A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses

Faten Khalil, Jiuyong Li and Hua Wang

Department of Mathematics & Computing
University of Southern Queensland
Toowoomba, Australia, 4350,
Email: {khalil,jiuyong and wang}@usq.edu.au

Abstract

The importance of predicting Web users' behaviour and their next movement has been recognised and discussed by many researchers lately. Association rules and Markov models are the most commonly used approaches for this type of prediction. Association rules tend to generate many rules, which result in contradictory predictions for a user session. Low order Markov models do not use enough user browsing history and therefore, lack accuracy, whereas, high order Markov models incur high state space complexity. This paper proposes a novel approach that integrates both association rules and low order Markov models in order to achieve higher accuracy with low state space complexity. A low order Markov model provides high coverage with low state space complexity, and association rules help achieve better accuracy.

Keywords: Association rules, Markov models, prediction.

1 Introduction

The need to predict the next Web page to be accessed by the user is apparent in most Web applications today whether they are search engines or e-commerce solutions or mere marketing sites. Web applications today are driven to provide a more personalized experience for their users. Therefore, it is extremely important to form some kind of interaction with Web users and always be one step ahead of them when it comes to predicting next accessed pages. For instance, knowing the user browsing history on the site grants us valuable information as to which one of the most frequently accessed pages will be accessed next. Also, it provides us with extra information like the type of user we are dealing with and the users preferences as well. There are various ways that can help us make such a prediction, but the most common approaches are Markov models and association rules. Each of the approaches used for this purpose has its own weaknesses when it comes to accuracy, coverage and performance. Lower order Markov models lack accuracy because of the limitation in covering enough browsing history; whereas higher order Markov models usually result in higher state space complexity. On the other hand, association rules have the problem of identifying the one correct prediction out of the many rules that lead to a large number of predictions (Mobasher, Dai, Luo & Nakagawa 2001, Yang, Li, & Wang 2004). This paper proposes an improved approach, based on a combination of Markov models and association rules that results in better prediction accuracy and more coverage. We use low order Markov models to predict multiple pages to be visited by a user and then we apply association rules to predict the next page to be accessed by the user based on long history data.

1.1 Related Work

The importance of Web usage mining has led to a number of research papers in the area. However, most of these papers were hindered by some kind of limitations. For instance, many of the papers proposed using association rules or Markov models for next page prediction, however, none of these papers have addressed the use of a combination of both methodologies. Some of the papers that proposed the use of association rules for better predicting the next page to be accessed by the user are (Mobasher et al. 2001, Spiliopoulou, Faulstich & Winkler 1999, Yong, Zhanhuai & Yang 2005); whereas, other papers like (Cadez, Heckerman, Meek, Smyth & White 2000, Deshpande & Karypis 2004, Dongshan & Junyi 2002, Garafalakis, Pastogi, Seshadri & Shim 1999, Gunduz & Ozsu 2003, Jespersen, Pedersen & Thorhauge 2003) covered Markov models.

Mobasher *et al.* (2001) were confronted with the problem of providing user personalisation at an early stage of the Web session. They proposed the use of collaborative filtering approaches like the k-Nearest Neighbour (KNN) approach. However, some problems were identified like scalability and efficiency. KNN requires that neighbourhood identification be performed online. This is not feasible most of the time because of the large amount of data. Another problem is the effectiveness in terms of coverage and precision. Low coverage is caused by larger user histories and low precision is due to the sparsity of Web data. The authors then proposed a solution that gives better results than the KNN approach in terms of scalability and effectiveness. They recommended an approach that uses association rules techniques that are based on storing the most frequent items used in a data structure and using an algorithm to identify the most suitable items to be used with online recommendations. The main problem associated with association rules in general is scalability due to the large number of itemsets. However, when the authors proposed a method that includes increasing the window size, it caused scalability problems as well as lower

This project was partially supported by Australian Research Council Discovery Grant DP0559090.

coverage. On the other hand, using multiple support thresholds resulted in better coverage but it did not improve on accuracy.

Yang *et al.* (2004) have studied five different representations of association rules which are: Subset rules, Subsequence rules, Latest subsequence rules, Substring rules and Latest substring rules. As a result of the experiments performed by the authors concerning the precision of these five association rules representations using different selection methods, the latest substring rules were proven to have the highest precision with decreased number of rules.

On the other hand, Yong *et al.* (2005) explored sequential association rules further and they proposed a new sequential association rule model for Web document prediction based on the comparison of different types of sequential association rules according to sequence constrains and temporal constrains. They proved through means of experimentation that both sequence constrains and temporal constrains affect the precision of Web document prediction and that temporal constrains have more influence than sequence constrains.

