Integrating Recommendation Models for Improved Web page prediction accuracy

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

Jiuyong Li
School of Computer & Information Science
University of South Australia
Mason Lakes, Australia
Email: Jiuyong.Li@unisa.edu.au

July 17, 2007

Abstract

Recent research initiatives have addressed the need for improved performance of Web page prediction that would profit many applications, e-business in particular. Despite the various efforts so far, there is still room for advancement in this field. This paper endeavors to provide an improved prediction accuracy by using a novel approach that involves combining clustering, association rules and Markov models. Each of these frameworks has its own strengths and weaknesses and their integration proves to provide better prediction than using each technique individually.

1 Introduction

Web page access prediction gained its importance from the ever increasing number of e-commerce Web information systems and e-businesses. Web page prediction that involves personalizing the Web users' browsing experiences assists Web masters in the improvement of the Web site structure, and helps Web users in navigating the site and accessing the information they need. Various attempts have been exploited to achieve Web page access prediction by preprocessing Web server log files and analyzing Web users' navigational patterns. The most widely used approach for this purpose is Web usage mining that entails many techniques like Markov model, association rules and clustering [21].

- Markov models are the most effective techniques for Web page access prediction and Many researchers stress out their importance in the field [2, 4, 5, 7, 28]. Other researchers use Markov models to improve the Web server access efficiency either by using object prefetching [17] or by helping reduce the Web server overhead [14]. Lower order Markov models are known for their low accuracy due to the limited availability of users' browsing history. Higher order Markov models achieve higher accuracy but are associated with higher state space complexity.

- Association rule mining is a major pattern discovery technique [15]. The original goal of association rule mining is to solve market basket problem but The applications of association rules are far beyond that. Using association rules for Web page access prediction involves dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions [10, 15, 25].

- Although clustering techniques have been used for personalization purposes by discovering Web site structure and extracting useful patterns [1, 3, 16, 18, 22], usually, they are not very successful in attaining good results. Proper clustering groups users sessions with similar browsing history together, and this facilitates classification. However, prediction is performed on the cluster sets rather than the actual sessions.

Therefore, there arises a need for improvement when using any of the aforementioned techniques. This paper integrates all three frameworks together, clustering, association rules and Markov model, to achieve better Web page access prediction performance specifically when it comes to accuracy efficiency.

2 Literature Review

A number of researchers attempted to improve the Web page access prediction precision or coverage by combining different recommendation frameworks. For instance, many papers combined clustering with association rules [11, 12]. Lai *et. al* [11], have introduced a customized marketing on the Web approach using a combination of clustering and association rules. The authors collected information about customers using forms, Web server log files and cookies. They categorized customers according to the information collected. Since k-means clustering algorithm works only with numerical data, the authors used PAM (Partitioning Around Medoids) algorithm to cluster data using categorical scales. They then performed association rules techniques on each cluster. They proved through experimentations that implementing association rules on clusters achieves better results than on non-clustered data for customizing the customers' marketing preferences. Liu *et. al* [12] have introduced MARC (Mining Association Rules using Clustering) that helps reduce the I/O overhead associated with large databases by making only one pass over the database when learning association rules. The authors group similar transactions together and they mine association rules on the summaries of clusters instead of the whole data set. Although the authors prove through experimentation that MARC can learn association rules more efficiently, their algorithm does not improve on the accuracy of the association rules learned.

Other papers combined clustering with Markov model [3, 28, 13]. Cadez *et. al* [3] partitioned site users using a model-based clustering approach where they implemented first order Markov model using the Expectation-Maximization algorithm. After partitioning the users into clusters, they displayed the paths for users within each cluster. They also developed a visualization tool called WebCANVAS based on their model. Zhu it et. al [28] construct Markov models from log files and use co-citation and coupling similarities for measuring the conceptual relationships between Web pages. CitationCluster algorithm is then proposed to cluster conceptually related pages. A hierarchy of the Web site is constructed from the clustering results. The authors then combine Markov model based link prediction to the conceptual hierarchy into a prototype called ONE to assist users' navigation. Lu *et. al* [13] were able to generate Significant Usage Patterns (SUP) from clusters of abstracted Web sessions. Clustering was applied based on a two-phase abstraction technique. First, session similarity is computed using Needleman-Wunsch alignment algorithm and sessions are clustered according to their similarities. Second, a concept-based abstraction approach is used for further abstraction and a first order Markov model is built for each cluster of sessions. SUPs are the paths that are generated from first order Markov model with each cluster of user sessions.

