Knowl Inf Syst (2006) 10(4): 515–528
DOI 10.1007/s10115-006-0004-z

Knowledge and Information Systems

SHORT PAPER

Mei-Ling Shyu · Choochart Haruechaiyasak · Shu-Ching Chen

Mining user access patterns with traversal constraint for predicting web page requests

Received: 23 January 2004 / Revised: 29 March 2005 / Accepted: 10 December 2005 /
Published online: 22 April 2006

Abstract The recent increase in HyperText Transfer Protocol (HTTP) traffic on the World Wide Web (WWW) has generated an enormous amount of log records on Web server databases. Applying Web mining techniques on these server log records can discover potentially useful patterns and reveal user access behaviors on the Web site. In this paper, we propose a new approach for mining user access patterns for predicting Web page requests, which consists of two steps. First, the Minimum Reaching Distance (MRD) algorithm is applied to find the distances between the Web pages. Second, the association rule mining technique is applied to form a set of predictive rules, and the MRD information is used to prune the results from the association rule mining process. Experimental results from a real Web data set show that our approach improved the performance over the existing Markov-model approach in precision, recall, and the reduction of user browsing time.

Keywords Web usage mining · Association rule mining · Mining user access patterns

M.-L. Shyu (✉)
Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33124, USA
E-mail: shyu@miami.edu

C. Haruechaiyasak
Information Research and Development Division (RDI), National Electronics and Computer Technology Center (NECTEC), Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand

S.-C. Chen
Distributed Multimedia Information System Laboratory, School of Computer Science, Florida International University, Miami, FL 33199, USA

1 Introduction

Web usage mining is the process of applying data mining techniques to discover usage patterns from Web data [15, 17]. The technique of mining user access patterns (also known as *browsing patterns* and *path traversal patterns*) has been applied in a wide range of applications including Web caching [12, 14], Web page recommendation [8, 10], Web search engine [18], and Web personalization [11, 13]. In general, mining user access patterns can be considered as a special type of mining sequential patterns in the field of knowledge discovery and data mining. Mining sequential patterns is the process of discovering subsequences of data such that the frequency of occurrence exceeds a user-specified minimum support [2]. For example, in the market-basket databases, each data sequence corresponds to the items bought by an individual customer over time. Applying a method for mining the sequential patterns on these databases yields some frequently occurring item sequences, which can be very useful for predicting the future customer purchasing behavior.

In the context of mining user access patterns, data sequences are typically user access sequences of Web pages. These access sequences are extracted from the server log records via some Web data preparation techniques [6]. Mining user access patterns on these access sequences reveals user browsing behavior on the Web. The main difference between a method for mining user access patterns and a method for mining sequential patterns lies in the data set characteristics. While both methods consider the data sets containing time-ordering sequences, a method for mining user access patterns, however, considers access sequences which are constrained by the hyperlink structure of the Web. On a Web site, a user normally retrieves Web pages by following the hyperlinks embedded within the Web pages. Thus, considering the structural constraint in the algorithm for mining access pattern could increase the prediction accuracy and also reduce the space complexity of the prediction model.

In this paper, a new approach of mining user access patterns based on *association rule mining* with the inclusion of the traversal constraint is proposed. Recently, the association rule mining technique has attracted strong attention and proven to be highly successful in extracting useful information from very large databases [1, 9]. Our proposed framework consists of two steps: (1) the construction of user-based structure of the Web site and (2) the association rule mining process. Figure 1 illustrates the proposed framework. The goal of the first step is to consider traversal constraints by determining the reachability between Web pages. To capture the user access behavior on the Web, an alternative structure of the Web is constructed from user access sequences, as opposed to static structural hyperlinks. Using this user-based structure, a shortest path algorithm in graph theory is applied to find the distances between Web pages. We refer to these distances as *Minimum Reaching Distance (MRD)* information. MRD information is used to determine the reachability between Web pages and to evaluate the performance of the algorithms. In the second step, association rule mining is applied to find a set of predictive rules that passes the user-specified minimum support. The MRD information is used to prune the results from association rule mining in order to reduce the space complexity.