Numerous papers dealt with the topic of Markov Model as a method to solve the prediction problem with higher coverage, better accuracy and performance than association rules. For example, Deshpande *et al.* (2004) addressed the reduced accuracy problem of the low-order Markov Models. They proposed an all-kth order model instead. They solved the state space complexity problem of the all-kth order model by pruning some of the states according to frequency, confidence and error representations. This proposed solution to the state space complexity of the all-kth order model may not be feasible in some instances, especially when it comes to very large data sets. It requires a lot of time and effort to build the all-kth order models and prune the pages according to the three criteria. It also involves a great deal of calculations (different types of thresholds for different pruning methods.)

Dongshan *et al.* (2002) proposed the use of a hybrid-order tree like Markov Model (HTMM) in order to solve the problems associated with traditional Markov Models especially the state space complexity and low coverage. They identified the suitability of HTMM with predicting the next pages to be accessed by the user and caching such pages in order to improve Web pre-fetching. HTMM combines two methods: a tree-like Markov model method and a hybrid order method. The k-order Tree-like Markov model is a tree constructed using a sequence of visited Web pages accessed by the user. Each node of the tree conforms to a visited page URL and a count that records the number of times the page was visited. The height of the tree is k + 2 where k is the order of the Markov model and the width of the tree is no more than the number of sequences of the visited pages. The tree-like Markov model results in low coverage that results in low accuracy. As a solution, the authors proposed training varying order Markov models and combining those models together for prediction. They used two methods for combining the models: accuracy voting and blending. To evaluate the results of these methods, the authors used Web server log files of an educational site and after cleaning and preprocessing the log data, they came up with the following results: When it comes to precision and accuracy, both HTMM methods showed better results than traditional Markov models. Also, when it comes to time associated with building the models and giving prediction, the HTMM methods showed better results than traditional Markov models. However, with prediction time, HTMM methods and traditional methods showed similar results. These results are apparent with HTMM in general. However, when it comes to building the tree, it is based on all-kth order model and it has the same complexity as the all-kth order model. This places a great limitation on the approach as a whole.

Related work was presented by Gunduz *et al.* (2003)where they proposed a new model that takes into consideration the time spent on the page as well as the sequence of visiting pages in a Web session. First, pages are clustered according to their similarities. Then, a click-stream tree is used to generate recommendations. This approach is rather complicated and data has to go in various stages before prediction takes place. Other researchers that went through similar work, like Mobasher *et al.* (2000) and Sarukkai (2000), did not take Web data complexity into consideration. Of course, more complex data would lead to higher storage space and runtime overhead.

Kim *et al.* (2004) presented a combination of association rules, Markov models, sequential association rules and clustering. This paper presented the use of four Web personalisation models in order to improve on their performance especially when it comes to precision and recall. The authors argued that both association rules and sequential association rules techniques can use All-Kth order model to increase coverage but this produced less precision.

1.2 Organisation of the Paper

This paper is organised as follows. In section 2, we cover Web access prediction using Markov model and association rules. In section 3, we introduce our proposed solution. In section 4 we analyse the data and produce the experiments results. Section 5 concludes our work.

2 Related Technologies

2.1 Markov Model

Markov models are becoming very commonly used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages (Deshpande et al. 2004).

Let $P = \{p1, p2, \ldots, pm\}$ be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages, then prob$(pi|W)$ is the probability that the user visits pages pi next. Page p_{l+1} the user will visit next is estimated by:

$$\begin{aligned} P_{l+1} &= \text{argmax}_{p\in\mathbb{P}}\{P(P_{l+1} = p|W)\} \\ &= \text{argmax}_{p\in\mathbb{P}}\{P(P_{l+1} = p|p_l, p_{l-1}, \ldots, p_1)\} \quad (1) \end{aligned}$$

This probability, $prob(pi|W)$, is estimated by using all W sequences of all users in history (or training data), denoted by W. Naturally, the longer l and the larger W, the more accurate $prob(pi|W)$. However, it is infeasible to have very long l and large W and it leads to unnecessary complexity. Therefore, to overcome this problem, a more feasible probability is estimated by assuming that the sequence of the Web pages visited by users follows a Markov process. The Markov process imposed a limit on the number of previously accessed pages k. In other words, the probability of visiting a page pi does not depend on all the pages in the Web session, but only on a small set of k preceding pages, where $k << l$.

The equation becomes:

$$P_{l+1} = \text{argmax}_{p\in\mathbb{P}}\{P(P_{l+1} = p|p_l, p_{l-1}, \ldots, p_{l-(k-1)})\} \quad (2)$$

where k denotes the number of the preceding pages and it identifies the order of the Markov model. The resulting model of this equation is called the Kth-Order Markov model. Of course, the Markov model starts calculating the highest probability of the last page visited because during a Web session, the user can only link the page he is currently visiting to the next one. The example is similar to Desphpandes Figure 1 (Deshpande et al. 2004):

Let S_j^k be a state containing k pages, $S_j^k = \langle p_{l-(k-1)}, p_{l-(k-2)}, \ldots, p_l \rangle$. The probability of $P\left(p_i | S_j^k\right)$ is estimated as follows from a history (training) data set.

$$P\left(p_i | S_j^k\right) = \frac{\text{Frequency}\left(\langle S_j^k, p_i \rangle\right)}{\text{Frequency}\left(S_j^k\right)} \ . \qquad (3)$$

This formula calculates the conditional probability as the ratio of the frequency of the sequence occurring in the training set to the frequency of the page occurring directly after the sequence.