Combining association rules with Markov model is novel to our knowledge and only few of past researches combined all three models together [10]. Kim *et. al* improve the performance of Markov model, sequential association rules, association rules and clustering by combining all these models together. For instance, Markov model is used first. If MM cannot cover an active session or a state, sequential association rules are used. If sequential association rules cannot cover the state, association rules are used. If association rules cannot cover the state, clustering algorithm is applied. Kim *et. al* work improved recall and it did not improve the Web page prediction accuracy. Our work proves to outperform previous works in terms of Web page prediction accuracy using a combination of clustering, association rules and Markov model techniques.

3 Proposed Model

Our work is based on combining clustering algorithm, association rules mining and Markov model during the prediction process. The new model is known as Integrated Prediction Model or IPM. The process is as follows:

```
Training:

    Cluster user sessions into l-clusters
    Build a k-Markov model for each cluster
    For Markov model states where the majority is not clear
    Discover association rules for each state

Prediction:

    For each coming session
        Find its closest cluster
        Use corresponding Markov model to make prediction
        If the predictions are made by states that do not belong to a
        majority class
            Use association rules to make a revised prediction
```

The majority class includes states with high probabilities where probability differences between two pages are significant. In other words, a Markov model state is retained only if the probability difference between the most probable state and the second probable state is above (ϕ_c) [5]. On the other hand, the minority class includes all other cases. In particular, the minority class includes:

1. States with high probabilities where probability differences between two pages are below (ϕ_c) or equal to zero.

2. States where test data does not match any of the Markov model outcomes.

4 Methdology

4.1 Clustering

The main purpose of clustering the Web server log file is to combine meaningful sessions together in order to improve the Markov model prediction accuracy. Performing clustering tasks can be tedious and complex due to the increased number of clustering methods and algorithms. Clustering could be hierarchical or non-hierarchical [9], distance-based or model-based [27], and supervised or unsupervised [6]. For the purpose of this paper, we use a straightforward implementation of the k-means clustering algorithm which is distance-based, based on user sessions, unsupervised and partitional non-hierarchical. K-means clustering algorithm involves defining a set of items (n-by-p data matrix) to be clustered, defining a chosen number of clusters (k) and randomly assign a number of items to each cluster. K-means clustering then repeatedly calculates the mean vector for all items in each cluster and reassigns the items to the cluster whose center is closest to the item. Because the first clusters are created randomly, k-means runs different times each time it starts from a different point giving different results. The different clustering solutions are compared using the sum of distances within clusters. In this paper, clusters were achieved using MatLab that considers the clustering solution with the least sum of distances. Therefore, k-means clustering depends greatly on the number of clusters (k), the number of runs and the distance measure used. The output is a number of clusters with a number of items in each cluster.

This can take place using a variety of distance measures, in particular, Euclidean, Squared Euclidean, City Block, Hamming, Cosine and Correlation [22]. K-means computes centroid clusters differently for different k-means supported distance measures. Therefore, a normalization step is necessary for Cosine and Correlation distance measures for comparison purposes. The points in each cluster, whose mean forms the centroid of the cluster, are normalized to unit Euclidean length. In this paper we use Cosine distance measure that, according to Strehl *et al.* [22] and Halkidi *et al.* [8], yields better clustering results than the other distance measures and is a direct application of the extended Jaccard coefficient [22, 8].

4.2 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 [5].

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:

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

This probability, $prob(pi|W)$, is estimated by using all 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:

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

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 all k^{th} 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 probability of $P\left(p_i|S_j^k\right)$ is estimated as follows from a history (training) data set.

