Web User Behaviors Prediction System Using Trend Similarity

[a,b]NIEN-YI JAN and [b]NANCY P. LIN
[a]Business & Marketing Strategy Research Department
Telcommunication Lab., Chunghwa Telecom Co., Ltd
[b]Department of the Information Engineering
Tamkang University
151 Ying-chuan Road, Tamsui
Taiwan 251, R.O.C.
http://mail.tku.edu.tw/125502/csie/pclin.htm

Abstract: - In this paper, a new prediction model without extra information is proposed to construct prediction models for a proxy server prefetcting the suitable pages for adapting the website structure and improving the website performance. The trend similarity is proposed to consider both the page similarity and the position similarity of all pages. A testing websites are established in our experimental environment to evaluate accuracy of our prediction system using page correctness rate and order correctness rate criteria. The experimental results show that the 70% prefetched pages are useful.

Key-Words: - Trend similarity, Prediction system, Web mining

1 Introduction

As the World Wide Web is growing at a rapid rate, the web-based services and applications become more popular. Users are interested in seeking useful information on Internet and are highly requesting for the correctness in quickly response. How to adapt to the structure of website to meet the needs of users becomes an important task for a website manager.

To build suitable models by analyzing user browsing logs to predict the future traveling path and to facilitate the perfecting in proxy servers. Proxy servers play a key role between users and websites, which could reduce the response time of user requests and save network bandwidth. Many researchers have proposed various effective caching and prefetching algorithms in proxy servers to improve the website performance. Web mining technology is widely used on constructing prediction models to dig deeply into the interests of web users by analyzing their navigation behaviors. Therefore, the web administrator can adapt the website structure and predict web user prospective traveling path to provide the suitable information for users through analyzing the user behaviors [15][11]. Many distinctive methods are proposed to construct a prediction model such as association rules [15] and Markov chain [16], most of them however need extra information to improve the performance, e.g., website topology and the similarity between pages. Some researchers cluster similar browsing behaviors using clustering technology, most of them however focus on clustering similar browsing behaviors or similar pages. These prediction systems only predict which pages will be accessed but ignores the navigation sequences. In this paper, we will propose a trend based prediction system including two phases (prediction model constructing phase and predicting phase) to analyze user behaviors and predict the future traveling path based upon the trend similarity. Each prediction model consists of prefix pattern and postfix pattern, where the prefix pattern is used to help system determine all similar behaviors and the postfix pattern is used to select the page candidates. The trend similarities between the new user browsing behavior and these prefix patterns can be computed to select suitable prediction page candidates with ordering. Moreover, new prediction modes will be incrementally constructed for improving the accuracy and efficient of prediction. Two measures including page correctness rate and order correctness rate are designed to evaluate the accuracy of our prediction system.

2 Related Works

Due to the dramatic growth of the Internet, the number of transaction becomes larger and larger. It results in the difficulty of analyzing these complicated data using traditional analysis method. Many researchers tend to discover the potential behaviors using data mining approaches to predict the behavior trend [10], to improve the performance of network usage, to support decision in EC, and to provide a guideline in designing website, etc.

In [6], web mining can be divided into three categories according to different data types: (1) web content mining, (2) web usage mining, and (3) web structure mining. In 2000, Cooley considered user profile in web mining to improve the accuracy of analyzing the user behavior with similar profile [3]. Web content mining [2] is proposed to discover the meaningful information in the content of web pages. Unlike web content mining, web structure mining is proposed to discover the relationship or link between web pages according to the characteristic of web content. It can be used to adjust the structure of website after discovering the structure of web content. Web usage mining [1][5] is proposed to monitor the interaction between users and website for discovering useful information through analyzing the collected user navigation patterns [8].

M. Spiliopoulou [9] concluded three goals in web usage mining according to different domains. (1) To predict the user behavior of a website can be applied to reduce the latency of network access. (2) To compare the realistic performance of the website with the ideal performance. The website during its initial designing version might be not suitable for users. However, designer can easily adjust the website to improve the efficiency of a website according to the log analysis results. (3) To adjust the website structure according to the user interest. It could be improve the performance of website after understanding the personality of each user.