The fundamental assumption of predictions based on Markov models is that the next state is dependent on the previous k states. The longer the k is, the more accurate the predictions are. However, longer k causes the following two problems: The coverage of model is limited and leaves many states uncovered; and the complexity of the model becomes unmanageable. Therefore, the following are three modified Markov models for Predicting Web page access.

1. All kth Markov model: This model is to tackle the problem of low coverage of a high order Markov model. For each test instance, the highest order Markov model that covers the instance is used to predict the instance. For example, if we build an all 4- Markov model including 1-, 2-, 3-, and 4-, for a test instance, we try to use 4-Markov model to make prediction. If the 4-markov model does not contain the corresponding states, we then use the 3-markov model, and so forth (Pitkow & Pirolli 1999).

2. Frequency pruned Markov model: Though all-kth order Markov models result in low coverage, they exacerbate the problem of complexity since the states of all Markov models are added up. Note that many states have low statistically predictive reliability since their occurrence frequencies are very low. The removal of these low frequency states affects the accuracy of a Markov model. However, the number of states of the pruned Markov model will be significantly reduced.

3. Accuracy pruned Markov model: Frequency pruned Markov model does not capture factors that affect the accuracy of states. A high frequent state may not present accurate prediction. When we use a means to estimate the predictive accuracy of states, states with low predictive accuracy can be eliminated. One way to estimate the predictive accuracy using conditional probability is called confidence pruning. Another way to estimate the predictive accuracy is to count (estimated) errors involved, called error pruning.

2.2 Association Rules

Association rule mining is a major pattern discovery technique as proved by Mobasher *et al.* (2000). The original goal of association rule mining is to solve market basket problem. For a data set containing shopping transactions, association rules summarise relationships illustrated by the following example. Customers who buy bread and milk will most likely buy eggs, or, bread and milk → eggs. Association rules are mainly defined by two metrics: support and confidence. The applications of association rules are far beyond market basket applications. Let us look at how association rules are used in Web data mining.

Let $P = \{p_1, p_2, , p_m\}$ be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit, and D includes a collection of user sessions. Let A be a subsequence of W, and p_i be a page. We say that W supports A if A is a subsequence of W, and W supports $\langle A, p_i \rangle$ if $\langle A, p_i \rangle$ is a subsequence of W. The support for sequence A is the fraction of sessions supporting A in D, denoted by $\text{supp}(A)$. An implication is $A \rightarrow p_i$. The support of implication $A \rightarrow p_i$ is $\text{supp}(\langle A, p_i \rangle)$, and the confidence of the implication is $\text{supp}(\langle A, P \rangle)/\text{supp}(A)$, denoted by $\text{conf}(A \rightarrow p_i)$. When we use the same terminologies of Markov model, $\text{supp}(\langle A, p_i \rangle) = \text{prob}(\langle A, p_i \rangle)$, and confidence $(A, p_i) = \text{prob}(p_i | A)$. An implication is called an association rule if its support and confidence are not less than some user specified minimum thresholds.

The minimum support requirement dictates the efficiency of association rule mining. One major motivation for using the support factor comes from the fact that we are usually interested only in rules with certain popularity. Support corresponds to statistical significance, and confidence is a measure of the rules strength.

There are four types of sequential association rules presented by Yang *et al.* (2004):

1. Subsequence rules: they represent the sequential association rules where the items are listed in order.

2. Latest subsequence rules: They take into consideration the order of the items and most recent items in the set.

3. Substring rules: They take into consideration the order and the adjacency of the items.

4. Latest substring rules: They take into consideration the order of the items, the most recent items in the set as well as the adjacency of the items.

The immense number of generated rules gives rise to the need of some predictive models that reduce the rule numbers and increase their quality by weeding out the rules that were never applied. Yang *et al.* (2004), introduced the following predictive models:

1. Longest match: This method assumes that longer browsing paths produce higher quality information about the user access pattern. Therefore, in the case where we have more than one rule, all with support above a certain threshold and they match an observed sequence, the rule with the longest length will be chosen for predication purposes and the rest of the rules will be disregarded.

2. Most-confidence matching: This is a very common method where the rule with the highest confidence is chosen amongst the rest of all the applicable rules whose support values are above a certain threshold.

3. Least error matching: This is a method to combine support and confidence, based on the observed error rate and the support of each rule, to form a unified selection measure and to avoid the need to set a minimum support value artificially. The observed error rate is calculated by dividing

the number of incorrect predictions by the number of training instances that support it. The rule with the least error rate is chosen amongst all the other applicable rules.

