http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (4): 774-781, 2010
ISSN 1812-5638

Hybrid Web Page Prediction Model for Predicting a User's Next Access

[1]S. Chimphlee, [2]N. Salim, [2]M.S.B. Ngadiman and [1]W. Chimphlee
[1]Faculty of Science and Technology, Suan Dusit Rajabhat University,
228-228/1-3 Sirinthorn Rd, Bangphlad, Bangkok, 10700, Thailand
[2]Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Abstract: The web user sessions are clustered with incorporating the sequence of web page visits. A sequence-based clustering is developed by proposing new sequence representations and new similarity measures. The resulting sequence representation allows for calculation of similarity between web user sessions and then, can be used as input of clustering algorithms. This study proposed a hybrid prediction model (HyMFM) that integrates Markov model, Association rules and Fuzzy Adaptive Resonance Theory (Fuzzy ART) clustering together. The three approaches are integrated to maximize their strengths. A series of experiments was conducted to investigate whether, clustering performance is affected by different sequence representations and different similarity measures. This model could provide better prediction than using each approach individually.

Key words: Web page prediction, web usage mining, markov model, association rules, fuzzy adaptive resonance theory

INTRODUCTION

Predicting a user's next access on a web site has attracted a lot of research work lately due to the positive impact of such prediction on different areas of web based applications (Khalil *et al*., 2007). In all of these applications the goal is the development of an effective and accurate prediction model. The most successful prediction algorithms use historical access data from web access logs, which records the information about all the visits by different users to different web sites and web pages. By having this information, many researchers have designed action systems that use the predictions from a learned model and have developed methods for dealing with specific aspects of web usage mining, like automatically discovering web personalization (Nasraoui and Petenes, 2003), recommender systems (Khalil *et al*., 2008), web prefetching (Yang *et al*., 2003; Alexandros *et al*., 2003), web presending (Li, 2001), design of adaptive web sites (Zhu *et al*., 2002).

The most widely approach is web usage mining that entails many models like Markov models, Association rules and clustering (Srivastava *et al*., 2000). However, there are some challenges with the current state of the art solutions when, it comes to accuracy, coverage and performance. A Markov model is a popular approach to predict what pages are likely to be accessed next (Cadez *et al*., 2003; Khalil *et al*., 2006; Deshpande and Karypis, 2004). A problem that faces Markov model users is the difficulty in identifying the optimal number of Markov model orders which affects the system accuracy, coverage and performance.

The second is Association rules (Agrawal *et al*., 1993). It is based on the relationship of co-occurrence of pages without considering the sequence of them. This makes Association rules generally produce low precision, but high recall in the prediction (Kim *et al*., 2005). Yang *et al*. (2003) 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. As, the advantages of both techniques, Khalil *et al*. (2006) have been proposed the combination of Association rules and Markov model Khalil *et al*. (2006). They used lower order all k-th Markov models to predict the next page to be accessed. In ambiguous predictions, association rules are used to compliment Markov models. The advantage of this

Corresponding Author: Siriporn Chimphlee, Faculty of Science and Technology, Suan Dusit Rajabhat University, 228-228/1-3 Sirinthorn Rd, Bangphlad, Bangkok, 10700 Thailand

combination is that when Markov models are unable to make the prediction, Association rules look further back at the previously visited pages and leads to the most appropriate page for prediction. The main problem is dependent on the length of the web user session and it is difficult to perform the analysis with short user sessions.

Later, Khalil *et al.* (2008) introduce the Integration Prediction Model (IPM) by combining Markov model, Association rules and clustering algorithm together. Then, the prediction is performed on the cluster sets rather than the actual sessions (Vakali *et al.*, 2004; Pallis *et al.*, 2007; Khalil *et al.*, 2007; Yang *et al.*, 2003; Borges and Levene, 2005; Lu *et al.*, 2005a). The IPM integration model is based on the different constraint. The web user sessions first are divided into a number of clusters using k-means clustering algorithm and cosine distance measure. Then, an integration model computes Markov model prediction on the resulting clusters. This algorithm improves the state space complexity because Markov model prediction carried out on the particular clusters as opposed to the whole data set. In the case of state absence in the training data or where, the state prediction probability is not marginal, Association rules are examined more states than Markov model by looking at more history. Lastly, if a new page is presented, the cosine distance is calculated and identifies an appropriate cluster that a new web page should belong to. The integration model has been proved through the experiments that improve the prediction accuracy. Moreover, implementing the prediction model on the clusters achieves better results than on the non-clustered data. Although, a web page access prediction performance was improved, however, it can be seen that their integrated algorithm has a complicated procedures and must repeatedly employ in order to increase their prediction performance.