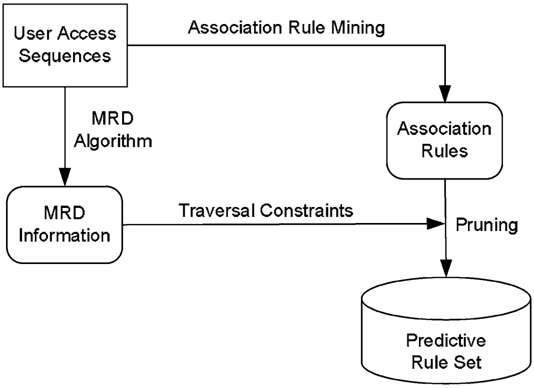


Fig. 1 Association rule mining framework for predicting Web page requests

Our approach aims at predicting Web page requests on the Web site in order to reduce the access time and to assist the users in browsing on the Web. Previous research studies in mining user access patterns for predicting Web page accesses have focused only on consecutive sequential access of Web pages. These methods are based on the notion of Markov model [3, 12, 13]. The ability to predict the next request from a given Web page is limited to the following adjacent Web pages determined by their access probabilities. By applying association rule mining, our method allows the prediction to include multiple nonconsecutive Web pages. The experimental results based on a real data set show that our approach yields better performance in terms of *precision* (prediction accuracy) and *recall* (prediction coverage) over the Markov-model based approach.

The remainder of this paper is organized as follows. In the next section, a review of related work in mining user access patterns, particularly, for predicting Web page accesses is given. In Sect. 3, the method of mining user access patterns based on association rule mining is explained in detail. The experiments and results are given in Sect. 4. The paper concludes in Sect. 5.

2 Related work

Various algorithms and techniques for mining user access patterns have been proposed in the literature. The focuses of these approaches are different in terms of implementation details and application usage. For example, an affinity-based clustering approach was proposed to capture the user access behavior on the World Wide Web (WWW), where the result can be used to re-organize the Web site for electronic-commerce purpose [16]. Chen et al. [5] proposed an algorithm for finding the maximal forward references from user traversal sequences was proposed to filter out the effect of backward references and then two algorithms based on association rule mining were proposed to capture the frequent traversal patterns. Several approaches applied the Markov models for mining user access patterns [7, 12, 13]. For example, Padmanabhan and Mogul [12] proposed a predictive model called the first-order Markov model to predict the users' requests on Web pages for the purpose of latency reduction. Pitkow and Pirolli [13] proposed a general form of the Markov model called All Kth-order Markov model to improve the coverage and accuracy over the first-order Markov model approach. In order to further improve the prediction accuracy and to reduce the space complexity, Deshpande and Karypis [7] later proposed the variations of the All Kth-order Markov model. However, the focus of most of these approaches is on the analysis of consecutive sequential access of Web pages. That is, given a currently visited Web page, the ability to predict the next request is limited to the following adjacent Web pages on the user access sequence. For example, given a user access sequence containing n Web pages in an ordered list $(p_1, p_2, \ldots, p_n)$, an approximation of the first-order Markov model would contain the transitional probabilities of two adjacent Web pages, $Pr(p_i \mid p_{i-1})$, in the user access sequence, where p_i represents a Web page and $1 < i \leq n$.

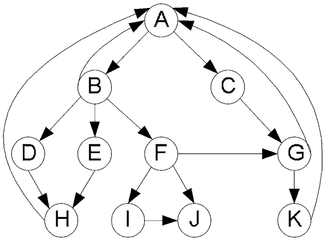
To allow the prediction to include multiple nonconsecutive Web pages, a new approach for mining user access patterns based on association rule mining is proposed in this paper. Under the Markov model notion, our method can be viewed as the All Kth-Order Markov model with "look-ahead" ability. Using a similar notation, the approximation of the first-order Markov model with the look-ahead ability would contain the following transitional probabilities of two Web pages including nonconsecutive ones, $Pr(p_j \mid p_i)$, where $1 \leq i < n$ and $i < j \leq n$. Therefore, our approach can be viewed as a more general form of the existing approaches that only considered the consecutive access of the Web pages.

