International Journal of Computer Applications Technology and Research
Volume 3– Issue 12, 827 - 830, 2014, ISSN:- 2319–8656

Prediction Model Using Web Usage Mining Techniques

Priyanka Bhart
U.I.E.T Kurukshetra University
Kurukshetra, India

Abstract: Popularity of WWW increasing day by day which results in the increase of web based services , due to which web is now largest data repository. In order to handle this incremental nature of data various prediction techniques are used. If the prefetched pages are not visited by user in their subsequent access there will be wastage network bandwidth as it is in limited amount. So there is critical requirement of accurate prediction method. As the data present on web is heterogeneous in nature and incremental in nature, during the pre-processing step hierarchical clustering technique is used. Then using Markov model category and page prediction is done and lastly page filtering is done using keywords.

Keywords: formatting Hierarchal clustering; markov model; page prediction; category prediction ;

1. INTRODUCTION

With the continued growth and proliferation of E-commerce there is need to predict users behavior. These predictions helps in implementing personalization, building proper websites, improving marketing strategy promotion, getting marketing information, forecasting market trends, and increase the competitive strength of enterprises etc.[1].

Web prediction is one of the classification problem where a set of web pages a user may visit are predicted on the basis of previously visited page which are stored in the form of web log files. Such kind of knowledge of users' navigation history within a slot of time is referred to as a session. This data is extracted from the log files of the web server which contains the sequence of web pages that a user visits along with visit date and time. This data is fed as the training data.

All the user's browsing behavior is recorded in the web log file with user's name, IP address, date, and request time etc.

S.no	Ip address	Req.	Timestamp	Protocol	Total bytes

Table 1.Common web log

Hierarchical clustering is a classification technique. It is an agglomerative(top down) clustering method , as its name suggests, the idea of this method is to build hierarchy of clusters, showing relations between the individual members and merging clusters of data based on similarity.

In order to analyze user web navigation data Markov model is used. It is used in category and page prediction. Only the set of pages which belong to the category which is predicted in first phase are used in second phase of page prediction. Here each Web page represent a state and ever pair of pages viewed in sequence represent a state transition in this model. The transition probability is calculated by the ratio of number of a particular transition is visited to the number of times the first state in the pair was visited.

This paper introduces an efficient four stage prediction model in order to analyze Web user navigation behavior .This model is used in identification of navigation patterns of users and to anticipate next choice of link of a user. It is expected that that this prediction model will reduce the operation scope and increase the accuracy precision.

2. RELATED WORK

Agrawal R and Srikant R [1] proposed a website access prediction method based on past access behavior of user by constructing first-order and second-order Markov model of website access and compare it association rules technique. Here by using session identification technique sequence of user requests are collected, which distinguishes the requests for the same web page in different browses. Trilok Nath Pandey [2] proposed a Integrating Markov Model with Clustering approach for user future request prediction. Here improvement of Markov model accuracy is done by grouping web sessions into clusters. The web pages in the user sessions are first allocated into categories according to the web services that are functionally meaning full. And lastly k-means clustering algorithm is implemented using the most appropriate number of clusters and distance measures. Lee and Fu proposed a two level prediction model in 2008[4]. Here prediction scope is reduced as it works in two levels. This model is designed by combining Markov Model and Bayesian theorem. In level one using Markov Model category prediction is done and page prediction is done by Bayesian theorem. Chu-Hui Lee [3] used the hierarchical agglomerative clustering to cluster user browsing behaviors due to the heterogeneity of user browsing features. The prediction results by two levels of prediction model framework work well in general cases. However, two levels of prediction model suffer from the heterogeneity user's behavior. So they have proposed a prediction model which decreases the prediction scope using two levels of framework. This prediction model is designed by combining Markov model and Bayesian theorem.

Sujatha [5] proposed the prediction of user navigation patterns using clustering and classification (PUCC) from web log data. In the first phase it separates the potential users in web log data, and in second t-stage clustering process is used to group the potential users with similar interest and lastly the results of classification and clustering is used to predict the users future requests. Sonal vishwakarma [6] analyzed all order Markov model with webpage keywords as a feature to give more accurate results in Web prediction.

3. OVERVIEW OF THE PROPOSED PREDICTION FRAMEWORK

The prediction model is designed by combining clustering and Markov model technique. During preprocessing step hierarchical clustering is done to group user's browsing behaviors and acquires many different clusters. The information of relevant to any cluster can be seen as cluster view that means every cluster has its own relevant matrix irrespective of having these matrixes for every user, so here global view is replaced by cluster view. After preprocessing category prediction is done by using Markov model. Here in this phase it is to predict category at time t which depends upon users category time at time t-1 and t-2. In the same way page prediction is done to predict the most possible web pages at a time t according to users state at a time t-1. Now the set of predicted pages are fed for keyword based filtering. Finally after this phase predicted results are released.

