An Efficient Algorithm Research to Web User Access Prediction

Shaorong Feng

School of Information Science and Technology, Xiamen University, Xiamen, China

shaorong@xmu.edu.cn

Keywords: Access prediction; Feedback mechanism; Web log mining; Markov prediction; Association rule

Abstract. In order to enhance accuracy of the algorithms on user access prediction. The prediction algorithm based on Markov chain and association rule(PAMA), and Markov prediction model with feedback(MPMF) were proposed. The PAMA integrates the advantage of Markov chain and association rule well. It corrects the Markov prediction result on forward and reverse perspective, and gets the last prediction page. The MPMF adjusts the prediction algorithm dynamically according to user feedback mechanism and history prediction information, and then gets the prediction page at last. Theoretical analysis proves these two prediction methods with linear time complexity. Experiments result shows that the accuracy of PAMA and MPMF is good, so the prediction efficiency is also meeting the requirement.

Introduction

The research of user access prediction is mainly focused on the following 3 aspects:

(1) Mathias Gery[1], etc. put forward to use association rules to predict the next access page of users[2]; Zhili Zhang[3], etc. put forward to combine rough set theory with association rules to conduct the prediction.

(2) Zuckerman[4] etc. put forward user access prediction algorithm based on Markov model. Xing Yongkang, etc. put forward multi Markov chain prediction algorithm[5-6].

(3) Sule Gunduz[7], etc. put forward prediction algorithm based on click stream data.

In order to solve the existing problems in the process of user access prediction, this paper puts forward user access prediction algorithm and Markov prediction algorithm with feedback based on Markov chain and association rules. According to the theoretical analysis and experimental results, the prediction accuracy is significantly improved.

Prediction Algorithm Based on Markov Chain and Association Rules

k order Markov chain prediction model ignores the earlier history access knowledge. Under the assumption that the next access page only has a relationship with the latest k pages, it simplifies prediction model and reduces the computing time, but the prediction accuracy also decreased. Aiming at this problem, this paper proposes a prediction algorithm based on Markov chain and association rules. Firstly, use Markov chain to predict the page set M which the model may visit after returning the next step and the Markov prediction probability correspond to these pages. Then, use two association rules to revise prediction results from two angles of forward and reverse. Assume that the current user access page sequence is v=$\{p_1, p_2, \ldots, p_s\}$, the result set predicted by Markov m=$\{r_1, r_2, \ldots, r_t\}\{\}$ and the corresponding Markov prediction probability are $mp=\{mp(r_1), mp(r_2), \ldots, mp(r_t)\}$. The reverse and forward revise processes are as follows:

Reverse revise process: For arbitrary $r_i \in$ m, separately calculate the credibility of all pages and r_i of v. If all the credibility is less than the threshold min_rule, no matter the value of $mp(r_i)$, delete page r_i from m. Therefore, the reverse revise process has veto power. Since all the credibility is less than the threshold min_rule, we consider that there is no hyperlink relation in all pages of page r_i and v, so users have little probability to access the next page.

Forward revise process: For arbitrary $r_i \in m$, separately calculate the credibility of all pages and r_i of v. If page $p_j \in v$ exists, and the credibility conf($p_j \to r_i$) of p_j and r_i is more than the threshold max_rule, the prediction probability of r_i is:

$$preditProb(r_i) = \lambda_1.mp(r_i) + \lambda_2 \sum_j \omega_j.conf(p_j \to r_i) \qquad (1)$$

In which, $\omega_j = \frac{j}{|v|}$, $\lambda_1+\lambda_2=1$. λ_1 is the predictive value of Markov. λ_2 is the predictive value of association rules. Their values will be discussed in details in the experimental analysis. ω_j is the weight of rule conf($p_j \to r_i$). Because in the access of sequence v, the closer p_j is to the current time, the greater the reference value of this rule, while the further, the smaller.

Algorithm 1 the Prediction Algorithm Based on Markov Chain and Association Rules. **Input:** Current access sequence v= (p_1, p_2, …, p_s), threshold min_rule and max_rule.