Another important mention is that most of the algorithms to deal with clustering web user sessions treat sessions as sets of visited pages within a time period and do not consider the sequence of the access. In order to improve the prediction ability using the different prediction techniques, a data representation is a key issue for representing the web user sessions to effectively support in web user session clustering. Different approaches have been adopted to represent the web user sessions for different mining tasks. Existing approaches vary on the web user session representation and similarity computation (Nichele and Becker, 2006). For web user session clustering, many studies have adopted a vector model with binary values representation of web user sessions which cannot represent sequence directly (Nasraoui and Petenes, 2003; De and Krishna, 2004; Rangarajan *et al.*, 2004). This representation is satisfying many existing data mining algorithms and relatively easy to apply. However, the major concern is the lost of page orders which the order of page sequences has significant consequence when computing similarities between session (Wang and Zaïane, 2002; Lu *et al.*, 2005a, b) and users' transitions from one page to another cannot be reflected (Park *et al.*, 2008). Another issue related to the clustering algorithm is the similarity measure. The similarity measures used to compare sessions are simply based on intersection between these sets, such as the cosine measure or the Jaccard coefficient. This similarity measure can only estimated and it is not measure the match between the prototype and the input. It is not adequate measure since, the sequence of events is not taken into account (Wang and Zaïane, 2002).

BACKGROUND OF THE PREDICTION TECHNIQUES

Association rules: Association rule mining technique, which was first introduced by Agrawal *et al.* (1993) is used to find the frequent itemsets of web pages from the user access sequences and constructs a set of rules based on those itemsets. Let, $P = \{P_1, P_2,..., P_m\}$ be a set of m distinct web pages. T is a user session that contains a set of web pages such that $T \subseteq P$, D is a database with different user sessions. An association is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset P$ are sets of web pages and $X \cap Y = \emptyset$. X is called antecedent while, Y is called consequent, the rules means X implies Y. Suppose one of the large itemsets is L_k, $L_k = \{P_1, P_2,..., P_m\}$ association rules with these itemsets are generated in the following way: First, the frequent 1-item is found. Secondly, candidates are generated for frequent 2-items retaining only those with counts greater than or equal to threshold θ. After this, the process is iterated until the algorithm repeatedly obtains frequent 3-items, 4-items etc., up to size k. There are two basic parameters of Association rule mining are support and confidence. Support of an Association rule is defined as the percentage of sessions that contain $X \cup Y$ to the total number of sessions in the database. The count for each web page is increased by one every time the item is encountered in different session T in database D during the scanning process. Confidence of an association rule is defined as the percentage of the number of sessions that contain $X \cup Y$ to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule $X \Rightarrow Y$ can be generated. After all frequent web pages are generated then strong association rules are defined. Once, the frequent itemsets from transactions in a database D have been found, it is straightforward to generate strong association rules from them where, strong association rules satisfy both minimum support and minimum confidence. Using the association rule mining technique, the modification of the predictive model can be done with ease by adjusting the support and confidence values.

Markov models: Markov models are a commonly used method for modelling stochastic sequences with an underlying finite-state structure and were shown to be well-suited for modelling and predicting a user's browsing behavior on a web site (Deshpande and Karypis, 2004). The precision of this technique comes from the consideration of consecutive orders of preceding pages. The goal is to build the user behavioral models that can be used to predict the web page that the user will most likely access next. The input for this problem is the sequence of web pages that were accessed by a user and it is assumed that it has the Markov property. In such a process, the past is irrelevant for predicting the future given knowledge of the present. 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. Assuming that the user has visited l pages then P (p_i/W) is the probability that the user visits page p_i next. The conditional probabilities are commonly estimated by assuming that the process generating sequences of the web pages visited by users follows a Markov process. That is, the probability of visiting a web page p_i does not depend on all the pages in the web session, but only on a small set of k preceding pages, where k << l. Using the Markov process assumption, the web page p_{l+1} will be generated next is given by