Many researchers proposed Markov Chain [16], clustering [12], and web mining to predict the browsing behaviors of users. The Markov chain is proposed to predict the maximal probability of next browsing page by calculating each probability of each link according the web structure and the historical browsing behaviors [16]. Most researchers proposed web mining to analyze the user behaviors for predicting the future browsing behaviors [11][8]. [7] focused on increasing the performance of prediction. [15] used association rule mining to discover the co-occurrence web pages embedded into the request pages in advance for reducing the latency of network access. [13] proposed a belief function to cluster similar user behaviors for predicting reference.

Three purposes for predicting network user behaviors are focused on making efficient recommendation for users [4] and reducing the latency of browsing web sites using prefeching [15][11]. Since the limitation of space in storing the perfecting pages, it is important to predict which pages could be re-used when perfecting.

3 Prediction System Architecture

Due to the variousness and uniqueness of user behaviors, it is inappropriate to predict the browsing behaviors of current user according to the similarity comparison with single browsing cluster of older users. Therefore, a trend-based prediction model is proposed to predict the further traveling path by generating ordering browsing sequence.

3.1 Framework of Prediction

In order to reduce the latency of browsing websites, a trend based prediction method is proposed in our prediction system to download some relevant web pages in advantage on local user's computer according to the previous browsing sequence. The prediction system is divided into two components which shown in Fig. 1. One is prediction model constructing phase, and the other is predicting phase.

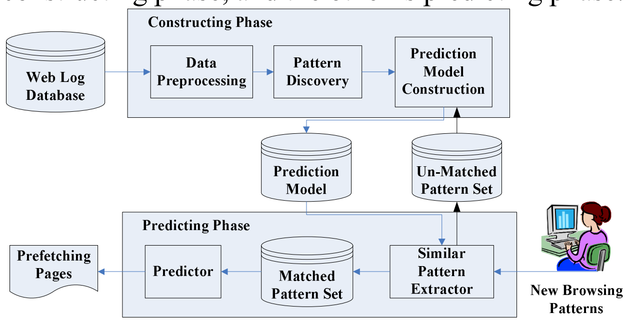

Fig. 1: Our prediction system architecture.

First, the prediction model of the website could be established in the constructing phase using web data mining approach after collecting enough web browsing log. In the predicting phase, the browsing behaviors of new users can be obtained to compare the similarity with the prediction model. Hence, the candidate pages could be prefetching to improve the browsing performance by applying on the replacement algorithm of proxy servers.

3.2 Prediction Model Constructing Phase

As mentioned above, the user behaviors are various, the prediction model constructing phase including data preprocessing, pattern discovery and prediction model construction steps is proposed to help experts discover useful common browsing patterns and then used to predict the further browsing sequences.

The browsing behavior can be treated as a sequence of browsing pages. The prediction model can be divided into prefix pattern and postfix pattern. Hence, the user browsing behavior can be predicted based upon the correlations between these patterns. The prediction model example can be shown as Fig. 2. In this example, the prefix pattern is {A0, B0, C3, D6}, which is length 4, and the postfix pattern is {D7, D8, D9, D16, D3, D9}, which is length 6. The model means that if an user browsed A0, B0, C3, D6 sequentially, the model will predict the D7, D8, D9, D16, D3, D9 pages might be browsed after several minutes.

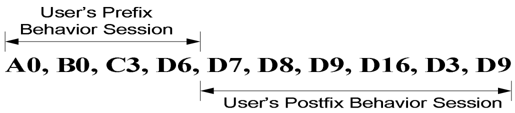

Fig. 2: User's behavior session example.

Since the original web log records all user request pages, which often result in several log entries, it is necessary to eliminate these irrelevant pages in order to obtain the actual user browsing behavior on the web server. Besides filtering the automatically generated pages, it is also important to distinct different behavior session for discovering different browsing behavior in a same user. In order to obtain the actual browsing behavior, we retain only the useful records such as .asp or .html could improve the accuracy and practicability of prediction model to reduce the complexity of post processing.

The behavior session is defined as one browsing behavior of a user. Different sessions could be generated by different times from same users or different users. Therefore, based upon user login and logout period to distinct different session of a user would be more suitable for session identification. However, we cannot enforce a user to log into a popular server before browsing. Hence, Cooley's [3] method is used to distinguish user navigation patterns, and give a name Behavior Session according to IP and agent log during a period of time, e.g., 30 minutes.

After data preprocessing phase, the user behavior could be represented as a sequential of browsing pages. Each browsing sequence can be separate into two parts: prefix pattern and postfix pattern. The cut point can be chosen by the application.

