Collaborative Filtering by Mining Association Rules from User Access Sequences

Mei-Ling Shyu
Department of Electrical and Computer Engineering, University of Miami
Coral Gables, FL 33124, USA
shyu@miami.edu

Choochart Haruechaiyasak
Information Research and Development Division (RDI)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
choochart.haruechaiyasak@nectec.or.th

Shu-Ching Chen and Na Zhao
Distributed Multimedia Information System Laboratory, School of Computer Science
Florida International University, Miami, FL 33199, USA
{chens, nzhao002}@cs.fiu.edu

Abstract

Recent research in mining user access patterns for predicting Web page requests focuses only on consecutive sequential Web page accesses, i.e., pages which are accessed by following the hyperlinks. In this paper, we propose a new method for mining user access patterns that allows the prediction of multiple non-consecutive Web pages, i.e., any pages within the Web site. Our approach consists of two major steps. First, the shortest path algorithm in graph theory is applied to find the distances between Web pages. In order to capture user access behavior on the Web, the distances are derived from user access sequences, as opposed to static structural hyperlinks. We refer to these distances as Minimum Reaching Distance (MRD) information. The association rule mining (ARM) technique is then applied to form a set of predictive rules which are further refined and pruned by using the MRD information. The proposed approach is applied as a collaborative filtering technique to recommend Web pages within a Web site. Experimental results demonstrate that our approach improves performance over the existing Markov model approach in terms of precision and recall, and also has a better potential of reducing the user access time on the Web.

Keywords: *Association Rule Mining, Collaborative Filtering, Web Data Extraction, Web Log/Navigation Path Analysis.*

1. Introduction

Due to the increase in HyperText Transfer Protocol (HTTP) traffic on the World-Wide Web (WWW), the amount of transaction log records generated and collected on the servers grows tremendously. In order to benefit from these server log records, *data mining* has emerged as a tool to extract any useful patterns and analyze user access behavior on the Web. This specific type of data mining technique is known as *Web usage mining* [15]. In particular, 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 [11, 13], Web page recommendation [6, 8, 9], and Web personalization [10, 12].

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. Association rule mining has recently attracted considerable attention and proven to be a highly successful technique for extracting useful information from very large databases [1, 2, 7, 14]. For the problem of mining user access patterns, data sequences are typically user access sequences of Web pages. These access sequences are extracted from server log records via some Web data preparation techniques [5]. Applying a method for mining user access patterns on these access sequences reveals the user browsing behavior on the Web.

Various algorithms and techniques for mining user ac-

cess patterns have been proposed in the literature. In [11, 12], variations of the Markov model such as first-order Markov model and all K^{th}-order Markov model were applied to construct a predictive model to predict the user requests on Web pages. Their work mainly focused on the analysis of consecutive sequential access of Web pages, and hence given a currently visiting 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, $(p_1, p_2, \ldots, p_n)$, where p_i represents a Web page, an approximation of the first-order Markov model would contain the transitional probabilities of two adjacent Web pages in the user access sequence, $Pr(p_i \mid p_{i-1})$, where $1 < i \leq n$.

In this paper, we propose a new approach for mining user access patterns. The 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 within the Web site. To capture the user access behavior on the Web site, an alternative structure of the Web is constructed from user access sequences obtained from the server logs, as opposed to static structural hyperlinks.

Most of the previous research work focused on the forward and backward accesses, where forward accesses are those accesses that browse the Web pages by following hyperlinks embedded within the Web pages, and backward accesses are those that access the Web pages by backtracking to the previous Web pages. For example, in [4], the user access sequence was divided into smaller sequences called the *maximal forward references*, and the effect of backward references was not considered. However, considering only the smaller sub-sequences of the user access patterns does not fully capture the user's intention of accessing a particular set of Web pages, since some of the Web pages may be put into a different access sequence. Another type of accesses is the jump accesses, which the user retrieves a Web page by entering the Uniform Resource Locator (URL) directly on the Web browser. In this paper, all three access types are considered when the model is constructed. We pruned out the duplicate Web pages in the access sequences, since our goal is to predict the Web pages which the user has not yet visited. A user access sequence is used to represent a data record during the mining process.