Firstly training data is fed for clustering where k number of cluster view will be obtained which include k similarity matrices S, k first-order transition matrices P and k second-order transition matrices between categories. Therefore, we get K relevant matrices R to represent K cluster views [4]. In step two, these matrices will be released out for creating index table which will be used for view selection based on user's browsing behavior at that time. In step three, after view selection testing data is fed into the prediction model and prediction results will be released as output.

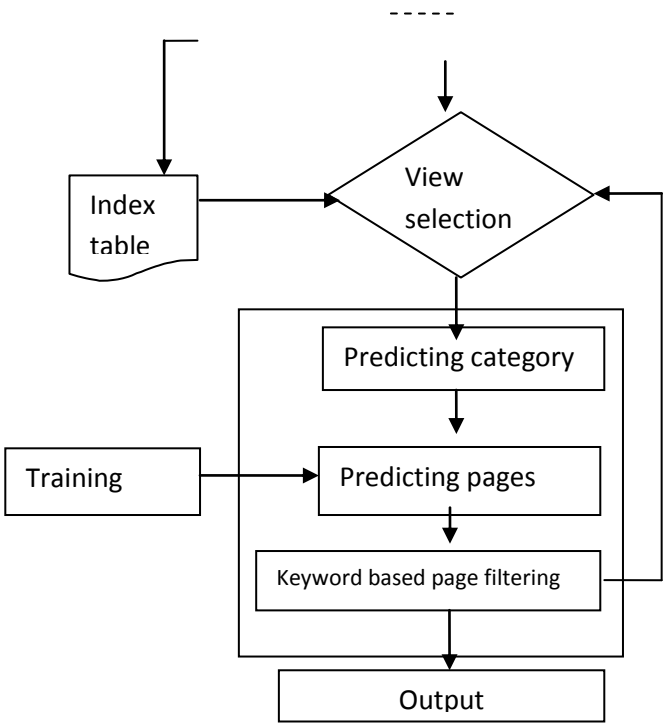
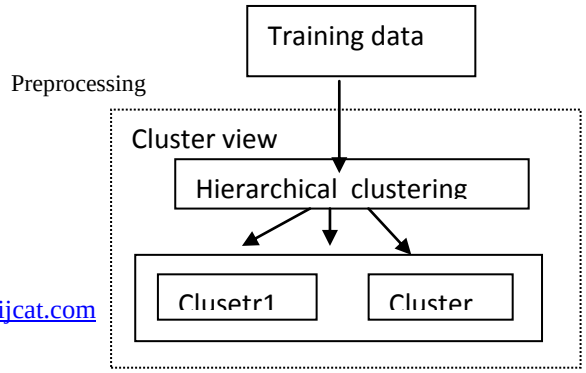


Figure 1. Proposed prediction framework

3.1 HIERARCHICAL CLUSTERING ALGORITHM

It is an agglomerative clustering method. The idea of this method is to build a hierarchy of clusters, showing relations between members and merging clusters of data based on similarity. For the clustering algorithm to work there is need to have some means by which similarity to be judged. This is generally called a distance measurement. Two commonly used metrics for measuring correlation (similarity/distance) are Euclidean and the Pearson correlations. The type of correlation metric used depends largely on what it is to be measured. Here we have used Euclidean correlation.

In the first step of clustering, the algorithm will look for the two most similar data points and merge them to create a new "pseudo-data point", which represents the average of the two data points. Each iterative step takes the next two closest data points and merges them. This process is generally continued until there is one large cluster covering all original data points. This clustering technique will results in a "tree", showing the relationship of all the original points. Here every user seems to be a cluster and grouped by most similar browsing feature into the cluster [12].



3.2 MARKOV MODEL

Markov is a probability based model which is represented by three parameters<A,S,T> where A is a set of all possible actions performed by any user; S is the set all possible states for which the Markov model is built; and T is a |A|× |S| Transition probability matrix, where represent the probability of performing the action j when the process is in state i. Markov model predicts the user's next action by looking at previously performed action by the user.

Here assume that D is the given database, which consists of user's usage records. It means users sessions are recorded and $D=\{session_1, session_2, \ldots\ldots\ldots, session_p\}$. each user session is a set of web pages recorded in time orderd sequential pattern and $session_p=\{page_1, page_2, \ldots\ldots, page_n\}$, where $page_i$ represents user's visiting page at time j. If a website has K categories, then the user session can be represented as $session_c=\{c_1, c_2, \ldots.., c_k\}$.