$$p_{1+1} = \text{argmax}_{p\in P}\{P(P_{1+1} = p/P_1, P_{1-1},\ldots,P_{1-(k-1)})\} \qquad (1)$$

where, k denotes the number of the preceding pages and it identifies the order of Markov model. The resulting model of this equation is called the kth-order Markov model. In order to use the kth-order Markov model, the learning of p_{l+1} is needed for each sequence of k web pages. Let S^k_j be a state with k as the number of preceding pages denoting the Markov model order and j as the number of unique pages on a web site, S^k_j =<$p_{l-(k-1)}$, $p_{l-(k-2)}$,..., p_l>, by estimating the various conditional probabilities P (P_{l+1} = p/P_l, P_{l-1},.., $P_{l-(k-1)}$). Using the Maximum likelihood principle (Richard, 2000), the conditional probability P (p_i/S^k_j) is computed by counting the number of times sequence S^k_j occurs in the training set and the number of times p_i occurs immediately after S^k_j. The conditional probability is the ratio of these two frequencies, that is:

$$P(p_i / S^k_j) = \frac{\text{Frequency}(<S^k_j, p_i>)}{\text{Frequency}(S^k_j)} \qquad (2)$$

Fuzzy adaptive resonance theory (fuzzy ART) clustering: A Fuzzy ART (Carpenter *et al*., 1991) is an unsupervised self-organizing network, it does not forget previously learned patterns and it deals with real-valued data as an input. A Fuzzy ART network is formed of three layers of neurons: input layer F_0, comparison layer F_1 and output layer F_2. Each layer has M, 2M and N nodes, respectively. The layers are fully interconnected, each neuron being connected to every neuron on the other layer. Two kind of weights are connected each other. Every connection between the two layers is weighted by a number lying between 0 and 1. The bottom-up weights connecting F_1 to F_2 are denoted by b_{ij} and the top-down weights connecting F_2 to F_1 are designed t_{ji}. Input vector a is pre-processed by complement coding in layer F_0. The operation consists on taking the input vector and concatenating it with its complement. The resulting vector is presented to layer F_1. Therefore, the dimension M of layer F_1 is the double of the input vector's dimension. Layer F_1 compares the similarity between the input vector and top-down weight vector and then layer F_2 chooses the node with the maximum competitive signal of bottom-up weight when, an input vector is presented. A representative vector w_j of Fuzzy ART contains both the bottom-up weights and top-down weights. When, the winning node J of F_2 layer is chosen, its top-down and bottom-up weights are updated in the same manner. Here, since both weights have same values, either weight can be considered to be representative pattern. Hence, fast learning is possible by updating only one of both weights as Eq. 3 and 4

$$t^{(new)}_j = \beta(I \wedge t^{(old)}_j) + (1-\beta)t^{(old)}_j \qquad (3)$$

$$b^{(new)}_j = \beta(I \wedge b^{(old)}_j) + (1-\beta)b^{(old)}_j \qquad (4)$$

where, β is learning rate parameter, I is input vector and t_j and b_j is top-down and bottom-up weight vectors associated with node j, respectively.

HYBRID MARKOV FUZZY MODEL

This study proposed the Hybrid Markov Fuzzy Model (HyMFM) which may help to close some obvious gaps. This approach combines the strengths of Markov model, Association rules and Fuzzy Adaptive Resonance Theory in order to achieve higher accuracy, better coverage and overall performance while, keeping the number of computations to a minimum. A HyMFM model first devises representation schemes that suitable for capturing sequence information. This is followed by two similarity measures and a description of the proposed HyMFM clustering.

