Web Page Prediction using Hybrid Model

Shreya Dubey
School of Information Technology
RGPV
Bhopal(M.P.), India
shreya.dubey07@gmail.com

Nishchol Mishra
School of Information Technology
RGPV
Bhopal(M.P.),India

Abstract— **One of the most important internet challenges in coming years will be the introduction of intelligent services and a more personalized environment for user. In this paper web page prediction is presented. We use several classification techniques, namely, Support Vector Machines (SVM), Association Rule mining (ARM), and Markov model in WWW prediction. We proposed a hybrid model by combining two or more of them using Dempster's rule [2] to enhance the efficiency of prediction Model.**

Keywords- Markov Model, Support Vector Machine(SVM), Association Rule Mining(ARM), Dempster Shafer Theory, and Hybrid Model.

I. Introduction

Web surfing prediction is a key aspect of web data Management and mining. Surfing prediction is an important research area on which many application improvements depend. Application such as latency reduction, web search and personalization systems utilize surfing prediction to improve their performance. This in turn improves application such as E-commerce, Social Networking and knowledge management.

There are several problems with current state-of-the-art solutions. First, the predictive accuracy using a proposed solution such as a Markov model is low; for example, the minimum training accuracy is 41%. Second, Prediction using association rule mining (ARM) and Longest repeating subsequence (LRS) pattern extraction is done by choosing the path with highest probability in the training set; hence, any new surfing path is misclassified because the probability of such a path occurring in training set is zero. Third, the sparse nature of user sessions used in training can result in unreliable predictors. Finally, many of the previous methods have ignored domain knowledge as means of the improving prediction. Domain knowledge plays a key role in improving predictive accuracy because it can be used to eliminate irrelevant classifiers during prediction or reduce their effectiveness by assigning them lower weights.

WWW prediction is a multiclass problem, and prediction can resolve into many classes, most multiclass technique, such as one-vs-one and one-vs-all, are based on binary classification. Prediction is required to check any new instance against all classes. In WWW prediction the number of classes is very large (11,700 classes in our experiment). Hence, prediction is low because it fails to choose the right class. For a given instance, domain knowledge can be used to eliminate irrelevant classes [1].

We use several classification techniques, namely, Support Vector Machine (SVM), Artificial Neural Network (ANN), Association Rule Mining (ARM) and Markov model in WWW prediction. We proposed a hybrid model by combining two or more of them using Dempster's rule [2]. The Markov model is a powerful technique for predicting seen data; however, it cannot predict unseen data. On other hand, SVM is powerful technique, which can predict not only for seen data but also for unseen data. However, when dealing with too many classes or when there is a probability that one instance may belong to many classes, SVM predictive power may decrease because such examples confuse the training process. To overcome these drawbacks with SVM, we extract domain knowledge from training set and incorporate this knowledge in training set, to improve prediction accuracy of SVM by reducing number of classifier during prediction.

ANN is also powerful technique, which can predict not only for seen data but also for unseen data. ANN has similar shortcoming as SVM when dealing with too many classes or when there is a probability that one instance may belong to many classes. Furthermore, the design of ANN becomes complex with large number of input and output nodes. To overcome these drawbacks with ANN, we employ domain knowledge from the training set and incorporate this knowledge in testing set by reducing the number of classifiers to consult during prediction. This improves prediction accuracy and reduces prediction time.

II. RELATED WORK

A number of researchers attempt to improve the web page access prediction precision or coverage by combining different recommendation framework. For instance many papers combined clustering with association rules (Lai and Yang 2000,Liu et al.2001) [2]. Lai & Yang (2000) 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. (2001) 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[3]. 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 (Cadez et al. 2003, Zhu et al. 2002, Lu et al. 2005)[4]. Cadez et al. (2003) 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 et al. (2002) 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. (2005) 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 firrst 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 (Kim et al. 2004). Kim et al. (2004) 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. (2004) 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.