Using this user traversal structure, a shortest path problem in the Graph Theory is applied to find the "access" distances between Web pages. We refer to these distances as Minimum Reaching Distance (MRD) information. The ARM technique is then applied to find a set of predictive rules that pass the user-specified minimum support. The MRD information is used to prune the results from ARM in order to increase the prediction accuracy and reduce the space complexity. The proposed method for mining user ac-

cess sequences was applied as a collaborative filtering technique. The results from the process of mining user access patterns are a predictive rule set that is used to recommend Web pages according to the users who accessed the Web site in the past [8, 9].

Under the Markov model notion, our method can be viewed as the All K^{th}-Order Markov model with the look-ahead ability, which allows the prediction to include multiple non-consecutive Web pages, i.e., any Web pages within the Web site which are not necessarily connected by hyperlinks. For example, the approximation of the first-order Markov model with the look-ahead ability would contain the following transitional probabilities of two Web pages including non-consecutive ones, $Pr(p_j \mid p_i)$, where $1 \leq i < n$ and $i < j \leq n$.

The remainder of this paper is organized as follows. In Section 2, the method of mining user access patterns based on association rule mining is explained in details. The experiments on a real Web data set are given in Section 3. In the same section, the experimental results are presented and analyzed. Conclusion is given in Section 4.

2. Association Rule Mining for User Access Patterns

We propose to apply the association rule mining (ARM) technique [1, 2] in mining user access patterns on the Web pages. The objective is to construct a model that predicts users' Web page requests to assist in browsing the Web pages and to reduce the access time.

In ARM, a *k-itemset* contains k items that pass the user-specified minimum *support* value. Hence, in essence, ARM enables one to discover interesting patterns or associations among the *k-itemset* in a given collection of records. Our framework applies the ARM 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 ARM is large. In the original algorithm, the minimum *confidence* value is used to prune the rules; while in our proposed framework, the number of rules is pruned by incorporating the MRD information, which reduces the state complexity of the model.

Our MRD calculation adopts the concept from the *shortest-path* problem in the Graph Theory. The original algorithm assumes that the traversal path follows the link structure of the graph, where a link structure is a representation of Web pages along with the embedded hyperlinks. However, the pages that a user accesses do not always follow the link structure of the Web pages, and hence an alternative Web representation based on the actual user browsing activity on the Web site needs to be constructed. Here, a user access sequence (also referred to as a browsing sequence or a traversal path) is
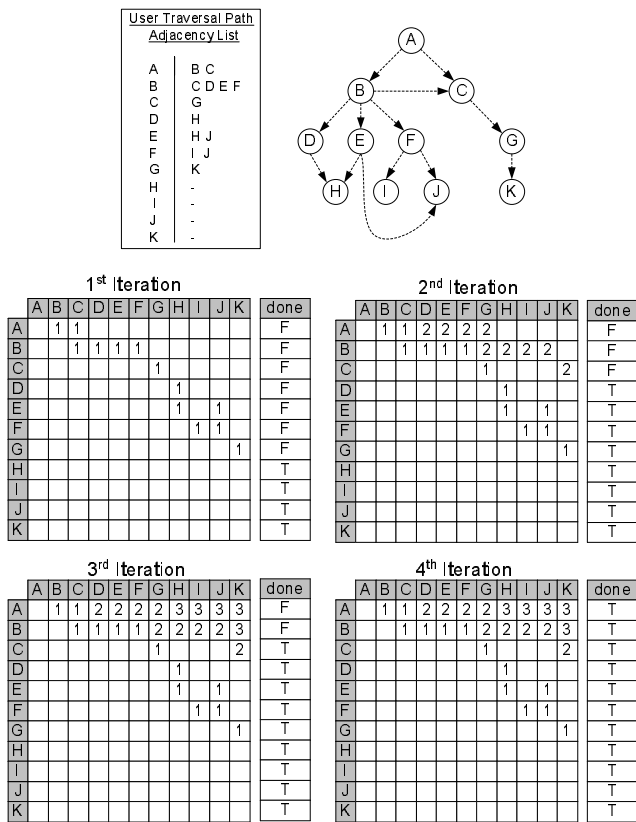
Figure 1. Minimum Reaching Distance (MRD) Construction