Output: Prediction page

```
1 Use second order Markov prediction algorithm to calculate the page set m which user is likely to access and the corresponding Markov prediction probability mp;
2 foreach ( $r_i \in m$ )
3 { $predictProb(r_i)=mp(r_i)$;          // Initial prediction probability
4 foreach ( $p_j \in v$ )
5 { $conf(p_j \to r_i)= R_{p_j r_i}$; }   // Find the credibility of the corresponding rules from the correlation matrix R
6 if ( foreach $p_j \in v$ $conf(p_j \to r_i)$<min_rule)
7 { remove $r_i$ from m; }        // Delete $r_i$ from m
8 foreach ($p_j \in v$)
9 { if ($conf(p_j \to r_i)$>max_rule)
10 { $predictProb(r_i) += \lambda \cdot \omega_j \cdot conf(p_j \to r_i)$;}      // Calculate prediction probability
11 } }
12 Return the page with the maximum value of the predictProb.
```

Theory 1 The time complexity of Algorithm 1 is $O(n)$, in which n is the number of web pages. **Prove** The time complexity of the first line of the algorithm is $O(n)$. The cycle execution of the second line is $|m|$ times and the cycle of line 4,6,8 are all |v| times. m is a subset of the page web site, its size is controlled by Markov prediction algorithm and are generally much less than n. |v| is users access length and is the random variable following the Poisson distribution, so its mathematical expectation is λ. So the time complexity of the algorithm is $O(n+\lambda|m|)=O(n)$. Over!

Markov Prediction Model with Feedback

The classic model of user access prediction has a common shortcoming that the results of the prediction cannot be decided and the prediction algorithm cannot be dynamically adjusted without considering the user feedback. This paper introduces the user feedback[8] mechanism and the preservation of a certain amount of historical records, according to the user feedback, determine whether the prediction is correct and thus establish a historical tree including prediction accuracy. In the process of prediction, the historical prediction results with high accuracy can be used directly as the current prediction; or the historical prediction accuracy can be taken as the calculation parameters of the current prediction. Markov prediction algorithm shall be dynamically adjusted, so as to improve the efficiency and accuracy of prediction.

History Prediction Tree. HPT (History Prediction Tree) is used to preserve the historical prediction information which can be used to assess the accuracy of the current prediction, and thus to dynamically adjust the prediction algorithm to improve the accuracy of prediction. History Prediction Tree is gradually generated in the prediction process. It has only one root node. From the leaf node to the root node, it is a access path. Each access path has one or more corresponding historical prediction records. HPT also contains a page table which records all the pages in HPT and the entrance address of these pages in HPT.

Definition 1 History Record is a four tuple (x, p, y, t). x represents the times of correct prediction, p represents prediction pages, y represents the times of wrong prediction and t represents the prediction time.

Definition 2 HPTNode is a 5 tuple (p, parent, childList, next, hrList), in which p represents page, parent represents the parent node of the node, childList represents the set of the child nodes of the node, next represents

Definition 3 The table entry for the page table is a two tuple (page, HPTNode). page represents the page which shows in HPT has uniqueness in page table; HPTNode represents the address of the first node of the page which shows in HPT.

Construction Algorithm of HPT. Input: The current access sequence of users v=(p_1, p_2, …, p_s) and prediction page p

Output: HPT

```
cnode = root;    // The current node points to the root node
foreach ( page pg in v )    // if pg is in v
{ if( pg in cnode. childList ) // if pg is the child node of the current node
{ cnode = node which node.p = pg in cnode. childList;}
else
{ new Node. p = pg;      // establish a new node
new Node. parent = cnode;
cnode.childList  ∪= newNode;
cnode = new Node;
Establish link between newNode and the node contains page pg:}
} Insert the prediction results ( 0, p, 0, currentTime ) into cnode.hrList;
```

History prediction tree increases with the number of prediction and needs to be reduced. It is controlled within an appropriate range. The analysis shows that the branch with earlier prediction has poor timeliness; it is hard to judge the effect of the branch with the same number of the times of correct and wrong prediction; the branch with fewer times of historical prediction (times of correct + times of wrong) is lack of statistical significance; therefore these branches can be cut.

Markov Prediction Model with Feedback. The model firstly queries the historical prediction information of the current session in HPT. If the prediction accuracy existing in HPT is greater than the prediction page of the specified threshold (goodT), the page can be directly considered as the prediction result of the current session; otherwise, find out the historical prediction results and prediction accuracy from HPT and revise the prediction results of Markov from forward and reverse. The user feedback is the foundation of HPT. The feedback process is to determine whether the current prediction is correct according to the next access pages of users.

Markov Prediction Algorithm with Feedback. If the prediction page is not found in the query process, then return to the historical prediction records and revise Markov prediction results from the forward and reverse. Assume that the current access page sequence of users is v={p1, p2, …, ps}, Markov prediction result set is m={r1, r2, …, rt} and the corresponding Markov prediction probability is mp={mp(r1), mp(r2), …, mp(rt)}. The following are the introduction of forward revise and reverse revise.