3 Mining user access patterns via association rule mining

In this section, the Minimum Reaching Distance (MRD) information among the Web pages is first described, and then the algorithm for mining user access patterns based on the association rule mining technique is explained. The MRD information is derived from the user access sequences and used to prune the rules generated from the association rule mining process. The pruned rule set cannot only increase the accuracy of the prediction of Web pages, but also reduce the space complexity.

3.1 Minimum reaching distance (MRD)

The main goal of the *shortest-path* problem in the Graph Theory is to find the path in a weighted graph connecting two given nodes, with the property that the sum of the weights of all the edges is minimized over all such paths. The original algorithm assumes that the traversal path follows the link structure of the graph. However, the user access does not always follow the link-structure property of the graph. Therefore, we must construct an alternative Web representation based on the actual user browsing activity on the Web site. For example, from Fig. 2, s1 contains the following three subpaths, A → B, B → D, and D → H. Using all the user access sequences, a graph can be constructed based on these user subpaths, as opposed to the link structure of the graph. The top of Fig. 3 shows the result of constructing a graph based on the user access sequences from the example data given in Fig. 2. The new graph contains some additional paths such as E → J, which does not previously exist in the link-structure graph in Fig. 2. Also, some



Link Adjacency List

A	B C
B	A D E F
C	G
D	H
E	H
F	G I J
G	A K
H	A
I	J
J	-
K	A

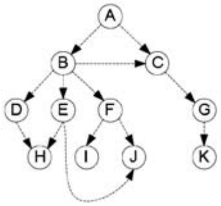
User Browsing Sequence

s1: (A, B, D, H)
s2: (A, B, D)
s3: (A, B, C)
s4: (A, B, E, H)
s5: (A, B, F, I)
s6: (B, F, J)
s7: (A, C, G, K)
s8: (A, B, E, J)

Fig. 2 Examples of link structure and user access sequences

User Traversal Path Adjacency List

A	B C
B	C D E F
C	G
D	H
E	H J
F	I J
G	K
H	-
I	-
J	-
K	-



1st Iteration

	A	B	C	D	E	F	G	H	I	J	K	done
A		1	1									F
B			1	1	1	1						F
C							1					F
D								1				F
E								1		1		F
F									1	1		F
G											1	F
H												T
I												T
J												T
K												T

2nd Iteration

	A	B	C	D	E	F	G	H	I	J	K	done
A		1	1	2	2	2	2					F
B			1	1	1	1	2	2	2	2		F
C							1				2	F
D								1				T
E								1		1		T
F									1	1		T
G											1	T
H												T
I												T
J												T
K												T

3rd Iteration

	A	B	C	D	E	F	G	H	I	J	K	done
A		1	1	2	2	2	2	3	3	3	3	F
B			1	1	1	1	2	2	2	2	3	F
C							1				2	T
D								1				T
E								1		1		T
F									1	1		T
G											1	T
H												T
I												T
J												T
K												T

4th Iteration

	A	B	C	D	E	F	G	H	I	J	K	done
A		1	1	2	2	2	2	3	3	3	3	T
B			1	1	1	1	2	2	2	2	3	T
C							1				2	T
D								1				T
E								1		1		T
F									1	1		T
G											1	T
H												T
I												T
J												T
K												T

Fig. 3 Construction of the minimum reaching distance (MRD) information

of the links in Fig. 2 such as F → G does not appear in Fig. 3, since no user has traversed these particular links. Therefore, using the user-based graph offers a better view of the user access behavior on the Web site than using the link-structure graph.

Table 1 gives our proposed MRD algorithm that includes the detection of an early completion of the algorithm. Many observations have been shown that the user requests for Web pages follow the Zipf-like distribution. As a result,