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

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. In this paper, we use the 2^{nd} order Markov model because it has better accuracy than that of the all 1^{st} order Markov model without the drawback of the state space complexity of the all 3^{rd} and all 4^{th} order Markov model. We also employ the confidence pruned Markov model introduced by Deshpande *et. al*, [5]. The confidence threshold was calculated as follows:

$$\phi_c = \hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4}$$

Where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution, and n is the frequency of the Markov state. Equation 5 stresses out the fact that states with high frequency would lead to smaller confidence threshold. That means that even if the difference between the two most probable pages is small, the state with higher probability will be chosen in the case of high frequency of the state occurrence. The smaller confidence threshold results in larger majority class. The effect of the confidence threshold value and, therefore, the majority class size on the prediction accuracy depends on the actual data set. A number of experiments took place to determine the optimal value of $z_{\alpha/2}$ and, as a result, the value of the confidence factor ϕ_c. As Table 1 depicts, the increase of the minority class or, in other words, the increase in the confidence factor is affected by the decrease of $z_{\alpha/2}$. During the prediction process, if the Markov model probability belongs to the minority class, association rules probability for the item is taken into consideration instead. Table 1 displays the results of the IPM accuracy using different values for $z_{\alpha/2}$. It is clear that the accuracy increases at first with lower confidence threshold and therefore, larger minority class. However, after a certain point, accuracy starts to decrease when the majority class is reduced to the extent where it looses the advantage of the accuracy obtained by combining Markov model and clustering. The optimal value for $z_{\alpha/2}$ is 1.15. Note that the number of states has dramatically decreased.

With $z_{\alpha/2}$=1.15, the most probable pages range approximately between 80% and 40% with ϕ_c ranging between 47% and zero respectively given n=2. This results in approximately 0.78 as the ratio of the majority class to the whole data set. This leaves space for 22% improvement using association rules mining not including instances that have zero matching states in the training data set.

4.3 Association Rules

Association rule mining, a major pattern discovery technique, is implemented with Markov model and clustering in order to improve the Web page access prediction accuracy. Association rules are mainly defined by two metrics: support and confidence. 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

Table 1: Accuracy according to $z_{\alpha/2}$ value

$z_{\alpha/2}$	Accuracy	# states
0	31.29	9162
0.75	33.57	2061
0.84	35.45	1932
0.93	37.80	1744
1.03	40.60	1729
1.15	44.91	1706
1.28	43.81	1689
1.44	40.93	1614
1.64	38.85	1557
1.96	37.91	1479
2.57	36.81	1304

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.

There are four types of sequential association rules presented by Yang *et al.* [24]:

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.* [24], 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 [24], the latest substring with the least error matching produces the most accurate models for Web document prediction.

In this paper, we perform sequential association rule mining on the whole data set and will only be referred to when the Markov model prediction leads to an instance that does not belong to the majority class.

4.4 Combining Clustering, Markov Model and Association Rules

The IPM process depends on three prediction fundamentals, Markov model, clustering and association rules. This process has significant advantages. For instance, association rule mining entails generating redundant rules that sometimes make it unsuitable for prediction purposes. Using IPM, association rules will only be used in the case where the predictions are made by states that do not belong to a majority class. Also, lower order Markov models do not look back at users' browsing history to make their prediction fully accountable. IPM has the advantage of allowing us to look at users' previous browsing history using association rule mining only in the cases where Markov model prediction confidence is not high. The added advantage of IPM is the use of clustering techniques to further reduce the high state space complexity associated with higher order Markov models and to further improve the prediction accuracy. Clustering combines similar Web page paths or user sessions together and the subsets of data are therefore more homogeneous.

For instance, consider table 2 that depicts data transactions performed by a user browsing a Web site.