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

Figure 2. Frequent itemsets and rule pruning

2-itemset: (A,B) (B,D) (A,C) (B,E) (A,D) (B,F) (A,E) (B,H) (A,H) (B,J)

rules:
A→B   B→D
B→A   D→B
A→C   B→E
C→A   E→B
A→D   B→F
D→A   F→B
A→E   B→H
E→A   H→B
A→H   B→J
H→A   J→B

3-itemset: (A,B,D) (A,B,E) (A,B,H)

rules:
B,D→A
A,D→B
A,B→D

B,E→A
A,E→B
A,B→E

B,H→A
A,H→B
A,B→H

is blank otherwise. The values of the $done[p_i]$ column are false (F) if row p_i contains at least one direct link from the user-based graph. In the second iteration, MRD finds the possible reaching distance from the Web page whose value in the *done* array is false. For example, considering 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. Please note that the distance between A and C remains **1** since C is reachable from A in one step, and the MRD algorithm is to keep the minimum reaching distance. For each iteration, the element $done[p_i]$ is set to true (T) if there is no change of the value in row p_i in an iteration. The algorithm terminates when all the elements in array *done* are true. In this example, it terminates in **4** iterations.

Once the MRD information for the Web pages are constructed, they are used in the rule pruning process. Assume that Figure 2 shows all the resulting 2-itemsets and 3-itemsets that passed a pre-specified minimum support value. 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 3-itemset (A, B, D): (B, D) → A, (A, D) → B, and (A, B) → D.

Next, the MRD information from Figure 1 is used to prune the resulting rules as follows. Consider a single-consequent rule of the form $(p_1, p_2, \ldots, p_{n-1}) \rightarrow p_n$, this rule would pass the pruning step if and only if $M[p_i][p_n]$ is greater than 0, $\forall\, i,\ 1 \le i < n$. That is, when the post-condition item is reachable from all the pre-condition items in the rule. As seen from Figure 2, using the MRD information, some of the rules 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 non-consecutive to a current Web page are also considered, which we believe can

an ordered list of Web pages accessed by a user during one session. User access sequences are extracted from the Web server log records as part of the data preprocessing step. The issues and analysis of the user access sequence are considered in many research works including [3, 5].

For example, the top of Figure 1 shows the result of constructing a graph based on the user access sequences. This representation allows us to capture some paths (such as E → J) which do not previously exist in the link structure graph. In addition, some of the links that exist in the link structure do not appear in the user-based graph 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. The bottom of Figure 1 also shows the iterations of constructing the MRD information using the user-based graph. As shown in this figure, in the first iteration, the algorithm first translates the user-based graph into an adjacency matrix structure M, where the value in the position $M[p_i][p_j]$ is **1** if p_j has a direct link from p_i in the user-based graph, and
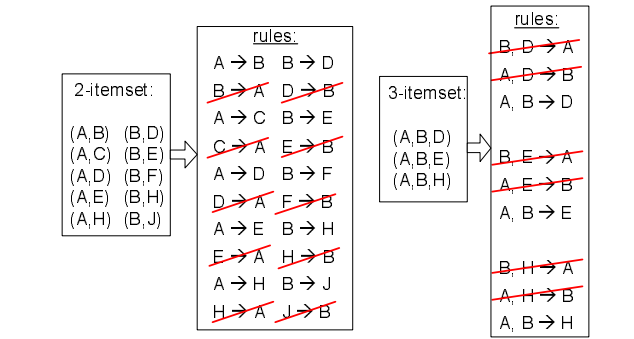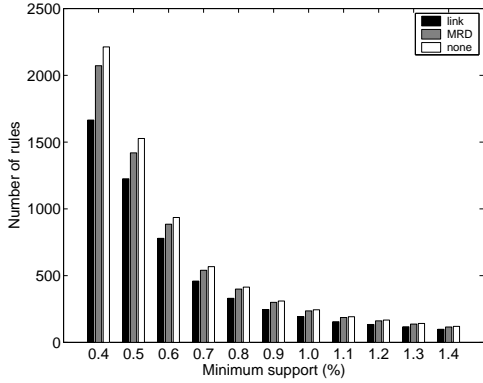
Figure 3. Number of rules comparison for three approaches: (1) without a pruning process (none), (2) via the MRD information (MRD), and (3) via the link structure (link)