From a previous study (Yang et al. 2004), the latest substring with the least error matching produces the most accurate models for Web document prediction. In this paper, we will use sequential association rule mining on user transaction data to discover Web page usage patterns. Prediction of the next page to be accessed by the user is performed by matching the discovered patterns against the user sessions. This is usually done online.

3 A framework for integration

The main problem associated with association rules that apply to large data item sets is the discovery of large number of rules and the difficulty in identifying the one rule that leads to the correct prediction. In regards to Markov models, low order Markov models lack web page prediction accuracy because they do not use enough history and high order Markov models have high state space complexity.

There is apparent a direct relationship between Markov models and association rules techniques. According to the Markov model pruning methods presented by Mobasher *et al.* (2004) and association rules selection methods presented by Yang *et al.* (2004), there exists a great resemblance between the two. The substring association rules with most confidence prediction model form a frequency pruned all kth order Markov model, where k is the number of maximum items in the association rules. They also share similar problems. For instance, the number of states (rules) becomes unmanageable when k is large. In contrast, short history is not enough for making accurate predictions.

We propose to use low order all kth Markov models to keep low state complexity and high coverage. The accuracy of low order Markov models is normally not satisfactory. For those Markov states that provide ambiguous predictions, we make use of association rules to sample long history. Association rules help those states to make more accurate predictions. Association rules are complicated as well, but we only use rules to complement Markov states that provide ambiguous predictions. Therefore, this does not add too much complexity to the system. We use the following example to show the idea of the integration.

Consider the set of Web page structure for an online computer shop in Figure 1.

Note that letters are assigned to nodes names in Figure 1 for simplicity purposes. Table 1 examines the following 6 user sessions:

Table 1: User sessions

T1	A,C,G,A,D,H,M,C,F,C,G,R,I,P,H,O,J
T2	A,G,T,A,C,S,G,J,R,A,D,H,M,D,J
T3	A,F,I,B,A,E,D,H,N,P,I,Q,F,J,D,H,N,G,C
T4	A,I,J,B,A,E,C,T,D,H,M,I,Q,G
T5	F,D,H,N,J,A,D,A,E,D,J,R,H,N,G,C,F,G
T6	F,L,S,D,H,N,J,Q,E,I,P,C,I,O,A,D,H,M

Calculating the frequencies of accessed pages, Table 2 lists the pageviews with their frequencies.

A 100% support results in a very large number of rules and is rather cumbersome. Therefore, assuming that the minimum support is 4; B, K, L, O, P, Q, R, S and T are removed from the itemsets. Table 3 lists the user sessions that pass the frequency and support tests.

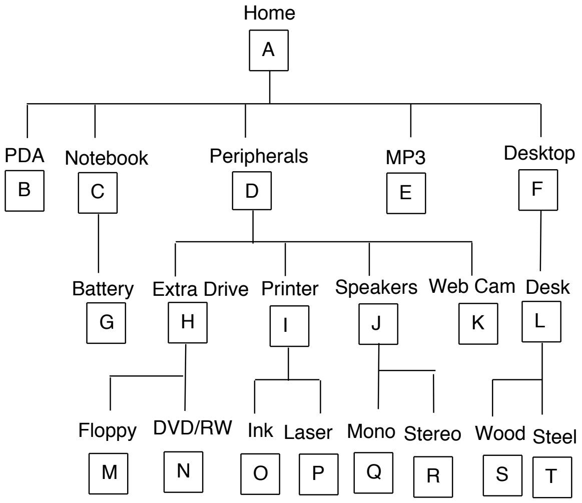

Figure 1: Online computer store Web page structure.

Table 2: Pageviews frequencies

Page	A	B	C	D	E	F	G	H	I	J
Freq	12	2	8	11	4	6	8	10	7	8

Page	K	L	M	N	O	P	Q	R	S	T
Freq	0	1	4	4	3	3	3	3	2	2

Table 3: User sessions after frequency and support pruning

T1	A,C,G,A,D,H,M,C,F,C,G,R,I,H,J
T2	A,G,A,C,G,J,A,D,H,M,D,J
T3	A,F,I,A,E,D,H,N,I,F,J,D,H,N,G,C
T4	A,I,J,A,E,C,D,H,M,I,G
T5	F,D,H,N,J,A,D,A,E,D,J,H,N,G,C,F,G
T6	F,D,H,N,J,E,I,C,I,A,D,H,M

Applying the 2^{nd} order Markov Model to the above training user sessions we notice that the most frequent state is $\langle D, H\rangle$ and it appeared 8 times as follows:

$$P_{l+1} = \operatorname{argmax}\{P(M|H, D)\} = M \ \text{ OR } \ N$$

Obviously, this information alone does not provide us with correct prediction of the next page to be accessed by the user as we have high frequencies for both pages, M and N. To break the tie and find out which page would lead to the most accurate prediction, we have to look at previous pages in history. This is where we use subsequence association rules as it shows in Table 4 below.