Table 1 Minimum reaching distance (MRD) algorithm

```
INPUT:
(1) $n$: number of Web pages;
(2) $M[n][n]$: an adjacency matrix for a user-based graph with an array $done[n]$;
(3) $iter$: a step number.

$iter = 1$;
while ($done[a]$ = FALSE, $\exists a$, $1 \leq a \leq n$) {
    for (all rows of $M[i][j]$, where $done[i]$ = FALSE) {
        $done[i]$ = TRUE;
        for (all columns of $M[i][j]$, where $M[i][j] = iter$) {
            for (all columns of $M[j][k]$, where $M[j][k] > 0$, $\forall k$, $1 \leq k \leq n$) {
                if ($M[i][k] = 0$) {
                    $M[i][k] = iter + 1$;
                    $done[i]$ = FALSE;
                }
            }
        }
    }
    $iter = iter+1$;
}
```

the user-based graph is likely to be sparse. Therefore, the detection of the early completion would increase the scalability when applied to a large Web site. The execution time of the MRD algorithm depends on the number of nodes (i.e., the Web pages) and the density (i.e., the averaged number of outgoing links per node) of the graph. In general, the time complexity of the algorithm is $O(n^3)$, where n is the number of nodes.

Figure 3 also shows the execution of finding the MRD information for the user-based graph. During the first iteration, the algorithm translates the user-based graph into the adjacency matrix structure M. The value in the position M$[p_i][p_j]$ is set to one if p_j has a direct link from p_i in the user-based graph. We denote all zero values by blank spaces. The element of $done[p_i]$ is set to false (F) if row p_i of the adjacency matrix contains at least one direct link from the user-based graph. During the next iteration, the algorithm finds the possible reaching distance from the Web page whose value in the *done* array is false. For example, consider Web page A, A has the direct links to B and C. In order to search for other possible reaching distances, the algorithm looks for the direct links from B and C. Since B has the direct links to C, D, E, and F; and C has the direct link to G. Therefore, A can reach D, E, F, and G in two steps. Note that since C is reachable from A in one step, therefore C is ignored from consideration during the second iteration. The element $done[p_i]$ is set to true (T), if there is no change of the value in row p_i of the adjacency matrix during an iteration. The algorithm terminated when all the elements in array *done* are true. For the example data set, the program terminates after the 4th iteration.

3.2 Mining user access pattern

In our proposed framework, the association rule mining technique [1, 2] is adopted to construct the predictive model for mining user access patterns. In association rule mining, a set of items that passes the user-specified minimum *support* value is referred to as an *itemset* and an itemset that contains k items is called a k*-itemset*. Our framework applies the association rule mining technique to find the frequent itemsets of Web pages from the user access sequences and to construct a set of rules based on those itemsets. Generally, the number of rules constructed from the association rule mining technique is large. In the original algorithm, it uses the user-specified minimum *confidence* value to prune the rules. In our proposed framework, the number of rules is pruned down by incorporating the MRD information, which reduces the state complexity of the model.

Consider the user access sequences in Fig. 2 as the data set to the association rule mining and suppose that the support value is set to 2. Therefore, only the itemsets which co-occurred in two or more data records would pass this support constraint. For example, A and B co-occurred in s1, s2, s3, s4, s5, and s8, and therefore itemset (A, B) has the support of 6 which passes the required minimum support of 2. For the example data set, the maximum number of items in an itemset is equal to 3. All resulting itemsets (2-itemsets and 3-itemsets) are shown in Fig. 4. The next step is to generate the rules from these itemsets. To build the model for predicting Web page accesses, we consider the rules with single-item consequence. For example, three single-consequence rules can be generated from the itemset (A, B, D): (B, D) → A, (A, D) → B, and (A, B) → D.

Next, the MRD information from Fig. 3 is used to prune the resulting rules as follows. Consider a single-consequent rule of the form $(p_1, p_2, \ldots, p_{k-1}) \rightarrow p_k$ generated from one k-itemset, this rule would pass the pruning step if and only if $M[p_i][p_k]$ is greater than 0, $\forall\ i,\ 1\ \leq\ i\ <\ k$. That is, when the post-condition item is reachable from all the pre-condition items in the rule. As can be seen from Fig. 4, using the MRD information, some of the rules one are pruned out. For instance, the rule B → A is pruned since A is not reachable from B. In addition, using our approach of constructing the predictive model, the Web pages which are nonconsecutive to a current Web page are also considered, which we believe can better predict the user access patterns. For example, the predictive rule of A → H is included in our approach, although A is not consecutive to H.