Performing clustering analysis on the data set using k-means clustering algorithm and Cosine distance measure where the number of clusters k=2 results in the following two clusters:

Cluster 1:	T1	A, F, I, J, E, C, D, H, N, I, J, G, D, H, N, C, I, J, G
	T2	F, D, H, N, I, J, E, A, C, D, H, N, I, J, G
	T3	F, D, H, I, J, E, H, F, I, J, E, D, H, M

Table 2: User sessions

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

Cluster 2:

T5	E, C, A, C, F, I, A, C, G, A, D, H, M, G, J
T4	G, E, A, C, F, D, H, M, I, C, A, C, G

Consider the following test data state I → J → ?. Applying the 2^{nd} order Markov Model to the above training user sessions we notice that the state $\langle I, J \rangle$ belongs to cluster 1 and it appeared 7 times as follows:

$$P_{l+1} = \text{argmax}\{P(E|J, I)\} = \text{argmax}\{E \rightarrow 0.57\}$$

$$P_{l+1} = \text{argmax}\{P(G|J, I)\} = \text{argmax}\{G \rightarrow 0.43\}$$

This information alone does not provide us with correct prediction of the next page to be accessed by the user as we have high probabilities for both pages, G and E. Although the result does not conclude with a tie, neither G nor E belong to the majority class. The difference between the two pages (0.14), is not higher than the confidence threshold (in this case 0.2745). In order to 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 appears in Table 3 below.

Table 3: User sessions history

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

Tables 4 and 5 summarise the results of applying subsequence association rules to the training data. Table 4 shows that F → E has the highest confidence of 100%. While Table 5 shows that C→ G has the highest confidence of 100%.

Using Markov models, we can determine that the next page to be accessed by the user after accessing the pages I and J could be either E or G. Whereas subsequence association rules take this result a step further by determining that if the user accesses page F before pages I and J, then there is a 100% confidence

Table 4: Confidence of accessing page E using subsequence association rules

A → E	AE/A	1/2	50%
F → E	FE/F	4/4	100%
D → E	DE/D	2/6	33%
H → E	HE/H	2/7	29%
N → E	NE/N	1/4	25%

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

C → G	CG/C	3/3	100%
D → G	DG/D	3/6	50%
H → G	HG/H	3/7	43%
N → G	NG/N	3/4	75%
A → G	AG/A	1/2	50%

that the user will access page E next. Whereas, if the user visits page C before visiting pages I and J, then there is a 100% confidence that the user will access page G next.

5 Experimental Evaluation

5.1 Experimental Setup

For our experiments, the first step was to gather log files from active web servers. Usually, Web log files are the main source of data for any e-commerce or Web related session analysis [20]. 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.

Before using the EPA log file data, it was necessary to perform data preprocessing [26, 19]. We removed erroneous and invalid pages. Those include HTTP error codes 400s, 500s, and HTTP 1.0 errors, as well as, 302 and 304 HTTP errors that involve requests with no server replies. We also eliminated multi-media files such as gif, jpg and script files such as js and cgi.

Next step was to identify user sessions. A session is a sequence of URLs requested by the same user within a reasonable time. The end of a session is determined by a 30 minute threshold between two consecutive web page requests. If the number of requests is more than the predefined threshold value, we conclude that the user is not a regular user; it is either a robot activity, a web spider or a programmed web crawler. Short sessions were also removed and only sessions with at least 5 pages were considered. the EPA preprocessing and filtering resulted in 799 web sessions. The sessions of the data set are of different length. They were represented by vectors with the number of occurrence of pages as weights.

Finally, EPA sessions were categorized according to feature selection techniques introduced by Wang et al. [23]. The pages, and not users, were grouped according to services requested which yield best results if carried out according to functionality [23]. The grouping of Web pages according to functionality could be done either by removing the suffix of visited pages or the prefix. In our case, we could not merge according to suffix because, for example, pages with suffix index.html could mean any default page like OWOW/sec4/index.html or OWOW/sec9/index.html or ozone/index.html. Therefore, merging was according to a prefix. Since not all Web sites have a specific structure where we can go up the hierarchy to a suitable level, we had to come up with a suitable automatic method that can merge similar pages automatically. A program runs and examines each record. It only keeps the delimited and unique word. A manual examination of the results also takes place to further reduce the number of categories by combining similar pages.