Table 4: User sessions history

A, C, G, A,	$\langle D, H\rangle$	M
A, G, A, C, G, J, A,	$\langle D, H\rangle$	M
A, F, I, A, E,	$\langle D, H\rangle$	N
I, F, J,	$\langle D, H\rangle$	N
A, I, J, A, E, C,	$\langle D, H\rangle$	M
F,	$\langle D, H\rangle$	N
F,	$\langle D, H\rangle$	N
J, E, I, C, I, A,	$\langle D, H\rangle$	M

Tables 5 and 6 summarise the results of applying subsequence association rules to the training data. Table 5 shows that C → M has the highest confidence of 100%. While Table 6 shows that F → N has the highest confidence of 100%.

Table 5: Confidence of accessing page M using subsequence association rules

A → M	AM/A	4/10	40%
C → M	CM/C	4/4	100%
E → M	EM/E	2/3	67%
F → M	FM/F	0/4	0%
G → M	GM/G	2/3	67%
I → M	IM/I	2/5	40%
J → M	JM/J	3/4	67%

Table 6: Confidence of accessing page N using subsequence association rules

A → N	AN/A	1/10	10%
C → N	CN/C	0/4	0%
E → N	EN/E	1/3	33%
F → N	FN/F	4/4	100%
G → N	GN/G	0/3	0%
I → N	IN/I	2/5	40%
J → N	JN/J	1/4	25%

Using Markov models, we can determine that there is a 50/50 chance that the next page to be accessed by the user after accessing the pages D and H could be either M or N. Whereas subsequence association rules take this result a step further by determining that if the user accesses page C before pages D and H, then there is a 100% confidence that the user will access page M next. Whereas, if the user visits page F before visiting pages D and H, then there is a 100% confidence that the user will access page N next.

Applying this result back to our example, we find that if the user buys a notebook, there is more chance that he/she will buy an external floppy drive. However, if the user buys a desktop, there is more chance that he/she will buy an extra DVD/RW drive. This extra bit of information is very important as knowing user browsing history gives us an added advantage of knowing the browsing habits of our users.

In this paper, we introduced the Integrated Markov and Association Model (IMAM) that inputs a database(D) and a session (s) and outputs the next page(p) that will be accessed by the user with high prediction. IMAM is summarised as follows:

Training:

```
Build a low order Markov model
FOR each state of the Markov model
    IF the prediction is ambiguous
        THEN
        Collect all sessions satisfying
            the state
        Construct association rules to
            resolve ambiguity
        Store the association rules with
            the state
    ENDIF
ENDFOR
```

Test:

```
Find a matching state of the Markov model
      for a test session
IF the matching state provides an non-
      ambiguous prediction
    THEN the prediction is made by the state
    ELSE
    Use its corresponding association
      rules to make prediction
ENDIF
```

In this work, we define an ambiguous prediction as two or more predictive pages that have the same conditional probability by a Markov model. The ambiguous prediction potentially has other definitions, for example, the certainty of a prediction is below a threshold. We did not explore other options in this paper.

4 Experimental Evaluation

4.1 Data Collection and Preprocessing

For our experiments, the first step was to gather a log file from an active Web server. Usually, Web log files are the main source of data for any e-commerce or Web related session analysis (Spiliopoulou et al. 1999). The log file we used as a data source for our experiments is a day's worth of all HTTP requests to the EPA WWW server located at Research Triangle Park, NC. The logs are an ASCII file with one line per request, with the following information: The host making the request, date and time of request, requested page, HTTP reply code and bytes in the reply. The logs were collected for Wednesday, August 30 1995. There were 47,748 total requests, 46,014 GET requests, 1,622 POST requests, 107 HEAD requests and 6 invalid requests. The gathered Web log data had to be cleaned and filtered (Zhao, Bhowmick & Gruenwald 2005, Sarukkai 2000).

Cleaning the data involved removing erroneous and invalid pages. Those included HTTP error codes 400s and 500s, HTTP 1.0 errors, and CGI entries. The total number of valid entries was diminished to 19,121. Then, 302 and 304 HTTP errors that involve requests with no server replies were also removed and the number of entries went down to 14,091. Filtering and cleaning the log files made them ready for further preprocessing and analysis. Pages links are converted to numbers for easy manipulation. Repeated pages are removed because it is uncommon for the same Web page to be accessed more than once and any internal links are irrelevant. Next step was to identify user sessions. Taking a 30-minute timeout into consideration, the number of user sessions amounted to 1,868. Short sessions were then removed and only sessions with at least 5 pages were considered. Distinct Web pages were identified and they amounted to 2,891 pages.