Extraction of feature matrix for the data representation: In order to capture sequence information, each web user session is represented as a transition matrix of size n×m, where, n is the number of web pages, referred to as session matrix and a set of n initial probabilities describing how likely that user is will begin a session in a given web page. Let, there be m sessions and the web user sessions be $S = \{s_1, s_2, ... s_m\}$. Let, P be the state and defines as a set of web pages accessed by the users, $P = \{P_1, P_2, ..., P_n\}$, each $s_i \in S$ is a non-empty subset of P. S_i is ith web user session (Eq. 5). The previously visited web pages, the current web page and the future visited web pages are $(X_1 = p_1) \rightarrow (X_2 = p_2) \rightarrow ... \rightarrow (X_{k-1} = p_{k-1})$, $(X_k = p_k)$ and $(X_n = p_n)$, respectively.

$$S_i = \{P_1, P_2, P_3, P_4 ...\} \qquad (5)$$

A first hybrid Markov model (1st HyMM) occurs if the next visited web page is dependant only on a current web page in the same session. A previous web page and next visited web page are not only the adjacent web page. The rule from a first hybrid Markov model can be expressed as $(X_k = p_k) \Rightarrow (X_n = p_n)$, where, $(X_k = p_k)$ is a current state and $(X_n = p_n)$ is the future visited state, the transition probability of going from state $(X_k = p_k)$ to state $(X_n = p_n)$ is generated as Eq. 6:

$$P[(X_k = p_k) \Rightarrow (X_n = p_n)] = P_{p_k \Rightarrow p_n} \qquad (6)$$

where, $P_{pk \rightarrow pn}$ is the transition probability that a user access from web page p_k to web page p_n and this probability is independent of i. The maximum likelihood can be used to find the transition probability as the number of times a transition from state $(X_k = p_k)$ to state $(X_n = p_n)$ is observed divided by the number of observation of state $(X_k = p_k)$.

$$P(p_k \Rightarrow p_n) = \frac{\mathrm{Num}(p_k \Rightarrow p_n)}{\mathrm{Num}(p_k)} \qquad (7)$$

where, Num (p_k) is total number of page request for state $(X_k = p_k)$, Num $(p_k \Rightarrow p_n)$ is the total number of page requests for state $(X_k = p_k)$ before state $(X_n = p_n)$, which the access of page $(X_n = p_n)$ does not come immediately after page $(X_k = p_k)$ in the same session.

This approach can capture more memory in the process. If the constraint of previous web pages is extended to k, it is called a k-Hybrid Markov Model (kth-HyMM), the transition probability of going from $(X_1 = p_1)$ $(X_2 = p_2)$ $\rightarrow ... \rightarrow (X_k = p_k)$ to state $(X_n = p_n)$ is generated as Eq. 8 and 9:

$$P[(X_1 = p_1) \rightarrow (X_2 = p_2) \rightarrow ... \rightarrow (X_k = p_k) \Rightarrow (X_n = p_n)] \qquad (8)$$

$$P(p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_k \Rightarrow p_n) = \frac{\mathrm{Num}(p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_k \Rightarrow p_n)}{\mathrm{Num}(p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_k)} \qquad (9)$$

In this study, a 1st HyMM algorithm is used to construct the transition probability matrix and its matrix is used as the input of HyMFM clustering instead of input vectors found in Fuzzy ART. The advantage of this representation is to eliminate the problem of unequal length of the sequence. In Eq. 10, the matrix S_i shows the structure of such a transition matrix, where, $p_{(pk \rightarrow pn)}$ is the probability for a user to transition from state $(X_k = p_k)$ to state $(X_n = p_n)$ in a session i. The entries in the matrix represent the probability of transitioning between any two states in the same session.

$$S_i = \begin{bmatrix} p_{p_1 \Rightarrow p_1} & p_{p_1 \Rightarrow p_2} & \cdots & p_{p_1 \Rightarrow p_n} \\ p_{p_2 \Rightarrow p_1} & p_{p_2 \Rightarrow p_2} & \cdots & p_{p_2 \Rightarrow p_n} \\ \cdots & & & \\ p_{p_n \Rightarrow p_1} & p_{p_n \Rightarrow p_2} & \cdots & p_{p_n \Rightarrow p_n} \end{bmatrix} \qquad (10)$$

Sequence similarity measures: The similarity measure is used to evaluate the extent to which one web user session matches from another. This study develops two similarity measures: Matrix norm similarity and Matrix distance similarity, to enhance the Fuzzy ART algorithm and the clustering performances of these similarity measures are compared. The detail is described as follows.