Figure 4. Performance evaluation under precision and recall

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.

3. Experimental Results

Experiments on a real Web data set from University of Miami are conducted to evaluate the performance of our proposed framework. A crawling program was developed to collect the hyperlinks embedded within the Web pages for the link structure. The total number of Web pages with unique URLs is equal to *3,948*. The user log records are used to construct the user access sequences. Once all the user access sequences are identified, two-third of the data set with *34,362* user access sequences is used as the training data set; while one-third of it have *17,182* user access sequences and are used as the test data set.

Figure 3 shows the reduction comparison of the resulting rules by using our MRD-based pruning approach (*MRD*), the original association rule mining without a pruning process (*none*), and with the pruning process using the link structure (*link*). 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 the pruning process reduces more rules than using the MRD information. The reason for this outcome is that in addition to browse Web pages by the forward accesses, the users are also likely to access Web pages by backward accesses and jump accesses.

Although, using the link structure could reduce more rules than using the MRD information, our further analysis
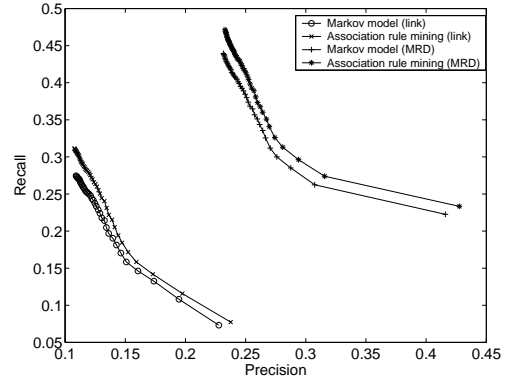
shows that using the link structure actually yields a worse performance than using the MRD information in terms of precision and recall. This implies that the pruning process using the link structure may over-prune the rule set and affect the performance. Whereas under the MRD information, the precision and recall are not affected compared to the case of the original association rule mining.

Next, we evaluate the performance based on the precision and recall. Assume we have 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 precision measures the accuracy of the predictive rule set when applied to the testing data set. It is defined as the ratio of the number of Web pages correctly predicted over the total number of Web pages presented to the user. That is, $precision = \frac{|U \cap V|}{m}$. On the other hand, the recall measures the coverage or the number of rules from the predictive rule set that match the incoming requests. It is defined as the ratio of the number of Web pages correctly predicted over the total number of user's visited Web pages. That is, $recall = \frac{|U \cap V|}{n}$.

In this experiment, the precision and recall values are compared for the following four different approaches:

1. association rule mining using link structure information (association rule mining (link));

2. Markov model using link structure information (Markov model (link));

3. association rule mining using the MRD information (association rule mining (MRD)); and

4. Markov model using the MRD information (Markov model (MRD)).

The buffer size (the limited number of predicted Web pages presented to the user) varies from 1 to 50. The precision and recall graph based on the support value of 0.8% is

Predictive Model	Averaged F_1 Value
association rule mining (MRD)	0.2911
Markov model (MRD)	0.2771
association rule mining (link)	0.1559
Markov model (link)	0.1524

Table 1. Averaged F_1 measures under four different approaches for mining user access patterns

shown in Figure 4. From this figure, it can be observed that, for both association rule mining and the Markov model approaches, using the link structure information to prune the results of the association rule mining gives a worse performance compared to those using the MRD (user-based) information. In addition, our proposed approach (association rule mining (MRD)) improves the performance over the existing Markov model approach. Using the combination of the precision and recall, the F_1 measure which is the harmonic average of precision and recall is defined in Equation 1.

$$F_1 = \frac{2 \times (precision \times recall)}{precision + recall} \qquad (1)$$

Table 1 shows the summarized results of Figure 4 under the F_1 measures. As shown in this table, the approach of association rule mining improves the performance under F_1 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 F_1 value improves about twice as much as of the hyperlink structure. This is since the MRD information is derived based on the user traversal constraint, and therefore it can better capture the user access behavior on the Web site.