5.2 Clustering, Markov Model and Association Rules

All clustering experiments were developed using MATLAB statistics toolbox. Since k-means computes different centroids each run and this yields different clustering results each time, the best clustering solution with the least sum of distances is considered using MATLAB k-means clustering solutions. Therefore, using Cosine distance measure with the number of clusters (k)=7 leads to good clustering results while keeping the number of clusters to a minimum. 7 clusters were obtained in 17 iterations with the least sum of distances of 99.1192. Figure 1, Figure 2, Figure 3, Figure 4 and Figure 5 represent clusters using Euclidean, Hamming, City Block, Pearson Correlation and Cosine distance measures respectively. They plot the silhouette value represented by the cluster indices displaying a measure of how close each point in one cluster is to points in the neighboring clusters. The silhouette measure ranges from +1, indicating points that are very distant from neighboring clusters, to 0, indicating points that do not belong to a cluster. The figures reveal that the order of distance measures from worst to best are Hamming, City Block, Euclidean, Pearson Correlation and Cosine respectively. For instance, the maximum silhouette value in Figure 3 for Hamming distance is around 0.5, whereas, the silhouette value of Figure 6 for Cosine distance ranges between 0.5 and 0.9. The larger silhouette value of the Cosine distance implies that the clusters are separated from neighboring clusters.

Merging Web pages by web services according to functionality reduces the number of unique pages from 2924 to 155 categories. The sessions were divided into 7 clusters using the k-means algorithm and according to the Cosine distance measure. For each cluster, the categories were expanded back to their original form in the data set. This process is performed using a simple program that

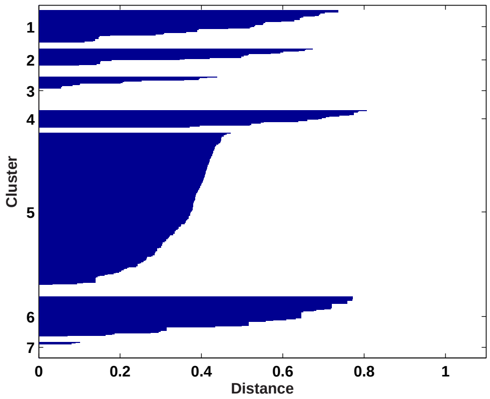

Figure 1: Silhouette value of Euclidean distance measure with 7 clusters.

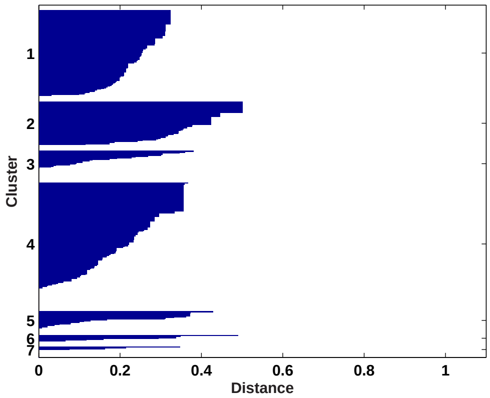

Figure 2: Silhouette value of Hamming distance measure with 7 clusters.

seeks and displays the data related to each category.

Markov model implementation was carried out for the whole data set. The data set was divided into training set and test set and 2-Markov model accuracy was calculated accordingly. Then, using the test set, each transaction was considered as a new point and distance measures were calculated in order to define the cluster that the point belongs to. Next, 2-Markov model prediction accuracy was computed considering the transaction as a test set and only the cluster that the transaction belongs to as a training set. Prediction accuracy 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. The Markov model accuracy was calculated using a 10-fold cross

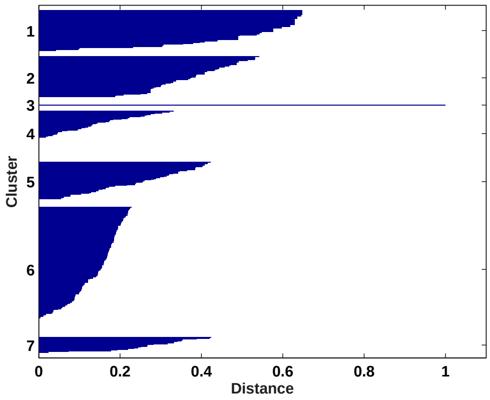

Figure 3: Silhouette value of City Block distance measure with 7 clusters.