Matrix norm similarity: Matrix norm similarity is the modification of the choice function of original Fuzzy ART. This similarity computes the compatibility between an input and a prototype and is defined as:

$$T_k = \frac{\|I \wedge w_k\|}{\alpha + \|w_k\|} \qquad (11)$$

where, I is an n×n matrix of input and each component in $I \in [0, 1]$, n is number of web pages, the norm $\|.\|$ is the sum of its components in the matrix and is defined by:

$$\|p\| = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij}$$

$\wedge$ is the fuzzy AND operator, $(I \wedge W)_{ij} = \min(I_{ij}, W_{ij})$ and α is the choice parameter, k is number of k^{th} output clusters in the output layer.

Matrix norm distance: This similarity modifies the fuzzy AND operator of the choice function by using the difference of two transition probability matrices and called Matrix norm distance. The Matrix norm distance is defined as:

$$T_k = \frac{\|I - w_k\|}{\alpha + \|w_k\|} \tag{12}$$

where, ||.|| is the Euclidean norm

$$\|A\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2}$$

(I-W) is the difference of an n×n matrix of the prototype and the input:

$$(I - w) = \sqrt{\sum_{i=1}^{n}\sum_{i=1}^{n}|I_{ij} - w_{ij}|^2}$$

Hybrid markov fuzzy model (HyMFM) clustering: HyMFM model is developed by proposing a new representation scheme and new similarity measures to enable the application of Fuzzy ART neural network method, which is described in previous section. Figure 1 and 2 show the architecture of the HyMFM clustering uses Matrix norm similarity (HyMFM-1) and the architecture of the HyMFM clustering uses Matrix norm distance (HyMFM-2). The model for the incremental learning of these two clustering algorithms is similar manner as Fuzzy ART.

The network consists of an input and output layer. A session matrix s_i which represents the access pattern of web user session i is presented at the layer F_1. An input data s_i is the web user session that represented by the session matrix of n×n dimensions and n is the number of web pages. The nodes at the Recognition layer F_2 represent the clusters formed. In step 1, all weight matrices, store session matrices of output nodes, are initialized to values of one. The number of possible clusters can be chosen arbitrarily large; the remaining output nodes are said to be uncommitted after feeding all inputs to the network. The three parameters are required to specify: choice parameter (α), learning parameter (β) and vigilance parameter (ρ). The choice parameter must be greater than zero, $\alpha>0$; $\beta \in[0, 1]$, $\rho\in[0, 1]$. Step 2 involves reading each session matrix, I. Each session matrix presents at the input layer F_1, where each component in $I\in[0, 1]$. The session matrix activates a node in the out put layer F_2. The F_2 layer consists of a variable number of nodes corresponding to the number of clusters. The F_2 layer reads out the top-down expectation to F_1 layer, where the winner is compared with the session matrix. Step 3 computes the choice function, T_k for each output node, k = 1 to K (expected maximum number of clusters). In this step, the sequence similarity measures can be replaced by any of two similarity measures that were explained. From these, the output node k, with the largest T_k values, is selected in step 4. It is necessary to check whether this best-match meets the specified level of similarity, called vigilance parameter (ρ). The vigilance parameter determines the degree of mismatch that is to be tolerated. Step 5 computes the similarity measure. If the similarity level is greater than the specified ρ value, then this matrix is assigned under this cluster. The best-matching exemplar is updated by modifying the associated weight matrices in step 7. If it does not pass the similarity test, F_1 layer sends a reset burst to F_2 layer, which shut off the current node and choose another uncommitted node. Once the network stabilizes, the top-down weights corresponding to each node in F_2 layer represented a new prototype matrix for that node is created. Thus, the process is repeated until all inputs are processed.