4. Conclusion

In this paper, the problem of mining user access patterns is considered. A new method is proposed based on the association rule mining and the shortest path algorithm in graph theory. To capture the user access behavior, we model the Web by using user access sequences instead of static hyperlinks. The association rule mining technique is then applied to approximate and construct the predictive model. The proposed Minimum Reaching Distance (MRD) information is used to prune the results from the association rule mining to reduce the state-space complexity of the model. The proposed approach improves the performance over the existing Markov model approach by allowing the prediction to include multiple non-consecutive Web pages. To demonstrate a potential usage, we applied the proposed approach for the collaborative filtering technique. Experimental results using a real Web data set show that our approach improves performance over the existing approaches in terms of both preci-

sion and recall, and also has a better potential of reducing the user browsing time on the Web.

References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *Proc. of ACM SIGMOD Conf. on Management of Data*, 1993, pp. 207-216.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proc. of the Eleventh Int. Conf. on Data Engineering*, 1995, pp. 3-14.

[3] P. Berkhin, J. D. Becher, and D. J. Randall, "Interactive Path Analysis of Web Site Traffic," *Proc. of the Seventh ACM SIGKDD Int. Conf. in Knowledge Discovery and Data Mining*, 2001, pp. 414-419.

[4] M.-S. Chen, J. S. Park, and P. S. Yu, "Efficient Data Mining for Path Traversal Patterns," *IEEE Trans. on Knowledge and Data Engineering*, 10(2), 1998, pp. 209-221.

[5] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," *Knowledge and Information Systems*, 1(1) 1999, pp.5-32.

[6] J. Dean and M. R. Henzinger, "Finding Related Pages in the World Wide Web," *Proc. of the Eighth Int. World Wide Web Conf.*, 1999, pp. 389-401.

[7] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[8] C. Haruechaiyasak, M.-L. Shyu, and S.-C. Chen, "A Web-Page Recommender System via a Data Mining Framework and the Semantic Web Concept," accepted for publication, *International Journal of Computer Applications in Technology*.

[9] C. Haruechaiyasak, M.-L. Shyu, and S.-C. Chen, "A Data Mining Framework for Building A Web-Page Recommender System," *2004 IEEE International Conference on Information Reuse and Integration (IRI'2004)*, Las Vegas, Nevada, USA, November 8-10, 2004, pp. 357-262.

[10] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," *Communications of the ACM*, 43(8), 2000, pp. 142-151.

[11] V. N. Padmanabhan and J. C. Mogul, "Using Predictive Prefetching to Improve World Wide Web Latency," *ACM SIGCOMM Computer Communications Review*, 26(3), 1996, pp. 22-36.

[12] J. Pitkow and P. Pirolli, "Mining Longest Repeating Subsequences to Predict World Wide Web Surfing," *Proc. of the Second USENIX Symposium on Internet Technologies and Systems*, 1999, pp. 139-150.

[13] S. Schechter, M. Krishnan, and M. D. Smith, "Using Path Profiles to Predict HTTP Requests," *Computer Networks and ISDN Systems*, 30(1-7), 1998, pp. 457-467.

[14] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "A Generalized Affinity-Based Association Rule Mining for Multimedia Database Queries," *Knowledge and Information Systems (KAIS): An International Journal*, vol. 3, no. 3, August 2001, pp. 319-337.

[15] J. Srivasta, R. Cooley, M. Deshpande, and P. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKKD Explorations*, (1)2, 2000, pp. 12-23.