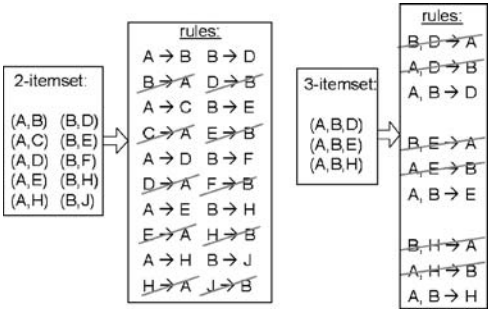


Fig. 4 The frequent itemsets and rule pruning process

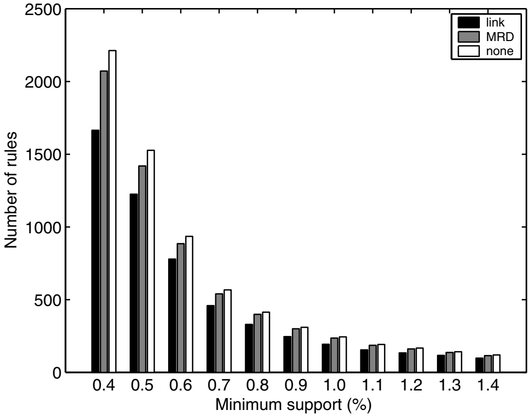


Fig. 5 Comparison of the numbers of rules for *link*, *MRD*, and *none* processes under various minimum support values

4 Experimental results and discussions

Several experiments on a real Web data set were conducted to evaluate the performance of the proposed approach. We applied our approach to approximate and construct the predictive model from the Web data set collected from the University of Miami's Web site. The link structure of the Web site was obtained by a crawling process to collect the hyperlinks embedded within the Web pages. The total number of Web pages with unique URLs is equal to 3,948. In the experiments, the user log records collected within the first week of April 2002 are used to construct the user access sequences. Once all the user access sequences are identified, two-third of the data set are separated as the training data set (with 34,362 user access sequences) and one-third are used for the test data set (with 17,182 user access sequences).

The first part of the experiments is to evaluate the actual reduction of the space complexity of the resulting rules with pruning (using either the link structure or the proposed MRD information) and without pruning. Figure 5 shows the number of rules generated by the association rule mining algorithm by varying the minimum support value. As shown in this figure, the result from the original association rule mining process without applying the pruning process is denoted by *none*, the pruning process using the link structure is denoted by *link*, and the pruning process using the MRD information is denoted by *MRD*, respectively.

The first observation is that by increasing the minimum support value for all different methods, the number of rules decreases at an exponential rate. This is because the minimum support value limits the number of itemsets, which are the basis for the rule construction. Another observation is that using the link structure for pruning reduces more rules than using the MRD information. The reason for this outcome is that in addition to browsing Web pages by following the hyperlinks (forward accesses), the users are also likely to access Web pages by the backward and jump accesses. A jump access occurs when the user retrieves a Web page without following a link, e.g., by using the search function on the Web site. Although, using the link structure could reduce more rules than using the MRD information, further analysis shows that using the link structure actually yields a worse performance than using the MRD information in terms of precision and recall. This implies that the pruning process using the link structure may over-prune the rule set and affect the performance. Whereas, using the MRD information, the precision and recall are not affected compared to the case of the original association rule mining technique.

The next experiment is to compare the performance of the following four different approaches:

1. Association rule mining using link structure information denoted by Association rule mining (link).
2. Markov model using link structure information denoted by Markov model (link).
3. Association rule mining using the MRD information denoted by Association rule mining (MRD).
4. Markov model using the MRD information denoted by Markov model (MRD).

To evaluate the performance of our approach, we measure the precision and recall values which are widely used in the information retrieval context [4, 19]. The precision measures the accuracy of the predictive rule set when applied to the testing data set; whereas the recall measures the coverage or the number of rules from the predictive rule set that match the incoming requests. Given a test access sequence $U = \{u_1, u_2, \ldots, u_n\}$, where n is the number of user's visited Web pages, and a list of predicted Web pages $V = \{v_1, v_2, \ldots, v_m\}$, where m is the number of predicted Web pages, the definitions of precision and recall for testing user access sequences are given as follows.