III. PROPOSED WORK

A. Feature Extraction

The available source of training data is the user’s sessions, which are sequence of pages that users visit within a period of time. In order to improve the predictive ability using different classification techniques, we need to extract additional features besides the Page’s IDs. In mining the web, the only source of training example is the logs that contain the sequence of pages or clicks users have visited or made, time, date, and period of time a user stays in each page. Many model such as Markov model, ARM, and our approach, apply a sliding window on the user sessions to make training instances the same length [5]. If we apply a sliding window, we will end up with many repeated instances that will participate the probability calculation. Furthermore, the page IDs in the user sessions is nominal attributes.

Nominal attributes have no internal structure and take one of finite number of possible values [6]. Many data mining techniques, such as SVM, require continuous attributes because they use a metric measure in their computations (dot product in case of SVM). In our implementation, we use bit vectors to represent the page IDs. We keep only the index of the page ID in vector and its numerical value, if that value is not zero. Missed attributes are assumed to have zero values.

To extract more knowledge from the user sessions, we use what we call frequency matrix in fig:-1.

	1	2	….	N
1	0	Freq(1,2)	….	Freq(1,N)
2	Freq(2,1)	0	….	Freq(2,N)
….	Freq(….,1)	Freq(….,2)	….	Freq(….,N)
N	0	Freq(N,2)	….	0

The first row and column represent the enumeration of web page IDs. Each cell in matrix represents the number of times (frequency) users have visited two pages in a sequence. Freq(x, y) is the number of time user have visited page y after page x. For example, cell (1, 2) contains the number of times users have visited page 2 after page 1. Note that freq(1,2) is not necessarily equals to freq(2,1), and freq(x,x) is always zero.

In this research, we apply a sliding window of size N to break long user sessions into N-size sessions. In our implementation, we use a sliding window of sizes 3 to 7. To further elaborate on sliding window concept, we present following example.

Example:- Suppose we have a user session A=<1,2,3,4,5,6,7> is the sequence of pages a user have visited. Suppose, also, that we use a sliding window of size 5. We apply feature extraction to A=<1,2,3,4,5,6,7> and end with the following user sessions of 5 page length: B=<1,2,3,4,5> , C=<2,3,4,5,6> and D=<3,4,5,6,7>. Note that the outcome or label of the sessions A,B,C and D are 7,5,6 and 7, respectively. This way, we end up with the following four user sessions: A, B, C, and D. In general, the total number of extracted sessions using a sliding window of size w and original session of size A is |A|-w+1.

B .Dempster Rule for Evidence Combination

Dempster's is a well known method for aggregating many different bodies of evidence in the same reference set. Suppose we want to combine evidence for a hypothesis C. In www prediction, C is assignment of a page during prediction for a user session. For example, what is the next page a user might visit after visiting p_1, p_3, p_4 and p_{10}? C is a member of $2^{\Theta,}$ that is . the power set of Θ, where Θ is our *frame of discernment*. A frame of discernment Θ is an exhaustive set of mutually exclusive elements (hypothesis, propositions). All of the elements in this power set, including the element of Θ, are propositions. Given two independent sources of evidence, m_1 and m_2, Dempster's rule combines them in the following frame:

$$m_{1,2}(C)= \frac{\sum_{A,B\sqsubseteq\theta,A\cap B=C} m_1(A)m_2(B)}{\sum_{A,B\sqsubseteq\theta,A\cap B\neq\emptyset} m_1(A)m_2(B)} \qquad \text{Eq.(1)}$$

Here A and B are supersets of C, but they are not necessarily proper subsets; that is, they may be equal to C or to the frame of discernment Θ. The independent sources of evidence m_1 and m_2 are functions (also known as a *mass of belief*) that assign a coefficient between 0 and 1 to different parts of 2^{Θ}. m_1 (A) is the portion of belief assigned to A by m_1, $m_{1,2}$(C) is the combined Dempster-Shafer probability for hypothesis C. To illustrate Dempster-Shafer theory, we present the following example,

EXAMPLE:- Consider a web site that contains three separate web pages B,C and D. Each page has a hyperlink to the two other page. We are interesting in predicting next web page (i.e., B,C or D) a user visits after the he or she visited several pages. We may form the propositions, which correspond to proper subset of Θ:

P_B – The user will visit Page B.
P_C – The user will visit Page C.
P_D – The user will visit Page D.
P_B,P_C – The user will visit either Page B or Page C.
P_D, P_B – The user will visit either Page D or Page B.
P_D, PC – The user will visit either Page D or Page C.