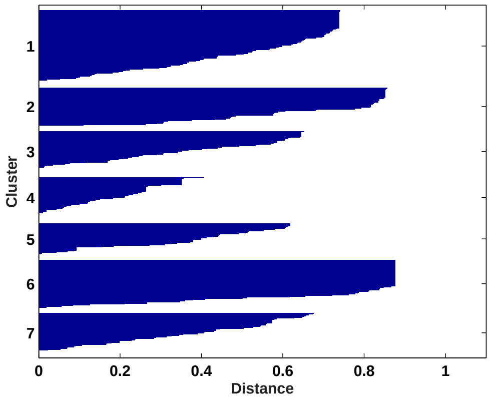

Figure 4: Silhouette value of Correlation distance measure with 7 clusters.

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.

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 6 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 in-

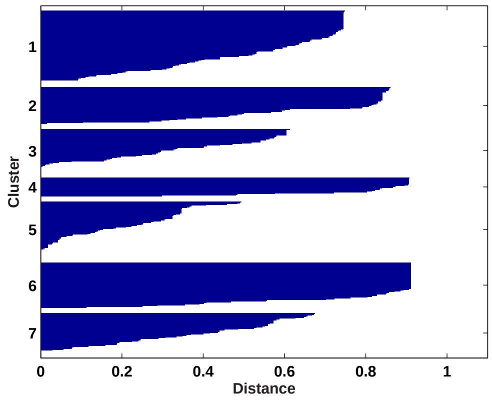

Figure 5: Silhouette value of Cosine distance measure with 7 clusters.

crease in the number of rules generated. This is apparent in Figure 7 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 6 and Figure 7 below that the influence of the minimum support factor is much greater on the number of rules than the influence of the confidence factor. IPM involves calculating association rules techniques prediction accuracy using the longest match precision method.

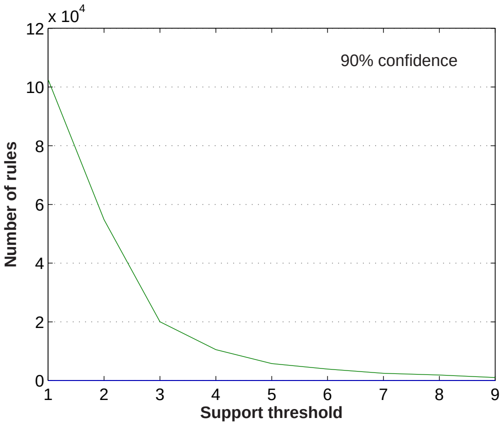

Figure 6: Number of rules generated according to different support threshold values and a fixed confidence factor: 90%.

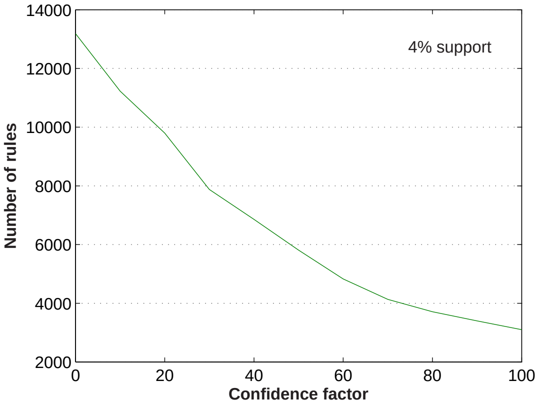

Figure 7: No. of rules generated according to a fixed support threshold: 4%.