The EPA data was further pre-processed before being used for our analysis purposes. The last page of each session was removed for testing purposes. Also, the frequency of each page visited by the user was calculated. The page access frequency is shown in Figure 2 which reveals that page number 3 is the most frequent page and it was accessed 73 times.

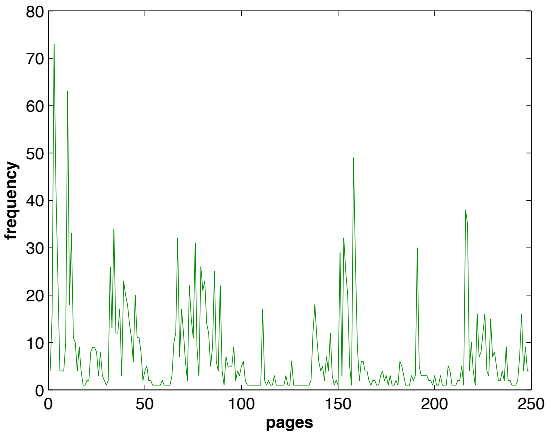

Figure 2: Frequency chart for the most frequent visited pages.

4.2 Experiments Results

Having all data sets processed, filtered and analysed, 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order Markov models were created. Then, all 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} order frequency pruned (Deshpande et al. 2004) Markov model analysis took place considering 4 as the frequency threshold. Prediction results were achieved using the maximum likelihood based on conditional probabilities as stated in equation 3 above. All predictions in the test data that did not exist in the training data sets were assumed incorrect and were given a zero value. All implementations were carried out using MATLAB. Figure 3 below illustrates the difference between Markov model orders and Frequency pruned all-kth Markov model results. The Figure demonstrates that as the order of Markov model increases, precision decreases due to the reduced coverage of the data. Coverage is defined as the ratio of the Web sessions in the test set that have a corresponding state in the training set to the number of Web sessions in the test set (Deshpande et al. 2004). Also, the increase of the frequency pruned Markov model precision is limited due to the elimination of states that could be of importance to the precision process. The frequency threshold parameter used was a fixed parameter of size 4.

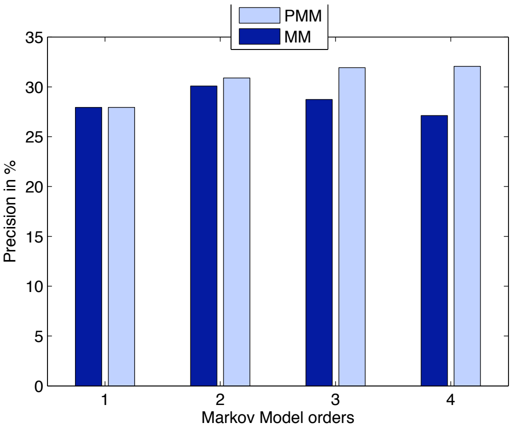

Figure 3: Precision of 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order Markov models and all 1^{st}, 2^{nd}, 3^{rd} and 4^{th} order frequency pruned Markov models.

Table 7 below reveals that the all- 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} order frequency pruned Markov models have considerably less states than the 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} order Markov models.

Table 7: Number of states of Markov model and frequency pruned Markov model orders.

Model	MM States	All-kth FP States
1^{st} order	1945	745
2^{nd} order	39162	9162
3^{rd} order	72524	14977
4^{th} order	101365	17034

The reported accuracies in this section are based on 10-fold cross validation. The data was split into ten equal sets. First, we considered the first nine sets as training data and the last set for test data. Then, the second last set was used for testing and the rest for training. We continued moving the test set upward until the first set was used for testing and the rest for training. The reported accuracy is the average of ten tests.

The 1^{st} order and 2^{nd} order Markov model results cannot be 100% reliable simply because we did not look back into the history of pages accessed by the user. We assumed that the pages visited long before the current page in a Web session do tend to influence the users actions. These previously accessed pages affect the prediction process as they interfere with the user browsing behaviour and are not mere information providers. Performing 3^{rd} and 4^{th} order Markov models techniques solves the problem of examining the users previous browsing behaviour, but it results in an increase in the number of states as it is obvious in Table 7 above that illustrates the number of states generated based upon non empty states. To overcome this shortcoming, we applied subsequence association rules techniques in order to generate the most appropriate rule. Before applying association rules techniques, the most frequent occurrences or the Markov model frequent states are removed.

Since association rules techniques require the determination of a minimum support factor and a confidence factor, we used the experimental data to help determine such factors. We can only consider rules with certain support factor and above a certain confidence threshold.

Figure 4 below shows that the number of generated association rules dramatically decreases with the increase of the minimum support threshold with a fixed 90% confidence factor. Reducing the confidence factor results in an increase in the number of rules generated. This is apparent in Figure 5 where the number of generated rules decreases with the increase of the confidence factor while the support threshold is a fixed 4% value. It is also apparent from Figure 4 and Figure 5 below that the influence of the minimum support factor is much greater on the number of rules than the influence of the confidence factor.