With these propositions, 2^{Θ} would be consisting of following:

$2^{\Theta} = \{\{P_D\},\{P_B\},\{P_C\},\{P_D,P_C\},\{P_B,P_C\},\{P_D,P_B\},\{P_B,P_C,P_D\},ø\}$

In many applications, basic probabilities for every proper subset of Θ may not be available. In these cases, a nonzero m (Θ) accounts for all those subsets for which we have no specific belief. Because we are expecting the user to visit only one web page (it is impossible to visit two pages at same time), we have positive evidence for individual pages only, that is,

m(A)>0:A€$\{\{P_D\},\{P_B\},\{P_C\}\}$

The uncertainty of the evidence m (Θ) in this scenario is as follows:

m (Θ) = 1- $\sum_{A\subset\theta} m(A)$

In equation (1), the numerator accumulates the evidence that supports a particular hypothesis, and the denominator conditions it on the total evidence for those hypotheses supported by both sources. Applying this combination formula to preceding example, assuming we have two bodies of evidence, namely, SVM and Markov Model would yield

$$m_{svm,markov}(P_B) = \frac{W}{\sum_{A,B\sqsubseteq\theta,A\cap B\neq\emptyset} m_{svm}(A)m_{markov}(B)}$$

$w=m_{svm}(\{P_B\})m_{markov}(\{P_B\}+m_{svm}(\{P_B\})m_{markov}(\{P_B,P_C\})+ m_{svm}(\{P_B\})m_{markov}(\{P_B,P_D\})$.

C. Using Dempster-Shafer theory in WWW Prediction

We have three sources of evidence: the output of SVM, ANN and Markov Model. These models operate independently. Furthermore, we assume that for any session X for which it does not appear in training set, the Markov prediction is zero. If we use Dempster's rule to combine SVM and Markov model as our body of evidence, we get the following equation:

$m_{svm,markov}(C) = \frac{\sum_{A,B\sqsubseteq\theta,A\cap B=C} m_{svm}(A)m_{markov}(B)}{\sum_{A,B\sqsubseteq\theta,A\cap B\neq\emptyset} m_{svm}(A)m_{markov}(B)}$ Eq.(2)

In the case of WWW prediction, we can simplify this formulation because we have only beliefs for singleton classes (i.e., the final prediction is only one web page, and it should not have more than one page) and the body of evidence itself (m(Θ)). This means that for any proper subset A of Θ for which we have no specific belief, m(A)=0.

For example, based on the example discussed earlier, we would have the following terms in the numeration of Eq.(2):

$m_{svm}(\{P_B\})m_{markov}(\{P_B\})$,
$m_{svm}(\{P_B\})m_{markov}(\{P_B,P_C\})$,
$m_{svm}(\{P_B\})m_{markov}(\{P_B,P_D\})$,
$m_{svm}(\{P_B\})m_{markov}(\Theta)$,
$m_{markov}(\{P_B\})m_{svm}(\{P_B,P_C\})$,
$m_{markov}(\{P_B\})m_{svm}(\{P_B,P_D\})$,
$m_{markov}(\{P_B\})m_{svm}(\Theta)$

Because we have nonzero basic probability assignments for only the singleton subsets of Θ and the Θ itself, this means that