- Precision is the ratio of the number of Web pages correctly predicted over the total number of Web pages presented to the user.

$$precision = \frac{\mid U \cap V \mid}{m} \tag{1}$$

- Recall is the ratio of the number of Web pages correctly predicted over the total number of user's visited Web pages.

$$recall = \frac{\mid U \cap V \mid}{n} \tag{2}$$

Using the combination of the precision and recall, the F_1 measure which is the harmonic average of precision and recall can be defined as follows [19].

$$F_1 = \frac{2 \times (precision \times recall)}{precision + recall} \tag{3}$$

In addition to the F_1 measure, we also use the *rule coverage* which determines how well the predictive rules capture the user access sequences. To find the best performance under each approach, a graph under the averaged F1 measure multiplied by the coverage is plotted. The result is shown in Fig. 6. The optimum performance for both predictive algorithms using link structure occurs when the minimum support value is equal to 0.7%. Using the MRD information, the peak occurs at a minimum support of 0.9%.

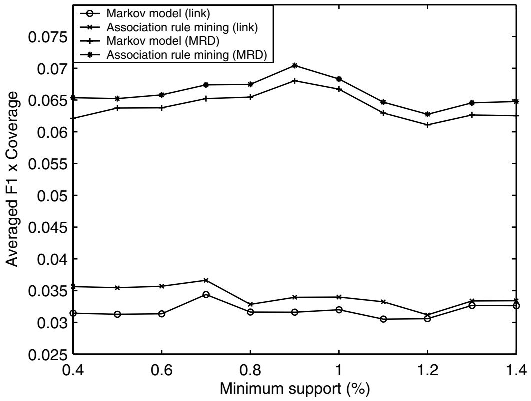


Fig. 6 Performance evaluation by (averaged F_1 × Coverage)

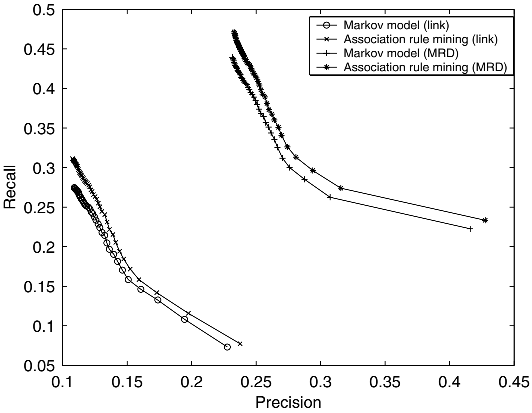


Fig. 7 Performance evaluation under precision and recall

Next, the precision and recall graph based on the optimum performance from each approach is plotted in Fig. 7. From this figure, it can be observed that for both approaches, using the link structure information to prune the results of association rule mining gives an inferior performance when compared to the MRD (user-based) information. The proposed approach based on the association rule mining yields a better performance than the existing Markov model approach.

Table 2 shows the summarized results of Fig. 7 under the averaged F1 measures. As shown in this table, the approach of association rule mining improves the performance of the averaged F1 value by 5.05% over the Markov model using the same MRD information. In addition, using the MRD information to prune the set of predictive rules, the averaged F1 value improves about twice as much as the hyperlink structure. This is because the MRD information is derived based on the user traversal constraint, therefore it can better capture the user access behavior on the Web site.

The performance using another different measure called averaged saving distance is also evaluated. Given a set of recommended Web pages, the averaged

Table 2 Averaged F1 values under four different approaches for mining user access patterns

Predictive model	Averaged F1 value
Association rule mining (MRD)	0.2911
Markov model (MRD)	0.2771
Association rule mining (link)	0.1559
Markov model (link)	0.1524