Reverse revise process: For arbitrary $r_i \in$ m, if the historical prediction accuracy rate of r_i is less than threshold min_p, delete r_i from m.

Forward revise process: For arbitrary $r_i \in$ m, if there is r_i in the historical prediction records, then calculate its historical prediction accuracy rate (which is marked as hpp(r_i)). The final prediction probability is:

$$predictProb(r_i)=\mathrm{mp}(r_i)+\lambda \cdot \mathrm{mp}(r_i) \tag{2}$$

In which, λ is the weight value of historical prediction, $\lambda \in [0,1]$. If there is no r_i in the historical prediction records, current Markov prediction probability shall be considered as the historical prediction probability which is :

$$\mathit{predict\ Prob}(r_i)=\mathrm{mp}(r_i)+\lambda\cdot\mathrm{mp}(r_i) \tag{3}$$

Algorithm 2 Markov Prediction Algorithm with Feedback. Input: The current access sequence of users v=(p_1, p_2, ……, p_s), the historical prediction records HRS and threshold min_p

Output: Prediction pages

```
(1) Use second order Markov prediction algorithm to calculate the page set m which user is likely to access and the corresponding Markov prediction probability mp;
(2) foreach (HR hr in HRS)
(3) { if ( hr.y/(hr.x+hr.y)<min_p && hr.p in m)
(4) { remove hr.p from m; remove hr from HRS;        }
(5) }
(6) foreach (page p in m)
(7) { if (p in HRS.p) // if p in HRS.p
(8) predictProb(p)=mp(p)+$\lambda$·hpp(p)
(9) else predictProb($r_i$)=mp($r_i$)+$\lambda$·mp($r_i$)
(10) }
(11) Return the page with the maximum predictProb value
```

Theory 2 The time complexity of Algorithm 2 is $O(n)$, where n is the total number of Web site pages.

Prove The time complexity of the first line of the algorithm is O(n). The cycle execution of the second line is |HRS| times and the cycle execution of the eighth line is $|m|$*|HRS| times. m is the page set which users are likely to access calculated by Markov prediction algorithm. m is less than n. HRS is a historical prediction record set corresponding to the current session which is less than n. So $|m|$*|HRS|<n and the time complexity of the algorithm is $O(n)$. Over!

The time complexity of the online part of the Markov prediction model with feedback is $O(n)+O(\lambda^2|\mathrm{pageSet}|)$, in which |pageSet| is the number at leaf nodes with the page of $\mathrm{p_s}$ in HPT and λ is the mathematical expectation of the length of the user access path. Therefore, if the scale of HPT is controlled in a suitable arrange and the branches with small predictive guiding significance in HPT are cut off, the prediction model would have good prediction efficiency.

User feedback is the foundation to establish HPT. The feedback process is to determine whether the current prediction is correct, based on the next page which users will access. If it is correct, add 1 to the correct field (y); if it is wrong, add 1 to the wrong field (y) in the corresponding historical prediction records.

Result Analysis

Prediction Analysis Based on Markov Chain and Association Rules. The experimental data is from the access log for one week in February of 1998 of Microsoft server users. Fig. 1 is the comparison chart of maximum 1- prediction accuracy rate of the two order Markov chain prediction algorithm (MPA), and Markov chain and association rules prediction algorithm (PAMA), Fig. 2 is the comparison chart of the maximum x- prediction accuracy rate.

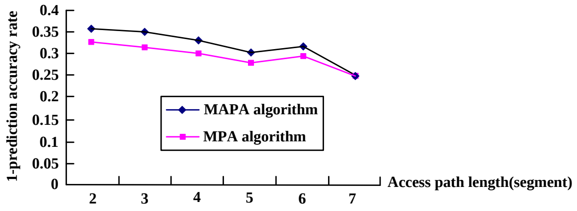

Figure 1. Max 1-Prediction accuracy comparison of the MPA and PAMA

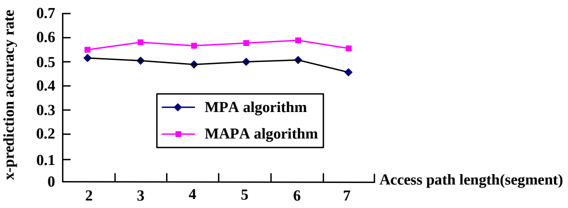

Figure 2. Max x - Prediction accuracy comparison of the MPA and PAMA

In the process of calculating x- prediction accuracy rate, when λ_1 and λ_2 are respectively 0.5 (other parameters unchanged), x- prediction accuracy rate of algorithm PAMA is the best. Fig. 2 is the largest x- prediction accuracy rate comparison chart of the MPA and PAMA algorithm