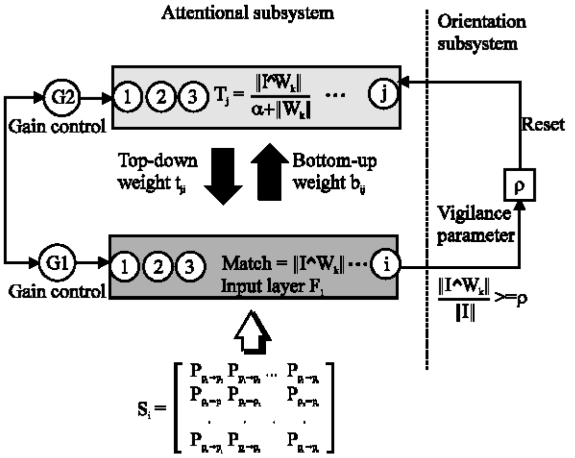

Fig. 1: Architecture of the HyMFM clustering uses matrix norm similarity (HyMFM-1)

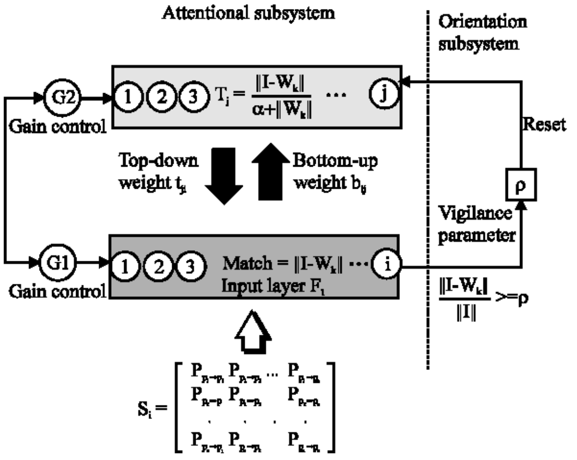

Fig. 2: Architecture of the HyMFM clustering uses Matrix norm distance (HyMFM-2)

THE EXPERIMENTAL RESULTS

The experiments compared the relative significance of each effect across different representation schemes and different similarity measures. The data representations of this experiment are Markov representation and Hybrid Markov representation. The first experiment uses Markov representation and Hybrid Markov representation with the matrix norm similarity measure, called Markov Fuzzy similarity model (referred to as MFM-1) and Hybrid Markov Fuzzy similarity model (referred to as HyMFM-1). The second experiment uses Markov representation and Hybrid Markov representation with the matrix norm distance measure, called Markov Fuzzy distance model (referred to as MFM-2) and Hybrid Markov Fuzzy distance model (referred to as HyMFM-2), respectively.

As, the experimental results that have compared between four Hybrid prediction models in Fig. 3 and 4, it can be seen that as the vigilance values is increased, the prediction performance of all Hybrid prediction models including the accuracy and coverage are increased. In particularly, for the vigilance value from 0.7, the performance of HyMFM-1 and HyMFM-2 are not much change. Due to the number of clusters, as increases the vigilance up to 0.7, the numbers of clusters are not very different. Thus, it can be implied that the vigilance value do not affect to the performance of HyMFM-1 and HyMFM-2 at the vigilance value equal to 0.7 or more.

Despite, the efficient prediction accuracy results that were achieved using HyMFM-2 model, it was necessary to compare the prediction accuracy results to the other algorithm. This study compared HyMFM-2 results to those of Association, Markov model, Hybrid Markov model and Integrated Prediction Model (IPM). It was observed that the accuracy of HyMFM-2 model is much higher than the accuracy of any other algorithms (Fig. 5). This revealed that the accuracy is depend on the actual data used. For HyMFM-2 model, the web user sessions were clustered into groups with similar patterns. This is assumed to be more homogeneous than the whole data set since, HyMM is significant for sub-group. As a consequence, performing a HyMM analysis on functionally related sessions leads to more accurate prediction than performing such analysis on the whole data set. With regard to IPM model, the web user sessions first are divided into a number of clusters using k-means clustering algorithm. The number of clusters must specify in advance. A high number of clusters would result in lower accuracy while low number of clusters would result in higher accuracy. There is no good way to a priori select optimal number of clusters. The difficulty of clustering with k-means is that if fails to identify clusters