5.3 Experiments Results

Figure 8 depicts better Web page access prediction accuracy by integrating Markov model, Association rules and clustering (IPM) as follows:

1. The data set is clustered according to k-means clustering algorithm and Cosine distance measure.
2. For each new instance, the prediction accuracy is calculated based on the 2-MM performed on the closest cluster.
3. If the prediction results in a state that does not belong to the majority class, global association rules are used for prediction.
4. The frequency of the item is also determined in that particular cluster.
5. ϕ_c is calculated for the new instance using $z_{\alpha/2}$ value to determine if it belongs to the majority class.
6. if the state does not belong to the majority class, global association rules are used to determine the prediction accuracy, otherwise, the original accuracy is used.

The experiments prove that combining Markov model with clustering techniques yields more significant accuracy improvement than combining Markov model with Association rules mining. However, combining the three models together yields best results.

All clustering runs were performed on a desktop PC with a Pentium IV Intel processor running at 2 GHz with 2 GB of RAM and 100 GB of hard disk memory. In our largest runs with K = 50, we exhausted around 6.1 MB of memory in 34 seconds. The runtime of the k-means algorithm, regardless of the distance

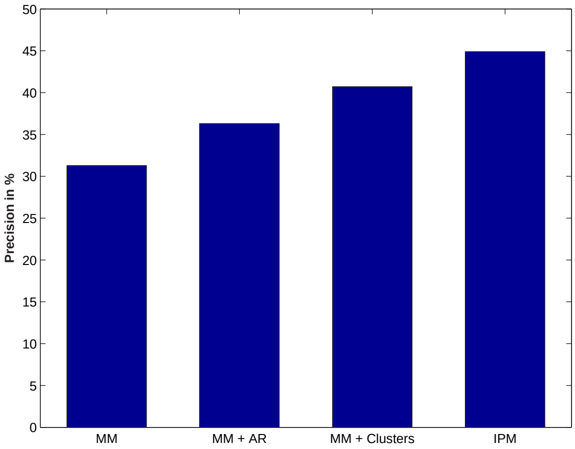

Figure 8: Precision of 2^{nd} order Markov model (2-MM) and 2-MM with Association rules mining and 2-MM with Clustering and all three models together (IPM) .

measure used, is equivalent to O(nkl) [9], where n is the number of items, k is the number of clusters and l is the number of iterations taken by the algorithm to converge. For our experiments, where n and k are fixed, the algorithm has a linear time complexity in terms of the size of the data set. The k-means algorithm has a $O(k+n)$ space complexity. This is because it requires space to store the data matrix. It is feasible to store the data matrix in a secondary memory and then the space complexity will become O(k). k-means algorithm is more time and space efficient than hierarchical clustering algorithms with O($n^2 logn$) time complexity and O(n^2) space complexity. As for all 2^{nd} order Markov model, the running time of the whole data set was similar to that of the clusters added together because the running time is in terms of the size of the data. i.e. T(n)=T(k1)+T(k2)+T(k3)+...T(ki) where time is denoted by T, the number of items in the data set is denoted by n, and the clusters are denoted by ki. The running time of association rule mining is the same in all cases above regardless of the size of the majority class. The association rules produced were for the whole data set. Accessing the appropriate rule is, however, performed online at time of prediction.

6 Conclusion

This paper improves the Web page access prediction accuracy by integrating all three prediction models: Markov model, Clustering and association rules according to certain constraints. Our model, IPM, integrates the three models using 2-Markov model computed on clusters achieved using k-means clustering algorithm and Cosine distance measures for states that belong to the majority class and performing association rules mining on the rest. Previous studies reveal the improved efficiency of combining the three models together but not accuracy. The experiments prove the significant improvement of the Web page access prediction accuracy.

References

[1] G. Adami, P. Avesani, and D. Sona. Clustering documents in a web directory. *WIDM'03, USA*, pages 66–73, 2003.

[2] C. Bouras and A. Konidaris. Predictive prefetching on the web and its potential impact in the wide area. *WWW: Internet and Web Information Systems*, (7):143–179, 2004.