In algorithm, no matter 1- prediction accuracy rate or x- prediction accuracy rate, when α_1 is 0.5 (means the weight values of the first order transition probability and the 2 order transition probability are the same), the rate is high, that is the current access page and the last access page have the equivalent role in Markov prediction process. However, in PAMA algorithm, whenλ_2 is 0.1 (λ_1 is 0.9), 1- prediction accuracy rate is the best, while when λ_2 is 0.5 (λ_1 is 0.5), x- prediction accuracy rate is the best. If λ_2 is 0.1, the effect of association rule modification on the results is small and when λ_2 is 0.5, the effect of association rule modification on the results is big. According to experience, in one step transfer process, between the current access page and the next access page, Markov is stronger; however, between the past access page and the next access page, the association is more close.

The experimental results show that the association rule revise can improve the prediction accuracy rate, especially x- prediction accuracy rate. Because the action that users access web site has Markov and there is an association between the visited pages, so the combination of Markov and association rules are well consistent with users' behaviors.

Since the final prediction results are influenced by multiple parameters, so this problem can be viewed as a linear programming problem with constraints. Some of the linear planning optimization algorithm which are put forward recently can be used to solve parameters repeated experimental adjustment problems, in order to reduce human intervention and influence in the prediction process as much as possible and to improve the prediction efficiency and accurate rate. At the same time, the accuracy of the prediction and the calculation of the workload shall be on a compromise. In addition, for the experiments on the algorithm, this paper only set experiment for the specific data, because the parameters involved in the prediction have different influence degrees on different data sources. In order to be more convincing, the variety and quantity of verification of data sources can be increased, so as to provide the theoretical reference for practical prediction, which will be one of the next tasks of the paper.

Experimental Analysis of Markov Prediction Model with Feedback. The experimental data is from the access log for one week in February of 1998 of Microsoft server users. 10954 sessions were randomly selected from the experimental training data as the training data of HPT, and then the original tested data is tested based on the existing HPT test data.

After many experiments, the threshold goodT in the algorithm is 0.8. It is predicted that when the threshold min_p in the algorithm is 0.1, the experiment results is the good. If goodT is 0.8 represents that for the current session, if there are prediction results with the prediction accuracy rate which is more than 0.8 in the historical prediction records, the results can be viewed as the current prediction results. The prediction algorithm is not needed. If goodT is 0.1 represents that for the current session, if there are prediction results with the prediction accuracy rate which is less than 0.1 in the historical prediction records and the results are in the candidate set for the current prediction, the results shall be deleted from the candidate set. When goodT is 0.8, min_p is 0.1 and the weight value λ of the historical prediction, 1- prediction accuracy rate is the best. Fig. 3 is the comparison chart of 1- prediction accuracy rate of second order Markov chain prediction algorithm, prediction algorithm based on Markov chain and association rules and MPMF.

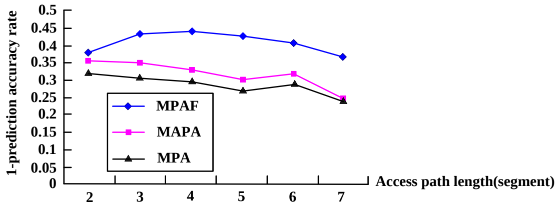

Figure 3. Max 1-Prediction accuracy comparison of the MPA, PAMA and MPMF

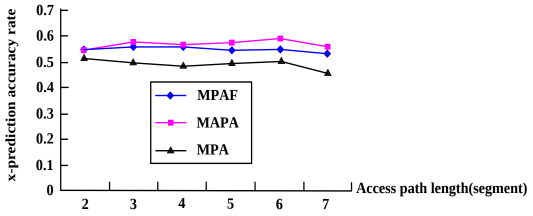

Figure 4. Max x-Prediction accuracy comparison of the MPA, PAMA and MPMF

When goodT is 0.8, min_p is 0.1, the weight of historical prediction λ is 0.1, x- prediction accuracy rate is the best. Fig. 4 is the x- prediction accuracy rate comparison chart of MPA, PAMA and MPMF.

According to Fig. 3, MPMF significantly improves the accuracy rate of 1- prediction. In theory, without considering the size of HPT, then with the increase of the times of historical prediction, the reference significance provided by HPT will be larger, so it can improve the prediction accuracy rate. In order to limit the size of the HPT, the experiment in this paper cut the branches with the times of historical prediction of less than 5 and with the prediction accuracy rate in the interval of [0.4, 0.6].