$m_{svm}(\{P_B\})m_{markov}(\{P_B\})>0$
$m_{svm}(\{P_B\})m_{markov}(\{P_B,P_C\}) = 0$, since $m_{markov}(\{P_B,P_C\}) = 0$
$m_{svm}(\{P_B\})m_{markov}(\{P_B,P_D\})=0$, since $m_{markov}(\{P_B,P_D\})=0$
$m_{svm}(\{P_B\})m_{markov}(\Theta)>0$,
$m_{markov}(\{P_B\})m_{svm}(\{P_B,P_C\})=0$, since $m_{svm}(\{P_B,P_C\})=0$
$m_{markov}(\{P_B\})m_{svm}(\{P_B,P_D\})=0$, since $m_{svm}(\{P_B,P_D\})=0$

After eliminating zero terms, we get simplified Dempster's combination rule as follows:

$m_{svm,markov}(\{P_B\})= \frac{m_{svm}(\{P_B\})m_{markov}(\{P_B\})+m_{svm}(\{P_B\})m_{markov}(\Theta)+m_{markov}(\{P_B\})m_{svm}(\Theta)}{\sum_{A,B\sqsubseteq\Theta,A\cap B\neq\emptyset} m_{markov}(A)m_{svm}(B)}$ Eq.(3)

We use Eq.(3) to rank hypothesis and resolve prediction. In choosing or top n hypotheses, instead of one hypothesis, we can generalize our approach to work as a recommendation system. Because we are interested in ranking the hypotheses, we can further simplify Eq.(3), making the denominator independent of any particular hypotheses as follows:

$m_{svm,markov}(\{P_B\})\alpha \quad m_{svm}(\{P_B\})m_{markov}(\{P_B\})+m_{svm}(\{P_B\})m_{markov}(\Theta)+ \quad m_{markov}(\{P_B\})m_{svm}(\Theta)$ Eq.(4)

The α is the "is proportional to" relationship. Msvm(Θ) and mmarkov(Θ) represent the uncertainty in the bodies of evidence for SVM and Markov model, respectively. The implication of involving the uncertainty of SVM and Markov model in Eq.(4).

The higher uncertainty value of one model is, the more credit or weight is given to the other model. For example, if an unseen path B is presented , the Markov model fails to predict next web page; hence, $m_{markov}(\Theta)=1$, $m_{markov}(\{P_B\})=0$, $m_{svm}(\Theta)m_{markov}(\{P_B\})=0$, and only SVM prediction, $m_{svm}(\{P_B\})$, will resolve the prediction.

We compute msvm(Θ) and mmarkov(Θ) in Eq.(4) as follows. For SVM, we use the margin values for SVM to compute the uncertainty. Uncertainty is computed on the basis of the maximum distance of training examples from the margin as in Eq.(5),(6) and (7).

$m_{svm}(\Theta)=\frac{1}{\ln(e+svm_{margin})}$ Eq.(5)

$m_{ANN}(\Theta) = \frac{1}{\ln(e+ANN_{margin})}$ Eq.(6)

$m_{markov}(\Theta) = \frac{1}{\ln(e+Markov_{Probability})}$ Eq.(7)

In Eq.(5) svm_{margin} is the maximum distance of training examples from the margin, and e is Euler's number.

D. Algorithm: WWW prediction using hybrid model

Input:
S← User session Data
Output:
Y_i: Next page prediction for testing session i.
Begin

```
1. S←Apply-Feature-Extraction(S)
2. svm-models←Train svm(S)// train SVM using one vs all.
3. Svm-prob-model←Map-SVM-model (svm-models)//map SVM output to a probability
4. Svm-uncertainty←ComputeUncertainty(SVM)// Eq.(5)
5. Construct Markov model
6. Markov-uncertainty←compute-Uncertainty(Markov)// see Eq.(7)
7. For each testing session X in S, do
   7.1 Compute and output SVM probabilities for different pages
   7.2 Compute and output Markov probability for different pages.
   7.3 Compute using Eq.(4) and the final prediction $Y_X$
```

IV. EXPERIEMENTAL EVALUATION

In this section, we present experimental results to evaluate the performance of our algorithm. For our experiments, the four data sets were relied upon and all preprocessing tasks including session identification and categorization were also used.