[3] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-based clustering and visualization of navigation patterns on a web site. *Data Mining and Knowledge Discovery*, 7, 2003.

[4] M. chen, A. S. LaPaugh, and J. P. Singh. Predicting category accesses for a user in a structured information space. *SIGIR'02, Finland*, pages 65–72, 2002.

[5] M. Deshpande and G. Karypis. Selective markov models for predicting web page accesses. *Transactions on Internet Technology*, 4(2):163–184, 2004.

[6] C. F. Eick, N. Zeidat, and Z. Zhao. Supervised clustering - algorithms and benefits. *IEEE ICTAI'04*, pages 774–776, 2004.

[7] M. Eirinaki, M. Vazirgiannis, and D. Kapogiannis. Web path recommendations based on page ranking and markov models. *WIDM'05*, pages 2–9, 2005.

[8] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis. Thesus: Organizing web document collections based on link semantics. *The VLDB Journal*, 2003(12):320–332, 2003.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[10] D. Kim, N. Adam, V. Alturi, M. Bieber, and Y. Yesha. A clickstream-based collaborative filtering personalization model: Towards a better performance. *WIDM '04*, pages 88–95, 2004.

[11] H. Lai and T. C. Yang. A group-based inference approach to customized marketing on the web - integrating clustering and association rules techniques. *Hawaii International Conference on System Sciences*, pages 37–46, 2000.

[12] F. Liu, Z. Lu, and S. Lu. Mining association rules using clustering. *Intelligent Data Analysis*, (5):309–326, 2001.

[13] L. Lu, M. Dunham, and Y. Meng. Discovery of significant usage patterns from clusters of clickstream data. *WebKDD '05*, 2005.

[14] V. Mathur and V. Apte. An overhead and resource contention aware analytical model for overloaded web servers. *WOSP'07, Argentina*, 2007.

[15] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. *WIDM'01, USA*, pages 9–15, 2001.

[16] N. K. Papadakis and D. Skoutas. STAVIES: A system for information extraction from unknown web data sources through automatic web warpper generation using clustering techniques. *IEEE Transactions on Knowledge and Data Engineering*, 17(12):1638–1652, 2005.

[17] A. P. Pons. Object prefetching using semantic links. *The DATA BASE for Advances in Information Systems*, 37(1):97–109, 2006.

[18] M. Rigou, S. Sirmakesses, and G. Tzimas. A method for personalized clustering in data intensive web applications. *APS'06, Denmark*, pages 35–40, 2006.

[19] R. Sarukkai. Link prediction and path analysis using markov chains. *9th International WWW Conference, Amsterdam*, pages 377–386, 2000.

[20] M. Spiliopoulou, L. C. Faulstich, and K. Winkler. A data miner analysing the navigational behaviour of web users. *Workshop on Machine Learning in User Modelling of the ACAI'99, Greece*, 1999.

[21] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGDD Explorations*, 1(2):12–23, 2000.

[22] A. Strehl, J. Ghosh, and R. J. Mooney. Impact of similarity measures on web-page clustering. *AI for Web Search*, pages 58–64, 2000.

[23] Q. Wang, D. J. Makaroff, and H. K. Edwards. Characterizing customer groups for an e-commerce website. *EC'04, USA*, pages 218–227, 2004.

[24] Q. Yang, T. Li, and K. Wang. Building association-rule based sequential classifiers for web-document prediction. *Journal of Data Mining and Knowledge Discovery*, 8, 2004.

[25] W. Yong, L. Zhanhuai, and Z. Yang. Mining sequential association-rule for improving web document prediction. *ICCIMA'05*, pages 146–151, 2005.

[26] Q. Zhao, S. S. Bhomick, and L. Gruenwald. Wam miner: In the search of web access motifs from historical web log data. *CIKM'05, Germany*, pages 421–428, 2005.

[27] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Machine Learning Research*, 4:1001–1037, 2003.

[28] J. Zhu, J. Hong, and J. G. Hughes. Using markov models for web site link prediction. *HT'02, USA*, pages 169–170, 2002.