According to Fig. 4, x- prediction accuracy rate of MPMF is sightly less than that of PAMA. Due to in MPMF, feedback is based on the user's next access, that is, in HPT, the historical prediction record saves 1- prediction accuracy rate, so MPMF has a poor prediction effect for multiple step transfer. Of course, the feedback can also be based on multiple step access, that is to say, the historical prediction records of HPT can save x- prediction accuracy rate, then 1- prediction accuracy rate will be decreased. If HPT saves both 1- prediction accuracy rate and x- prediction accuracy rate, computation complexity of the prediction will be increased. Therefore, according to the real needs, two HPT can be established to save separately 1- prediction accuracy rate and x- prediction accuracy rate.

Conclusion

Mining Web log can predict the user's future accessing page, so as to provide users with personalized service[9-12]. This paper proposes PAMA prediction algorithm based on Markov chain and association rules and Markov prediction model MPMF with feedback. Both the two methods are aiming at improving the accuracy of the prediction by modifying and adjusting the prediction candidate set. PAMA uses the two association rules to revise the prediction results from forward and reverse, which reduces user intervention in the process of prediction and shortens the predicted time, but flexibility is poor. According to the user feedback, MPMF combines with historical prediction information to adjust the prediction results, which emphasizes the user's participation in prediction process and the results of the prediction are more in line with the user's personalized requirements, but the prediction processing time is slightly longer than that of PAMA. Therefore, if the two methods can be perfectly integrated and coordinated to provide users with more choosing space in the operation, the result will be good. What’s more, if efficient processing method is considered to use to preprocess Web log and the combing method of web content mining and web log mining to solve the lagging problem of user access prediction based on Web log mining, the prediction accuracy and prediction efficiency will be further improved, which will be the next step in the future.

Acknowledgement

Fund Project: Supported by the Major Program of National Social Science Foundation of China(Grant No. 13&ZD148) (Major project of 2013 National Social Science Fund)

References

[1] Mathias Géry, Hatem Haddad. Evaluation of web usage mining approaches for user's next request prediction[C]. Proceedings of the 5th ACM international workshop on Web information and data management, New Orleans, Louisiana, USA.2003: 74-81.

[2] Wang Yan, xu Hong-Bin, Yang Zi-Rong. Research on the Application of Association Rule in Web Log Mining[J]. Computer Era, 2008, 12:29-31.

[3] Zhili Zhang, Lei Shi, Shen Guo, Deyu Qi, Fufang Li. Appling Association Rule to Web Prediction[C]. Proceedings of the First International Multi-Symposiums on Computer and

Computational Sciences(IMSCCS'06), 20-24, June, 2006, Vol.2: 522-527.

[4] I. Zukerman, D. W. Albrecht, A. E. Nicholson. Predicting user's requests on the WWW[C]. Proceedings of the seventh international conference on User modeling, Banff, Canada, 1999: 275-284.

[5] LIN Wen-Long, LIU Ye-Zheng, J IANG Yuan-Chun. Web Navigation Prediction Based on Markov Model—A Survey[J]. Computer Science, 2008, Vol.35, No.1: 9-14.

[6] XING Yongkang, MA Shaoping. Modeling user navigation sequences based on multi-Markov chains [J]. Chinese Journal of Computers, 2003, 26(11): 1510-1517.

[7] Sule Gündüz, M.Tamer Özsu. A Web page prediction model based on click-stream tree representation of user behavior[C]. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, USA. August, 2003:535-540.

[8] Yin-Fu Huang, Jhao-Min Hsu. Mining Web Logs to Improve Hit Ratios of Prefetching and Caching[J]. Knowledge-Based Systems, Vol.21, No.1, Feb. 2008: 62-69.

[9] Puntheeranurak Sutheera, Tsuji Hidekazu. Mining Web logs for a personalized recommender system[J]. Joho Shori Gakkai Zenkoku Taikai Koen Ronbunshu, Vol.67, No.3, 2005:19-20.

[10] Xian Xue-Feng, Yang Xue. Design and Implementation of Individuality E-Learning System Based on Web Mining[J]. Journal of Computer Applications, Vol.27, Jun. 2007:31-33.

[11] Subhash K.Shinde, U.V.Kulkarni. A New Approach For On Line Recommender System in Web Usage Mining[C]. Advanced ICACTE'08. International Conference on Computer Theory and Engineering, 2008, 20-22 Dec. 2008: 973-977.

[12] Haibin Liu, Vlado Kešelj. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests[J]. Data & Knowledge Engineering. Elsevier Science Publishers, Netherlands. Vol.61, No.2, May 2007: 304-330.