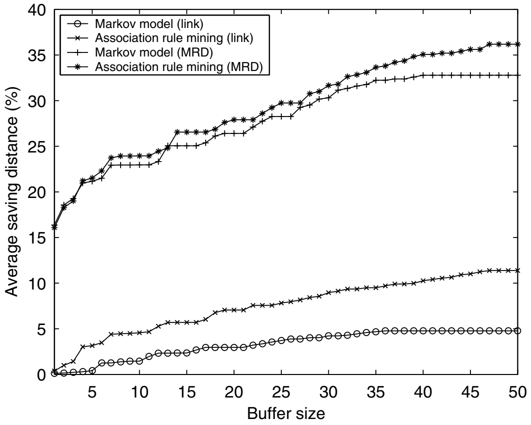


Fig. 8 Performance evaluation based on the averaged saving distance (%)

saving distance refers to the averaged number of *clicks* that a user can avoid from the actual browsing activity by directly selecting from the Web pages on the recommended list. The Web page recommended list is generated by ranking all the single items (i.e., Web pages) in the single-consequence rules obtained from the association rule mining algorithm. The ranking is based on the confidence values of those single-consequence rules.

For this performance comparison, the link structure information is needed in order to check the distance among Web pages. There are two assumptions for the calculation of the averaged saving distance. First, the saving distance is equal to zero for a Web page which has some direct hyperlinks to its descendant pages. This is based on the fact that users may directly select the hyperlinks without using the recommended list. All distances between Web pages can be looked up from the MRD information of the link structure. The second assumption is that the saving distance is set to 4 for the jump access, where the access does not follow the hyperlinks. This value is based on the empirical analysis of the averaged distance between a Web page to another one on the entire Web site. Figure 8 shows the performance evaluation based on the averaged saving distance. In this experiment, we varied the buffer size (the limited number of predicted Web pages presented to the user) from 1 to 50. As expected, the proposed approach yields a higher potential of browsing time reduction for the users. Another observation is that the averaged saving distance converges to a stable value after a certain buffer size is reached. This is due to the fact that once the allocation size is large enough to hold all the recommended Web pages, increasing the buffer size will not affect the performance.

5 Conclusion

In this paper, a new framework based on the association rule mining algorithm with traversal constraint to discover user access patterns is proposed. The proposed framework can be applied, for example, as a collaborative filtering technique in a recommender system. To capture the user access behavior, Web pages are modeled as a user-based graph by employing user access sequences instead of static hyperlinks. The shortest path algorithm in graph theory is applied on the user-based graph to find the Minimum Reaching Distance (MRD) between Web pages. The association rule mining technique is then applied to approximate and construct a predictive rule set. The MRD information, which imposes the traversal constraint, is used to prune the results from the association rule mining technique to reduce the space complexity of the rule set. The proposed approach improves the performance over the existing Markov model approach by allowing the prediction to include multiple nonconsecutive Web pages. Experiments on a real Web data set were conducted to demonstrate the effectiveness and efficiency of the proposed framework. Experimental comparisons were performed on (1) the number of rules generated, (2) the rule coverage percentage, (3) the precision and recall values, (4) the averaged F_1 value, and (5) the averaged saving distances. It can be concluded from the experimental results that the proposed framework outperforms the existing approaches in terms of both precision and recall, and furthermore there is a potential reduction of the user browsing time on the Web.

Acknowledgements For Mei-Ling Shyu, this research was support in part by NSF ITR (Medium) IIS-0325260. For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562 and NSF HRD-0317692.

References

1. Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD conference on management of data, Washington, D.C., pp 207–216
2. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the 11th international conference on data engineering, Taipei, Taiwan, pp 3–14
3. Anderson C, Domingos P, Weld D (2002) Relational Markov models and their application to adaptive web navigation. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, Canada, pp 143–152
4. Baeza-Yates R, Ribeiro-Neto B (eds) (1999) Modern information retrieval. ACM Press, Addison Wesley
5. Chen MS, Park JS, Yu PS (1998) Efficient data mining for path traversal patterns. IEEE Trans Knowl Data Eng 10(2):209–221
6. Cooley R, Mobasher B, Srivastava J (1999) Data preparation for mining world wide web browsing patterns. Knowl Inf Syst 1(1):5–32
7. Deshpande M, Karypis G (2001) Selective Markov models for predicting web-page accesses. In: Proceedings of the 1st SIAM international conference on data mining, Chicago, IL