Although the user behaviors are various and distinctive, the browsing patterns of a structure website could be similar. These behavior patterns could be used to predict the further browsing behaviors of new users. Hence, it is important to discover the significant user patterns within web log database.

In order to obtain the significant patterns for predicting the user browsing patterns, a maximal closed itemsets algorithm [14] is chosen in this paper to discover the frequent prefix patterns using the part of prefix patterns dataset, i.e., satisfy the support threshold.

3.3 Prediction Model Constructing

As mentioned above, the browsing behaviors are various, it means that users many browsing different postfix behavior patterns when they have the same prefix behavior patterns. On the other hand, they have the same postfix behavior patterns when they have similar prefix behavior patterns. Both mean that these users have similar browsing behaviors. Hence, it is reasonable to predict the postfix patterns through collecting all browsing behavior patterns of similar users. Two steps (including prediction page set determination and top-k prediction pages sorter) are designed in prediction model constructing phase.

All pages in user's postfix behavior sessions whose occurrences reach Confidence threshold will be discovered as prediction set. To build the representative prediction model, all the page candidates could be discovered after setting a confidence value based upon each same prefix pattern sessions. For example, let the confidence value is set to 30%. There are 11 users with the prefix patterns {A0, B0, C10, D6} in Table 1. Therefore, we can obtain six pages {D2, D10, D29, D34, D41, D43}, which satisfies the confidence.

Table 1: The partial processed web log.

User	Prefix behavior session				Postfix behavior session					
	1	2	3	4	1	2	3	4	5	6
U1	A0	B0	C10	D6	D10	D44	D3	D0	D29	D28
U2	A0	B0	C10	D6	D10	D34	D43	D41	D30	D19
U3	A0	B0	C10	D6	D23	D6	D34	D43	D41	D30
U4	A0	B0	C10	D6	D2	D9	D42	D28	C9	D29
U5	A0	B0	C10	D6	D34	D37	D40	D32	D0	D29
U6	A0	B0	C10	D6	D2	D9	D42	D19	D39	D15
U7	A0	B0	C10	D6	D34	D43	D41	D10	D0	D37
U8	A0	B0	C10	D6	D10	D4	D16	D32	D8	D12
U9	A0	B0	C10	D6	D2	D23	D36	D44	D3	D1
U10	A0	B0	C10	D6	D10	D34	D37	D38	D16	D29
U11	A0	B0	C10	D6	D2	C14	D18	D13	D43	D41
U12	A1	B0	C10	D6	D2	D9	D2	C14	D4	D16
U13	A1	B0	C10	D6	D23	D11	D31	D2	D41	D30
U14	A1	B0	C10	D6	D34	D6	D23	D36	D44	D27
U15	A1	B0	C10	D6	D23	D36	D39	D3	D0	D29
U16	A1	B0	C10	D6	D23	D11	D31	D2	D38	D4

Since each prefix behavior session can lead to various number of postfix behavior sessions, the number of picked up prediction pages would be different. Besides predicting the next pages, the ordering of page sequence are also important when building prediction model. Hence, we proposed top-*k* prediction pages sorter to determine the ordering of prediction sequence. The length of Postfix behavior Patterns is equal to the number of page prediction set.

In order to determine the ordering of page set, a weight vector is used to indicate that which page occurs in which position in each page candidate. The weight vector is set to increasing order. Equation (1) is used to calculate ordering grade to determine the ordering of page candidates.

$$O_{v_i}^{j} = (\sum S_{v_i}^{j}) / |PS_j| \qquad (1)$$

where PS_j is the postfix behavior sessions with the same prefix behavior pattern j, v_i is the i^{th} page in a page candidate set, $O_{v_i}^{j}$ is the ordering grade of v_i in PS_j, and $|PS_j|$ is the number of postfix page in PS_j. In Equation (2), $S_{V_i}^{j}$ is the function of determining position weight before confirming the ordering of page candidates. The PenaltyWeight = $|PS_j|$ + 1.

$$S_{v_i}^{j} = \begin{cases} PositionWeight & v_i \in PS_j \\ PenaltyWeight & v_i \notin PS_j \end{cases} \qquad (2)$$

For example, as mentioned above, {D2, D10, D29, D34, D41, D43} are the prediction page candidate shown in Table 1, the weight vector is set as {1, 2, 3, 4, 5, 6, 7} if the maximal length of postfix is 6. The ordering score of D10 is 4.5, which can be computed using Equations (1) and (2). Therefore, we can obtain an predicting sequence D10, D34, D2, D43, D41 if we choose only top-5 postfix pages. The prediction model will be set to {A0, B0, C10, D6} → {D10, D34, D2, D43, D41}.

