A New Web Usage Mining Approach for Next Page Access Prediction

A. Anitha
Member, IEEE,
UGC-Senior Research Fellow,
Centre for Information Technology and Engineering, Manonmaniam Sundaranar University,
Tirunelveli, Tamil Nadu – 627 012, INDIA

ABSTRACT
To engage users of a website at an early stage of surfing, a novel web access recommendation system is essential. In this paper, a new web usage mining approach is proposed to predict next page access. It is proposed to identify similar access patterns from web log using pair-wise nearest neighbor based clustering and then sequential pattern mining is done on these patterns to determine next page accesses. The tightness of clusters is improved by setting similarity threshold while forming clusters. In traditional recommendation models, clustering by non-sequential data decreases recommendation accuracy. In this paper it is proposed to integrate Markov model based sequential pattern mining with clustering. A variant of Markov model called dynamic support pruned all k^{th} order Markov model is proposed in order to reduce state space complexity. Mining the web access log of users of similar interest provides good recommendation accuracy. Hence, the proposed model provides accurate recommendations with reduced state space complexity.

Keywords: Similarity measure, pair-wise nearest neighbor, Markov model.

1. INTRODUCTION
The web is an important source of information retrieval now-a-days, and the users accessing the web are from different backgrounds. The usage information about users are recorded in web logs. Analyzing web log files to extract useful patterns is called web usage mining. Web usage mining approaches include clustering, association rule mining,sequential pattern mining etc.,To facilitate web page access by users, web recommendation model is needed. The web usage mining approaches can be applied to predict next page access [3].

The web recommendation models provide access friendliness for users while browsing a website. It also reduces network latency by pre-fetching the recommended pages. The prediction models play a vital role in e-commerce, to give advertisement at specific pages of commercial website. Web access prediction is useful in personalization to send personalized web content to specific type of users.

Clustering can identify similar access patterns. If mining is done on those patterns, recommendation accuracy will be improved rather than mining dissimilar access patterns. To form highly dense clusters, it is proposed to do pair-wise nearest neighbor base clustering approach using only k-neighbors. Clustering on non-sequential access data is responsible for degradation of recommendation accuracy in traditional models. Hence, in this work a sequential mining technique called Markov model is suggested in combination with pattern discovery.

The pair-wise nearest neighbor approach suffers from its slowness due to merging every pair of clusters, and updating distance values after every merge. Therefore, to reduce the number of distance updations, instead of considering all neighbors of a cluster in every step, only first k neighbors are considered. the candidate that participate in merging process must fall in the k-neighbor list. Hence number of computations is reduced from o (N-1) to O (K) at every step. More over, instead of using distance measure similarity measure involving simple operations is suggested.

The major benefit of hierarchical clustering approach (PNN) is that every object must be the candidate of only one cluster. Hence by applying this approach for pattern discovery, the click-sequences that participate in prediction module are exact and have good resemblance with one another. These are the factors that lead to improvement in recommendation accuracy.

2. BACKGROUND OF STUDY
Different combinations of mining techniques were already suggested for web access recommendation. Pawlak[4] introduced web access prediction model by integrating roughest clustering with Markov model. It has major drawback that lack of prediction accuracy due to approximation while forming clusters. The possibility of an object for belonging to a cluster can reduce the cluster tightness, which in turn affects prediction accuracy. The sequential mining suggested in that work is all k-th order Markov model.

For high order Markov models, if coverage is very less, it affects accuracy of prediction. Hence in this paper it is proposed to do support pruning. By which, the states that have less support or low coverage are eliminated while making recommendations. Another model[5] based on Markov process for web access prediction has drawback of high complexity due to consideration of all access sequences through out the prediction process. Sometimes, in this approach noisy or irrelevant access sequences participate in web page access prediction affects the accuracy. Hence it is suggested to do pattern discovery before sequential mining. Some works based on association rule were made, where prediction accuracy is affected by contradictory predictions

due to the generation of too many rules and participation of huge number of access sequences in mining process. A combined approach like integrating Markov model and association rule [6] is also affected by mining all types of access sequences i.e) without clustering. In agglomerative clustering using k neighbors, at later stage of clustering process, distant neighbors may also fall into k neighbors and they decrease the cluster tightness. Hence threshold is set to eliminate such objects.

Hence, it is proposed to find out highly homogeneous access patterns by pair wise nearest neighbor based clustering. The resultant patterns are highly relevant , and the size dataset that are utilized for sequential mining process is highly reduced.