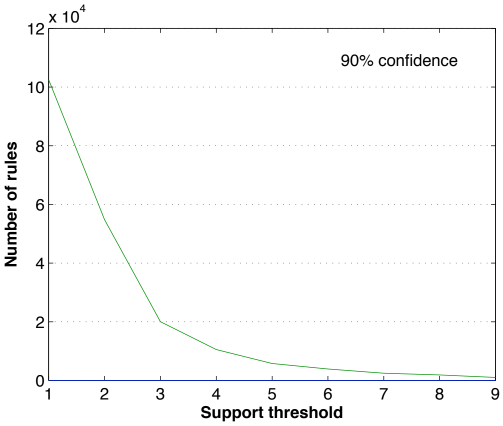

Figure 4: Number of rules generated according to different support threshold values and a fixed confidence factor: 90%.

Referring back to Figure 4 and Figure 5, we considered a minimum support threshold of 4%. The integration model, IMAM, involves calculating association rules techniques prediction accuracy using the longest match precision method. In IMAM, association rules were applied in two cases:

1. When we were unable to make a correct prediction in the case of a 2^{nd} order Markov model because of a tie. In such a case, using association rules techniques to look further back at previously visited pages, we were able to break the tie by looking at the page in history that leads to the most appropriate page for prediction. Looking at Figure 6, using 1^{st} order Markov

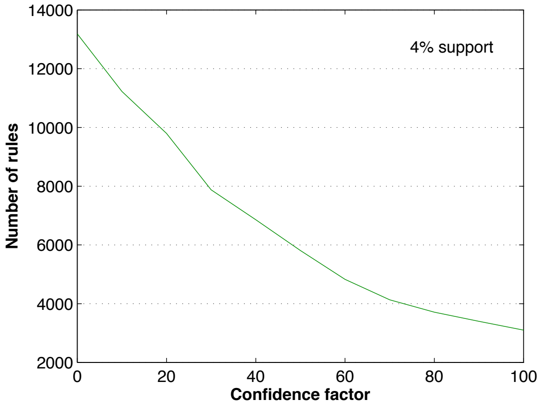

Figure 5: No. of rules generated according to a fixed support threshold: 4%.

model, the most frequently accessed page after EPA-PEST1995Aug23 is EPA-PEST1995Aug17 with 100% probability. Using 2^{nd} order Markov model, the most frequently accessed pages after EPA-PEST1995Aug17 are EPA-PEST1995July and OOPTPubs with 50% probability each. To decide which of the two pages would result in higher prediction precision, we look further back. Using association rules we find out that there is 100% chance that if EPA-PEST1995Aug16pr-373 is accessed before EPA-PEST1995Aug23, EPA-PEST1995July will be accessed next. And, there is 100% chance if PressReleases1995Aug is accessed before EPA-PEST1995Aug23, OOPTPubs will be accessed next. As a result, precision is calculated according to the results of association rules.

The precision of the proposed IMAM model was calculated by adding all successes and dividing the result by the number of states in the test data. According to Figure 7, the proposed IMAM model shows better precision than the 2^{nd} order Markov model (MM) and the frequency pruned all 2^{nd} order Markov model (PMM).

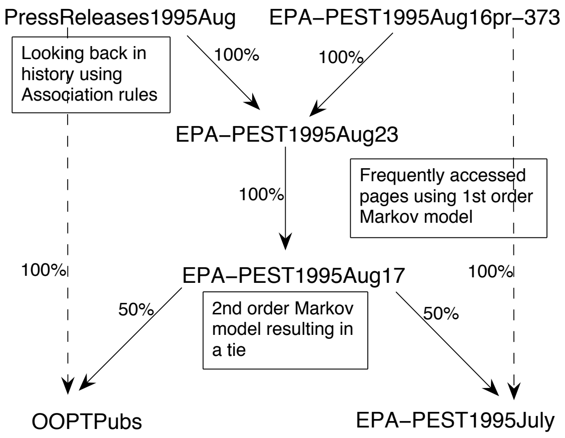
PressReleases1995Aug
EPA–PEST1995Aug16pr–373
Looking back in history using Association rules
100%
100%
EPA–PEST1995Aug23
100%
Frequently accessed pages using 1st order Markov model
100%
EPA–PEST1995Aug17
100%
50%
2nd order Markov model resulting in a tie
50%
OOPTPubs
EPA–PEST1995July

Figure 6: Portion of association rules results.

2. If the test data does not match any of the 2^{nd} order Markov model outcomes, we use the globally generated association rules to look back at previous user browsing history. Users have different browsing experiences, some of them get to the page they request using a shorter path than others depending upon the web site structure and internal links. For example, the same page could be accessed by a user after visiting 5 pages and by another user after visiting 2 pages.