3.4 Predicting Phase

The browsing behaviors of new visitors in the website can be predicted according to the prediction model. Since the browsing behavior might be similar with many prefix behavior sessions or not, the predictable patterns should be merged or the new prediction pattern should be constructed, respectively. The browsing behavior of a new visitor can be extracted to compare with all matching patterns in prediction model for predicting the future traveling path.

Later browsing behaviors would be more meaningful than the previous ones since the user's navigation path is the time series. Hence, a half decreasingly grading heuristic is proposed to implement the trend similarity to select all proper prediction patterns as Reference Patterns. The initial similarity is set to 0. To predict the browsing behaviors of new visitors with the generated browsing patterns, a pre-defined similarity threshold would be picked to filter some suitable prefix patterns to generate the probably patterns. Extracting the prediction patterns by trend similarity can get higher predicting accuracy.

$$Sim_{trend}(M_n^i, U_n^j) = \begin{cases} Sim_{Trend}(M_{n-1}^i, U_{n-1}^j) + n, & \text{if } m_n = u_n \\ Sim_{Trend}(M_{n-1}^i, U_{n-1}^j)/2, & \text{if } m_n \neq u_n \end{cases}, \qquad (3)$$

where n is the length of the ith comparing pattern. However, the similarity should be normalized by dividing the sum of position weight. For example, let M1 = {ABCD} be the Match Pattern and U1 = {ABCI}, U2 = {IBCD} are user's navigation patterns. By equation (3), the similarity between M1 and U1 is 0.3, and the similarity between M1 and U2 is 0.9. Therefore, the prediction pattern that contains M1 is proper for U2 than U1. Lets assume the similarity threshold is set to 0.5, the reference pattern can be obtained when the similarity is large than the threshold. Therefore, we can obtained the reference pattern set is P1, P6, and P10. The partial prediction pattern models are shows in Table 2.

Table 2: The partial prediction pattern models

	Match Pattern				Probable Pattern						Similarity
P1	A0	B0	C10	D6	D10	D17	D2	D29	D41	D43	0.61
P2	A1	B2	C3	D5	D22	D13	D9	D28	D13	D15	0
P3	A1	B4	C2	D6	D9	D5	D8	D17	D19	D11	0
P4	A0	B0	C4	D6	D34	D17	D18	D41	D38	D37	0.32
P5	A3	B5	C7	D9	D18	D36	D14	D35	D23	D40	0
P6	A4	B0	C10	D6	D22	D17	D9	D2	D43	D13	0.72
P7	A5	B5	C9	D24	D17	D20	D16	D39	D19	D21	0
P8	A0	B2	C10	D6	D10	D13	D3	D9	D29	D1	0.36
P9	A3	B9	C12	D6	D17	D9	D10	D15	D26	D29	0.22
P10	A0	B3	C10	D5	D10	D9	D17	D34	D41	D5	0.68

At the previous step, all the postfix page candidates are chosen to decide the sequence of prediction pages. In Predictor, a prediction principal will be proposed to sort the page sequence and to decide the web user prospective traveling path. Similar to construct the prediction model, the ordering between page candidates should be ranked. Since the similarity between reference patterns and match pattern are usually different, the similarity should be considered as a merged weight when sorting these candidate pages. The grading function to sort the candidate pages is shown as function (4).

$$G_{v_i} = \sum (Sim_{Trend}(M^k) \times S_{v_i}^k) / \sum Sim_{Trend}(M^k) \qquad (4)$$

where G_{v_i} is the ordering grade of predicting page vi, SimTrend(M^k) is the trend similarity between the match pattern M^k and the browsing pattern, $S_{v_i}^k$ is the position weight which is mentioned above. In this paper, we assume that the weight vector is {1, 2, 3, 4, 5, 6, 7} while the length of postfix pattern is 6.