8. Haruechaiyasak C, Shyu ML, Chen SC (2005) A web-page recommender system via a data mining framework and the semantic web concept. Int J Comput Applic Technol (in press)
9. Lin W, Alvarez S, Ruiz C (2002) Efficient adaptive-support association rule mining for recommender systems. Data Mining and Knowledge Dis 6(1):83–105
10. Mobasher B, Dai H, Luo T, Nakagawa M (2002a) Using sequential and nonsequential patterns for predictive web usage mining tasks. In: Proceedings of the IEEE international conference on data mining, Maebashi City, Japan
11. Mobasher B, Dai H, Tao M (2002b) Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining Knowl Dis 6(1):61–82
12. Padmanabhan VN, Mogul JC (1996) Using predictive prefetching to improve world wide web latency. ACM SIGCOMM Comput Commun Rev 26(3):22–36
13. Pitkow J, Pirolli P (1999) Mining longest repeating subsequences to predict world wide web surfing. In: Proceedings of the 2nd USENIX Symposium on internet technologies and systems, Boulder, CO, pp 139–150
14. Schechter S, Krishnan M, Smith MD (1998) Using path profiles to predict HTTP requests. Comput Networks ISDN Syst 30(1–7):457–467
15. Shyu ML, Haruechaiyasak C, Chen SC (2003) Category Cluster Discovery from Distributed WWW Directories. J Inf Sci (Special issue on knowledge discovery from distributed information sources) 155(3–4):181–197
16. Shyu ML, Chen SC, Haruechaiyasak C (2001) Mining user access behavior on the www. In: Proceedings of the IEEE international conference on systems, man, and cybernetics, Tucson, AZ, pp 1717–1722
17. Srivasta J, Cooley R, Deshpande M, Tan P (2000) Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor (1)2:12–23
18. Tan P, Kumar V (2002) Discovery of web robot sessions based on their navigational patterns. Data Mining Knowl Discov 6(1):9–35
19. Yang Y (1999) An evaluation of statistical approaches to text categorization. J Inf Retr 1(1/2):67–88

Mei-Ling Shyu received her Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN in 1999, and three Master's degrees from Computer Science, Electrical Engineering, and Restaurant, Hotel, Institutional, and Tourism Management from Purdue University. She has been an Associate Professor in the Department of Electrical and Computer Engineering (ECE) at the University of Miami (UM), Coral Gables, FL, since June 2005, Prior to that, she was an Assistant Professor in ECE at UM dating from January 2000. Her research interests include data mining, multimedia database systems, multimedia networking, database systems, and security. She has authored and co-authored more than 120 technical papers published in various prestigious journals, refereed conference/symposium/workshop proceedings, and book chapters. She is/was the guest editor of several journal special issues.

Choochart Haruechaiyasak received his Ph.D. degree from the Department of Electrical and Computer Engineering, University of Miami, in 2003 with the Outstanding Departmental Graduating Student award from the College of Engineering. After receiving his degree, he has joined the National Electronics and Computer Technology Center (NECTEC), located in Thailand Science Park, as a researcher in Information Research and Development Division (RDI). His current research interests include data/ text/ Web mining, Natural Language Processing, Information Retrieval, Search Engines, and Recommender Systems. He is currently leading a small group of researchers and programmer to develop an open-source search engine for Thai language. One of his objectives is to promote the use of data mining technology and other advanced applications in Information Technology in Thailand. He is also a visiting lecturer for Data Mining, Artificial Intelligence and Decision Support Systems courses in many universities in Thailand.

Shu-Ching Chen received his Ph.D. from the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA in December, 1998. He also received Master's degrees in Computer Science, Electrical Engineering, and Civil Engineering from Purdue University. He has been an Associate Professor in the School of Computing and Information Sciences (SCIS), Florida International University (FIU) since August, 2004. Prior to that, he was an Assistant Professor in SCIS at FIU dating from August, 1999. His main research interests include distributed multimedia database systems and multimedia data mining. Dr. Chen has authored and co-authored more than 140 research papers in journals, refereed conference/symposium/workshop proceedings, and book chapters. In 2005, he was awarded the IEEE Systems, Man, and Cybernetics Society's Outstanding Contribution Award. He was also awarded a University Outstanding Faculty Research Award from FIU in 2004, Outstanding Faculty Service Award from SCIS in 2004 and Outstanding Faculty Research Award from SCIS in 2002.