3.3 TRANSITION MATRIX

P is transition probability matrix represented as in equation(1) where P_{ij} represent the transition probability between any two pages/category ie. from P_i to P_j. It is calculated by the ratio of number of transition between category/page i and category/page j to the total number of transition between category/page i and every category/page k.

$$P = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1k} \\ P_{21} & P_{22} & \cdots & P_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ P_{k1} & P_{k2} & \cdots & P_{kk} \end{bmatrix} \end{array} \quad (1)$$

Web sessions

WS1={P3,P2,P1}

WS2={P3,P5,P2,P1,P4}

WS3={P4,P5,P2,P1,P5,P4}

WS4={P3,P4,P5,P2,P1}

WS5={P1,P4,P2,P5,P4}

1st Order	P1	P2	P3	P4	P5
S1={P1}	0	0	0	2	1
S2={P2}	4	0	0	0	1
S3={P3}	0	1	0	1	1
S4={P4}	0	1	0	0	2
S5={P5}	0	3	0	2	0

2nd Order	P1	P2	P3	P4	P5
{P1,P4}	0	1	0	0	0
{P1,P5}	0	0	0	1	0
{P2,P1}	0	0	0	1	1
{P2,P5}	0	0	0	1	0
{P3,P2}	1	0	0	0	0

2nd Order	P1	P2	P3	P4	P5
{P2,P5}	0	1	0	0	0
{P2,P4}	0	0	0	0	1
{P4,P5}	0	2	0	0	0
{P5,P2}	3	0	0	0	0
{P3,P4}	0	0	0	0	1

Figure 2. Sample web sessions with corresponding 1_{st} and 2^{nd} order transition probability matrices [7].

3.4 SIMILARITY MATRIX

Similarity between any two user user i and user j, can be calculated using Euclidean distance given in equation (3) .

$$sim(user_i, user_j) = (session^i, session^j) \quad \textbf{(2)}$$

Euclidean distance

$$D(user_i, user_j) = \sqrt{\sum_{i=1}^{k} (P_{i,i} - P_{j,l})^2} \quad (3)$$

Euclidean distance is further normalized by (4) equation, by this k×k similarity matrix as given in equation (5) will be obtained.

$$ND(user_i, user_j) = 1 - \sqrt{\frac{\sum_{i=1}^{k} (P_{i,i} - P_{j,l})^2}{k}} \quad (4)$$

$$S = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1k} \\ S_{21} & S_{22} & \cdots & S_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ S_{k1} & S_{k2} & \cdots & S_{kk} \end{bmatrix} \end{array} \quad (5)$$

3.5 RELEVANCE MATRIX

Last matrix to create is relevance matrix represented as in equation (6),which is equal to the product of transition and similarity matrix. Here relevance is an important factor of prediction between any two category and pages. It concludes the behavior between pages and categories. It is represented as follows:

$$R_n = \begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_k \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_k \end{array} \\ \begin{bmatrix} {R_{11}}^n & {R_{12}}^n & \cdots & {R_{1k}}^n \\ {R_{21}}^n & {R_{21}}^n & \cdots & {R_{2k}}^n \\ \vdots & \vdots & \ddots & \vdots \\ {R_{k1}}^n & {R_{k2}}^n & \cdots & {R_{kk}}^n \end{bmatrix} \end{array} \quad (6)$$

Where

$${R_{ij}}^n = {P_{ij}}^n \times S_{ij} \quad (7)$$

4. CONCLUSION

As there is large amount of data on web pages on many websites, So it is better to place them according to their category. In this paper users browsing behavior is firstly preprocessed using hierarchical clustering then prediction is done in three phases. In first phase category prediction is done using Markov model then in second phase page prediction is done. And lastly keyword based filtering is done which gives more accurate results.

5. REFERENCES

.

[1] Agrawal R, Imielinski T and Swami A "Mining Association Rules between Sets of Items in Large Databases", ACM SIGMOD Conference on Management of Data, pp.207-216.

[2] 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 International Journal of Computer Science Issues, Vol. 9, Issue 6, No 1, November 2012.

[3] Chu-Hui Lee, Yu-Hsiang Fu "Web Usage Mining based on Clustering of Browsing Features**"** Eighth International Conference on Intelligent Systems Design and Applications,IEEE, 2008. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[4] Chu-Hui Lee , Yu-lung Lo, Yu-Hsiang Fu, "A Novel Prediction Model based on Hierarchical Characteristic of Web Site", Expert Systems with Applications 38 , 2011.

[5] V. Sujatha, Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", Procedia Engineering 30 ,2012.

[6] Sonal Vishwakarma, Shrikant Lade, Manish Kumar Suman and Deepak Patel "Web User Prediction by: Integrating Markov with Different Features",vol2 IJERST , 2013.

[7] Deshpande M and Karypis G (2004), "Selective Markov Models for Predicting Web-Page Accesses", ACM Transactions on Internet Technology (TIOIT),Vol.4,No.2,pp.163-184.

[8] UCI KDD archive, http://kdd.ics.uci.edu/

[9] V.V.R. Maheswara Rao, Dr. V. Valli Kumari" An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining" 2010 International Conference on Advances in Recent Technologies in Communication and Computing IEEE.

[10] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction", International Journal of Computer Applications, Volume8–No.11,October2010.

[11] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat, "WebPUM: A Web-Based Recommendation System to Predict User Future Movements" Expert Systems with Applications 37 , 2010.

[12] www.microarrays.ca/services/hierarchical_clustering.pdf