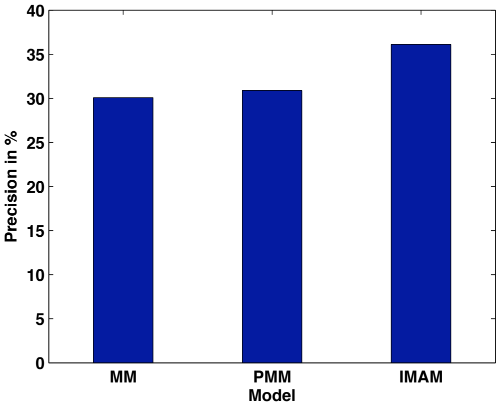

Figure 7: Precision of 2^{nd} order Markov model (MM), Frequency Pruned all 2^{nd} order Markov model (PMM) and IMAM model.

The main problem associated with this approach is that it is dependent on the length of user session of data available. This is usually not a problem when modelling a particular site with long user sessions and therefore, more history. But it becomes more difficult when performing multi-site analysis with shorter user sessions.

5 Conclusion

In this paper, we proposed a method to integrate Markov model and association rules for predicting Web page accesses. The integration is based on a low order Markov model. Sets of subsequence association rules are used to complement the Markov model for resolving ambiguous predictions by using long history data. The integration avoids the complexity of high order Markov model and the limitation of Markov model using short history. This model also reduces the large number of association rules since association rules are only used when ambiguous predictions occur. The experimental results show that the combined model increases the accuracy of the Web page access prediction of Markov model and association rules.

References

Cadez, I., Heckerman, D., Meek, C., Smyth P. & White S. (2000), Visualization of Navigation Patterns on a Web Site Using Model Based Clustering, *in* 'ACM SIGMOD International Conference on Knowledge Discovery and Data Mining', ACM Press, Washington DC, USA, pp. 280-284.

Deshpande, M. & Karypis, G. (2004), 'Models for Predicting Web Page Accesses', *Transactions on Internet Technology* **4**(2), 163–184.

Dongshan, X. & Junyi, S. (2002), 'A New Markov Model for Web Access Prediction', *Computing in Science and Engineering* **4**(6), 34–39.

Garafalakis, M., Rastogi, R., Seshadri, S. & Shim, K. (1999), Data Mining and the Web: Past, Present and Future, *in* 'WIDM Conference', Kansas City, USA, pp. 43-47.

Gunduz, S. & Ozsu, M. (2003), A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior, *in* 'SIGKDD 03 Conference', Washington, DC, USA, pp. 24-27.

Jespersen, S., Pedersen, T. B. & Thorhauge, J. (2003), Evaluating the Markov Assumption for Web Usage Mining, *in* '5th international workshop on WIDM03', New Orleans, USA, pp. 82-89.

Kim, D., Lm, L., Adam, N., Atluri, V., Bieber, M. & Yesha, Y. (2004), A Clickstream-Based Collaborative Filtering Personalization Model: Towards A Better Performance, *in* 'WIDM 04 Conference', Washington DC, USA, pp. 12-13.

Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2000), Discovery of Aggregate Usage Profiles for Web Personalisation, *in* 'WebKDD Workshop 2000', USA, pp. 61-82.

Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2001), Effective Personalization Based on Association Rule Discovery from Web Usage Data, *in* 'WIDM 01, 3^{rd} ACM Workshop on Web Information and Data Management', Atlanta, Georgia, USA, pp. 9-15.

Pei, J., Han, J., Mortazavi-asl, B. & Zhu, H. (2000), Mining access patterns efficiently from Web logs, *in* 'PAKDD conference', USA, pp. 396-407.

Pitkow, J. & Pirolli, P. (1999), Mining longest repeating subsequence to predict world wide Web surfing, *in* 'the 2^{nd} USENIX Symposium on Internet Technologies and Systems', Boulder, CO., pp. 139-150.

Sarukkai, R. (2000), Link Prediction and path analysis using Markov Chains, *in* 'the Ninth International World Wide Web Conference', Amsterdam, pp. 377-386.

Spiliopoulou M., Faulstich L. C. & Winkler K. (1999), A Data Miner analyzing the Navigational Behaviour of Web Users, *in* 'Workshop on Machine Learning in User Modelling of the ACAI99 International Conference', Creta, Greece.

Yang, Q., Li, T. & Wang, K. (2004), 'Building Association-Rule Based Sequential Classifiers for Web-document Prediction', *Journal of Data Mining and Knowledge Discovery* **8**(3), 253–273.

Yong, W., Zhanhuai, L. & Yang, Z. (2005), Mining Sequential Association-Rule for Improving WEB Document Prediction, *in* 'Sixth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA05)', pp. 146-151.

Zhao, Q., Bhowmick, S. S. & Gruenwald, L. (2005), WAM:Miner: In the Search of Web Access Motifs from Historical Web Log Data, *in* 'CIKM05 conference', Germany, pp. 421-428