For example, P1, P6, and P10 are the reference pattern set which are similar with the new user behavior. Therefore, the predicting page candidate set will be {D2, D5, D9, D10, D13, D17, D22, D29, D34, D41, D43}. Hence, the ordering grade of page D2 is 4.710, which can be calculated using function (4).

Depending on the ordering grade result, the prediction sequence will be {D17, D10, D9, D2, D22, D41, D34, D43, D29, D13, D5}. The constructing prediction model can prefectch the necessary page for proxy server according to the ordering grade.

Although the similar pattern extractor can not find any suitable prediction pattern due to the various behavior patterns, these unpredictable user's navigation patterns can be stored in the unmatch pattern database to learn the novel prediction models.

4 Experimental Result

4.1 Accuracy Measurement

In the prediction phase, each testing pattern is divided into two parts. First part is the browsed pages of this new user and the second part is the further browsing pages after verification. The similarity between matching pattern and the browsed pattern will be first calculated. The reference patterns will then be selected to generate the prediction result. To evaluate the accuracy of prediction result in our prediction system, two measuring criteria includes Page Correctness Rate (α) and Order Correctness Rate (β) are proposed to examine the predicting sequence.

The page correctness rate, computed as *countA/l*, is used to calculate the coverage of predicting sequence in the realistic further browsing patterns, where *countA* is the number of the same pages between testing pattern and predicting results, and *l* is the length of the testing pattern.

Since the ordering of a sequence can be treated as the set of multiple ordering pairs, the ratio of the same pairs of postfix patterns between predicting patterns and verification patterns can be calculated the correctness of the sequence ordering. The order correctness rate is shown in function (5), where $\sum_{x=1}^{n-1} x$ is the total number of order-pairs, *n* is the number of page candidates, and *countB* is the same order pairs between testing pattern and predicting results.

$$\beta = countB / \sum_{x=1}^{n-1} x \qquad (5)$$

For example, T = {A, B, C, D} is one user actual browsing pattern, and N = {B, D, C, I} is the predicting result by our prediction system. The Page Correctness Rate (α) is 3/4 = 0.75. Order-pairs included in T are {(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)}. The order-pair (A, B) means that A comes before B and so on. In this example, the prediction system has predicted B, C, D pages. The actual order-pairs are (B, C), (B, D), (C, D). Only (B, D), (B, C) has been predicted and this results in that the Order Correctness Rate (β) is 2/3.

4.2 Experimental Results

We simulated 20000 various user behavior sessions based upon website topology. The length of each session is set to 10 after data preprocessing, where the patterns will be ignored if the length is less than 10. In our experimental environment, 18000 user patterns are used to train the prediction model and the remaining 2000 patterns are used to test the accuracy of prediction models. 10-fold cross verification is used to enhance the confidence of evaluating our prediction model.

Predict:	A4	B9	C18	D8	D42	D35	D14	D25	D5	D32
Predict:	A2	B6	C10	D23	D22	D3	D29	D43	D11	D19
Predict:	A2	B4	C10	D2	D40	D15	D36	D12	D4	D31
Predict:	A3	B10	C17	D8	D42	D32	D37	D25	D35	D5
Predict:	A3	B10	C13	D6	D9	D38	D16	D20	D18	D28
Predict:	A1	B1	C0	D3	D29	D22	D39	D43	D19	D23
Predict:	A2	B6	C10	D19	D22	D3	D29	D43	D23	D11
Predict:	A1	B1	C0	D0	D40	D4	D36	D31	D15	D12
Predict:	A3	B7	C15	D37	D5	D14	D25	D42	D8	D35

Fig. 3: Partial prediction patterns.

Lets assume the support is set to 0.002 to filter the maximal frequent prefix patterns and the confidence is set to 50% to choose the candidate pages of postfix for each discovered prefix to construct prediction model. Fig. 3 shows the partial prediction patterns in our experiment. In our experiment, each prediction model is divided into two parts. First is match pattern whose length is 4. Second is probable pattern whose length is 6.

Fig. 4 shows the relationship between prediction correctness and the trend similarity. The result shows that the latest browsing behaviors can lead to more predictable pages for browsing. In Fig. 4, we observed that the rate of page correctness is less than 35% and the ordering correctness is less than 30% if the last page is unmatched. On the other hand, the page correctness increases to 73% and the order correctness increases to 46% when the last page is matched.