3. BLOCK DIAGRAM

3.1 Data preparation

The web log data is used for mining process is integrated, cleaned and relevant attributes like IP and web page accesses are selected. The clickstream is a sequence of mouse click made by every user. The transactions are generated by eliminating noisy, and very short or very long access sequences.

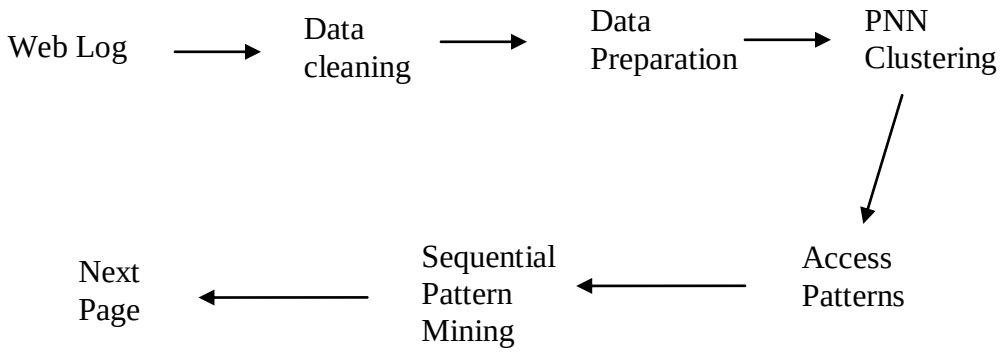

Figure 1. Block diagram of web recommendation system

3.2 PNN based clustering

Clustering is the process of grouping the objects in such a way that, intra cluster similarity is high and inter cluster similarity is low. The pair wise nearest neighbor approach is a bottom-up hierarchical clustering technique, by which every object belongs to individual clusters initially, pair wise merging of objects is done at every step based on their similarity. Finally, resulting in a single cluster. The distance calculations [2] are replaced by similarity measure.

Similarity between two transactions are given by the ratio of, total number of unique pages referenced by them to the number of common references. Then a table representing similarity value between every pair of transaction is created. For every transaction, its first k nearest access sequences are identified. Among the whole set of k-neighbors, the pair of sequences having high similarity is identified and merged. For this new cluster, the new k –neighbors are identified from 2k neighbors and its back neighbors are also updated[1]. The merging is continued until no more merging is possible. In this process , only the pair of access sequence having similarity value greater than a pre-defined threshold is selected for merging process. By this approach , the distant objects that are irrelevant to mining process are eliminated resulting in homogeneous access patterns.

3.3 Sequential Pattern mining

As clustering involves non-sequential access patterns,in order to improve the accuracy of prediction process, sequential mining using markov model is done on next stage. Let x1,x2…….xk be the set of pages in a web site, the probability of accessing next page is defined by markov model as,

The page to accessed next is,

$$X_{k+1} = \text{argmax}_{x \in IP} \{P (X_{k+1} = x \mid x_k, x_{k-1}, \ldots\ldots\ldots x_1)\} \quad (1)$$

Where k is the order of markov model or number previous page accesses considered .

By the above equation, among the set of all possible next pages obtained from training set ,the next page will be the one having highest possibility to be accessed next[6].

The traditional Markov models have serious limitations that, low order Markov models have good coverage but they lack accuracy due to poor history .And high order Markov models suffers from high state space complexity ,as they use long browsing history, but high order Markov models provides good prediction accuracy. To combine the advantages of all of those approaches, a dynamic support pruned kth order Markov model is suggested. In the proposed approach, the high state space complexity due to summing up of low order states of high order

Markov model[8] is eliminated. Since in the proposed dynamic support pruned all k-th order Markov model, when k-th order states does not have enough support then only $(k-1)^{th}$ order states are generated .Hence good prediction accuracy of high order Markov model is achieved with reduced state space complexity.

4 ALGORITHMS

4.1 Steps

1. Collect the access log information from web servers
2. Retrieve only IP address and URL details from access log by removing noise and filtering irrelevant details
3. Form click stream transactions and place each transaction in individual cluster
4. Perform pattern discovery by pair-wise nearest neighbor method on k neighborhood as follows:
    - (i) Find out similarity values between transactions
    - (ii) Identify first k neighbors having similarity greater than threshold, for every transaction and remove other neighbors
    - (iii) Cluster the pair with highest similarity
    - (iv) Update similarity for objects in the neighborhood of merged pair
    - (v) Find out new set of k neighbors from 2k neighbors of merged pair
    - (vi) Update the neighbors in the back list of merged pair
    - (vii) Repeat steps (iii) to (iv) until no more merging is possible
5. Get new test session
    - (i) Find out its cluster, the one having access sequences similar to test session
    - (ii) Start with highest possible value of k.
    - (iii) Apply Markov model and find out the k-th order states for the test session from its cluster
    - (iv) If the support is very less, calculate next lower order states for the test session from its cluster
    - (v) Repeat step (iv) until states are generated with enough support
6. Display the page with highest probability as the recommended page