Figure 1, Figure 2, Figure 3 and Figure 4 depict better Web page access prediction accuracy for all four data sets by integrating Markov model, association rules and clustering (Hybrid) than by employing the clustering, Markov model and association rules individually. Prediction accuracy was computed 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 Markov model prediction performed on the closest cluster.

3. The frequency of the item is also determined in that particular cluster and *c* is calculated for the new instance using $Z\alpha/2$ value to determine if it belongs to the majority class.

4. If the prediction results in a state that does not belong to the majority class, association rules are used for prediction, otherwise, Markov model accuracy is employed.

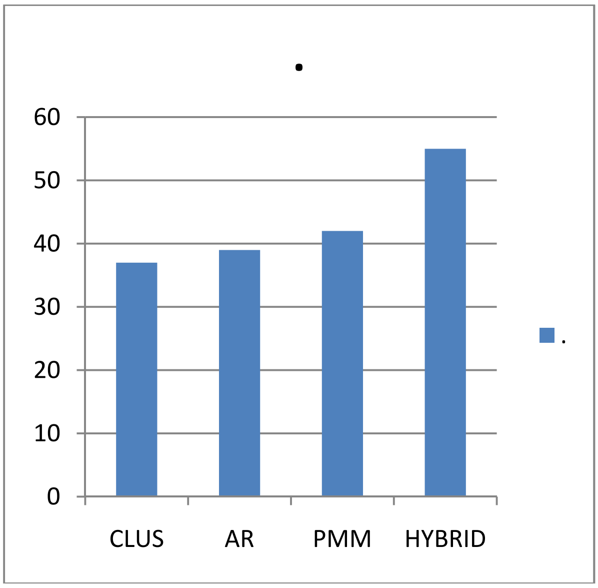

Figure 1: Accuracy of Clustering, AR, PMM, and IPM for data set D1.

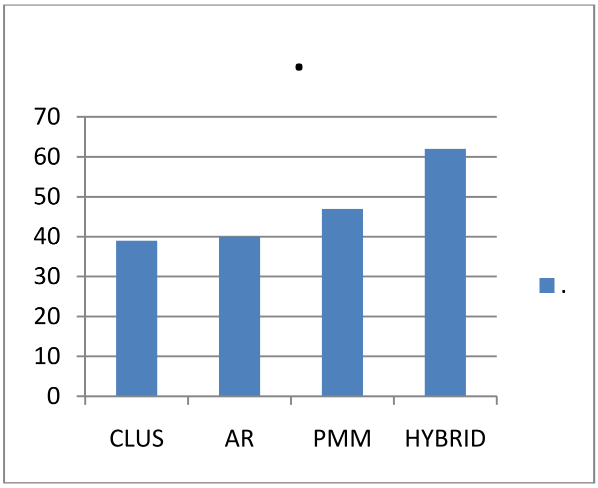

Figure 2: Accuracy of Clustering, AR, PMM, and IPM for data set D2.

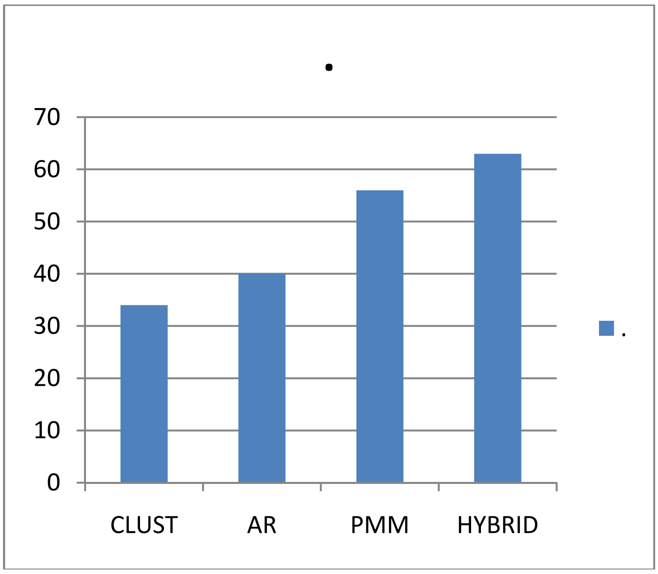

Figure 3: Accuracy of Clustering, AR, PMM, and IPM for data set D3.