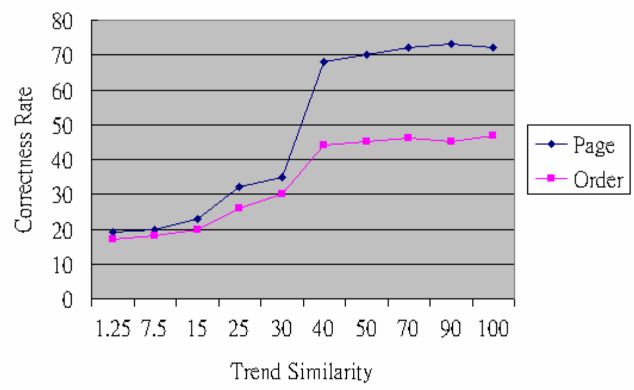

Fig. 4: The prediction accuracy between different trend similarities.

5 Conclusion

We proposed a trend based prediction system was proposed to predict the various web user behaviors. We design a trend similarity to select the proper prediction patterns for predicting a new user browsing behavior. Our prediction patterns can be collected to learn the new prediction models to achieve the flexibility and adaptation of our prediction system. The results show the performance of our proposed model is useful to prefetch candidates pages.

Reference:

[1] S. Araya, M. Silva, and R. Weber, "A methodology for web usage mining and its application to target group identification," *Fuzzy Sets and Systems*, Vol. 148, No. 1, Nov., 2004, pp. 139-152.

[2] L. H. Chen, W. L. Chue, "Using web structure and summarisation techniques for web content mining," *Information Processing and Management*, Vol. 41, No. 5, Sep 2005, pp. 1225-1242.

[3] R. W. Cooley, "*Web usage mining: discovery and application of interesting patterns from web data*," Ph. D. dissertation, Dept. of Computer Science, University of Minnesota, May 2000.

[4] Y. T. Lin and S. S. Tseng, "A characterized rating recommend system," in *Proc. of the 5th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 2001, pp. 41-46

[5] H. B. Liu and V. Kešelj, "Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data & Knowledge Engineering*, Vol. 61, No. 2, May 2007, pp. 304-330.

[6] S. Madria, S. S. Bhowmick, W. K Ng, and E. P. Lim, "Research issues in Web data mining," in *Proc. of the 1st Int'l Conference on Data Warehousing and Knowledge Discovery*, 1999, pp. 303-312.

[7] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos, "Effective prediction of web-user accesses: A data mining approach," in *Proc. of the Workshop WEBKDD*, 2001.

[8] M. L. Shyu, S. C. Chen, and C. Haruechaiyasak. "Mining user access behavior on the WWW," in *Proc. of IEEE Int'l Conf. on Systems, Man, and Cybernetics*, Vol. 3, 2001, pp. 1717-1722.

[9] M. Spiliopoulou, "Data mining for the web," in *Proc. of the 3rd European Conf. on Principles of Data Mining and Knowledge Discovery*, 1999, pp. 588-589.

[10] V. S. Tseng and K. W. Lin, "Efficient mining and prediction of user behavior patterns in mobile web systems," *Information and Software Technology*, Vol. 48, No. 6, June 2006, pp. 357-369.

[11] Y. H. Wu and A. L. P. Chen, "Prediction of web page accesses by proxy server log," *World Wide Web*, Vol. 5, No. 1, 2002, pp. 67-88.

[12] J. T. Xiao and Y. C. Zhang, "Clustering of web users using session-based similarity measures," in *Proc. of the 2001 Int'l Conf. on Computer Networks and Mobile Computing*, 2001, pp. 223-228.

[13] Y. J. Xie and V. V. Phoha, "Web user clustering from access log using belief function," in *Proc. of The 1st Int'l Conf. on Knowledge Capture*, 2001, pp. 202-208.

[14] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining closed sequential patterns in large database," in *Proc. of SIAM Int'l Conf. on Data Mining*, 2003, pp. 166-177.

[15] Q. Yang, H. H. N. Zhang, and T. Y. Li, "Mining web logs for prediction models in WWW caching and prefetching," in *Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2001, pp. 473-478.

[16] J. H. Zhu, J. Hong, and J. G. Hughes, "Using Markov models for web site link prediction," in *Proc. of the 13th ACM Conf. on Hypertext and Hypermedia*, 2002 , pp. 169-170.