5. RESULTS

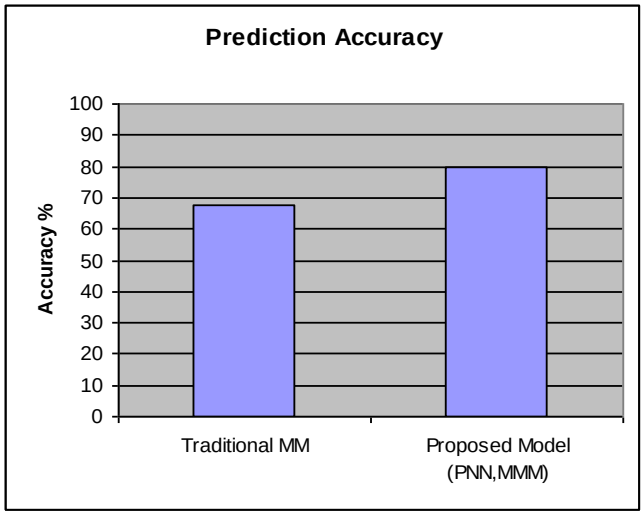

Fig 2. Prediction accuracy by PNN clustering with support pruned all k-th order markov model

The proposed model is applied on a log file of www.annauniv.edu where all graphics are eliminated, noise is removed and transactions are constituted. Sample training set chosen. Then, different set of clusters are formed. The PNN algorithm resulted in 5 clusters. The algorithm is tested in such a way that, by hiding the last page in every test session, a prediction is made. For the test session, its corresponding cluster is chosen, the dynamic support pruned all kth order markov model is applied only on that cluster. Fig 2. Shows that the percentage of accuracy by proposed model is higher that that of the traditional markov model. Out of the 40 test sessions, accurate predictions were made for 32 sessions by the proposed model, whereas without clustering, a simple traditional all k-th order markov model made only 27 correct predictions.

The accuracy of prediction is given by

$$Acc = Te \cap Tr / Te \qquad (2)$$

Where Te - is the number of test cases

Tr - is the number of training cases

6. CONCLUSION

The proposed method resulted in good prediction accuracy with less state space complexity. The drawback of this work is, loosely connected access sequences are not considered for mining process. Hence, it is suggested to extend this work by considering non-contiguous access sequences also.

REFERENCES

[1] Pasi Franti,Olli Virmajoki, and Ville Hautamaki "Fast Agglomerative Clustering Using a k Nearest Neighbor graph",IEEE transaction on pattern analysis and machine intelligence.Vol 28,No 11. November 2006, pp 1875-1880

[2] Pasi Franti, Timo Kaukoranta,Day-Fann Shen and Kuo-Shu Chang "Fast and Memory Efficient Implementation of exact PNN",IEEE Transaction on image processing,Vol 9,No 5,May 2000.pp 773-777

[3] Mathias G´ ery, Hatem Haddad," Evaluation of Web Usage Mining approaches for user's next request prediction" *WIDM* '03 Boston, USA,ACM

[4] Siripom chimphlee,Naomie Salim,Mohd Salihin Bin Ngadiman, Witcha,Surat ,"Rough Sets Clustering and Markov Model for Web Access Prediction" ,Proceedings of post graduate annual seminar 2006, pp. 470-474

[5]Devanshu Dhyani, Sourav S Bhowmick, Wee-Keong Ng, "Modelling and predicting web page accesses using Markov Processes", IEEE,Computer Society, 2003,1529-4188

[6]Faten Khalil, Jiuyong Li,Hua Wang, "Integerating Recommendation Models for Improved Web Page Prediction Accuracy", Australian Computer Society, 2008,Conferences in Research and Practice in Information Technology, Vol 74.

[7]Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa,"Effective Personalizaion based on association rule discovery from Web Usage Data", ACM workshop on Web Information and Data management, Nov 2001.

[8] Mukund Deshpande and George Karypis, "Selective markov model for predicting web-page accesses",Army High performance Computing Research Center, pp.1-15

[9] Faten Khalil, Jiuyong Li,Hua Wang, " Integrating Markov model with clustering for predicting web page accesses", Australian Conference, Mar 2007 ,pp 1-26