The above Figures display that IPM results in better prediction accuracy than any of the other techniques individually using experiments based on all four data sets. They also reveal that the increase in accuracy depends on the actual data set used. For instance, D1 and D4 reveal a more significant accuracy increase using IPM over the individual models. On the other hand, D2 and D3 display a more consistent improvement in prediction accuracy.

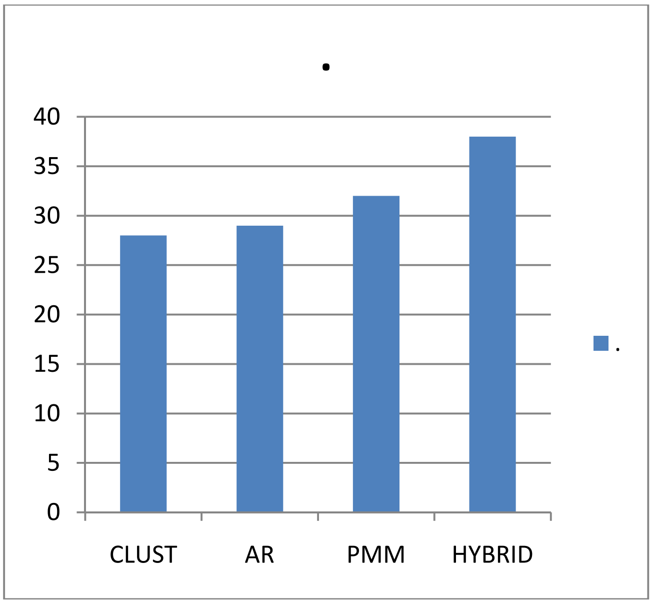

Figure 4: Accuracy of Clustering, AR, PMM, and IPM for data set D4

V. CONCLUSION

This paper improves the Web page access prediction accuracy by integrating all three prediction models: Markov model, Clustering and association rules according to Dempster Shafer Theory. Our model, Hybrid Model, integrates the three models using lower order 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 rule mining on the rest. User sessions are first clustered using some meaningful measures. Then Markov models are implemented using the outcome of the clustering effectuation. Association rules are used for prediction only in the case of certain stipulations. IPM proves to outperform all three models implemented individually, as well as, the IMAM and IMC integrated models when it comes to accuracy. Also, IPM improves the state space complexity of a higher order Markov model.

VI. FUTURE WORK

Future work will include examining other classification techniques for Web page prediction. These may include decision tree as well as nearest-neighbor algorithms. In addition, we also need to approve on our knowledge-based classification so that we can obtain better accuracy as well as reduce false positive and false negatives.

REFERENCES:-

[1] Piton, J. and Pirolli, P., Mining longest repeating subsequence to predict World Wide Web surfing, in Proceeding of the 2nd USENIX Symposium on Internet Technologies and Systems (USITS'99), Boulder, Colorado, October 1999, pp.139-150.

[2] Faten Khalil, Jiuyong Li, Hua Wang 'Integrating Recommendation Models for Improved Web Page Prediction Accuracy'

[3] Fang Liu, Zhengding Lu, Songfeng Lu 'mining Association rule using Clustering'

[4] Cadez et al. 2003, Zhu et al. 2002, Lu et al. 2005 'Clustering with Markov Model'

[5] Brin, S. and Page, L., The anatomy of a Large-scale hypertexual Web Search Engine, in proceedings of the 7th international WWW conference, Brisbane, Australia, 1998, pp.107-117.

[6] Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. Effective personalization based on association rule discovery from web usage data, in proceeding of the ACM workshop on Web Information and Data Management (WIDM01),2001, pp. 9-15.