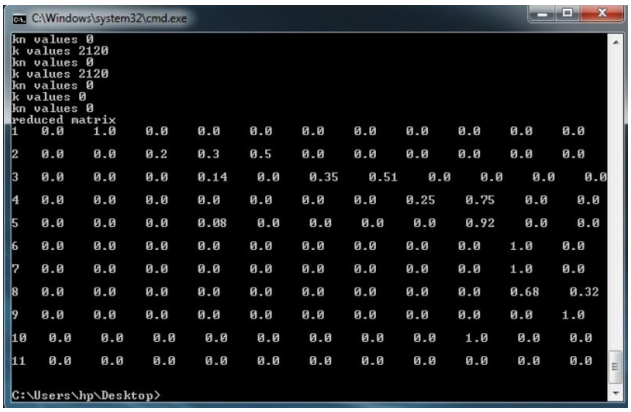

Fig.3 Reduced Transition Probability Matrix

Fig.3 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 demonstrated that size of reduced matrix has finally lessened the problem of dimensionality in Markov model to some extent. Size of matrix has decreased from 14 to 10, because dimensionality is of second order, there has been a significant level of reduction in space complexity of Matrix

5. CONCLUSION

In this paper, we have discussed the Web Personalization is essential application of Web Mining. Markov model's pros and cons are discussed. High dimensionality problem of Markov model is reduced to some extent by carefully studying the users' access behavior 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*
[6] *Web Environment*, in Proc. of the 16th Intl. Conference on Distributed Computing
[7] Systems (1996)
[8] [6] B. Berendt, *Using site semantics to analyze, visualize and support navigation*, in
[9] Data Mining and Knowledge Discovery Journal, 6: 37-59 (2002)
[10] A.G. Buchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, J.G. Hughes,
[11] *Navigation pattern discovery from Internet data*, in Proc. of the 1st WEBKDD