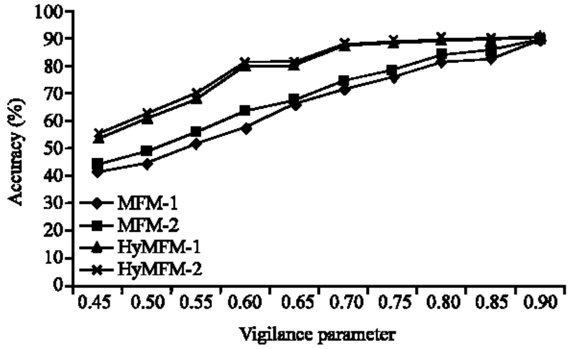

Fig. 3: Comparison of accuracy of four Hybrid clustering techniques: HyMFM-1, HyMFM-2, MFM-1 and MFM-2 by varying the value of the vigilance parameter

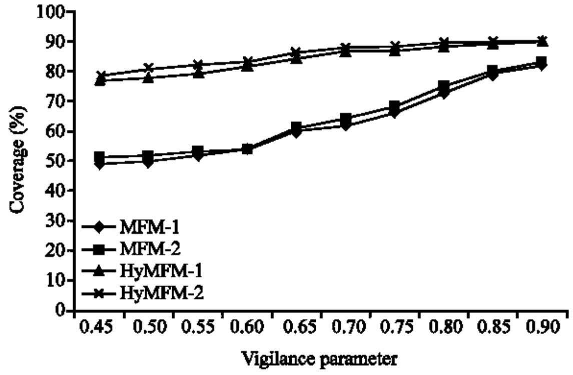

Fig. 4: Comparison of coverage of four Hybrid clustering techniques: HyMFM-1, HyMFM-2, MFM-1 and MFM-2 by varying the value of the vigilance parameter

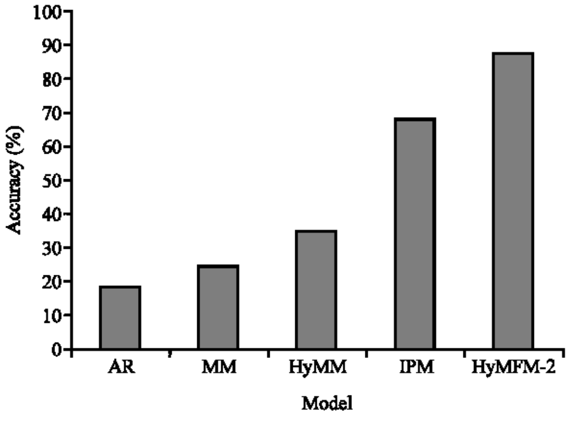

Fig. 5: Comparison of accuracy of AR, MM, HyMM, IPM and HyMFM-2

with large variation in sizes. It takes time to trial error until get optimal number of clusters and increases the number of computation.

Based on Fig. 3-5, the HyMFM-2 model showed a larger increase in prediction performance. Another observation is the representation of the web user session and the similarity measures affect to the prediction performance. In particularly, the representation showed more affect than the similarity measure. As it can be seen from the results, HyMFM-2 model results in highest prediction performance than MFM-2 in term of accuracy and coverage. This is because HyMFM-2 use first Hybrid Markov model to represent the web user session while MFM-2 use first Markov model to represents the web user session. Using first Markov model is over restrict while first Hybrid Markov model is flexible and keeps the orders of web pages. With regard to HyMFM-1 and HyMFM-2, the prediction performance of HyMFM-2 is better than HyMFM-1. It was confirmed that using matrix norm distance measure along with the Hybrid Markov representation showed significantly better prediction performance than using matrix norm similarity along with Markov representation and Hybrid Markov representation.

Finally, the experimented results are concluded that using HyMFM-2 clustering with the vigilance value of 0.7 lead to good clustering results while keeping the number of clusters to a minimum with a high degree of accuracy.

CONCLUSIONS

This study presented a Hybrid Markov Fuzzy Models (HyMFM) that are obtained by integrating the advantages of all three prediction models: Markov model, Association rules and Fuzzy Adaptive Resonance Theory (Fuzzy ART). HyMFM algorithm was developed for the web user sessions clustering by proposing the new sequence representations and the new similarity measures in incremental learning of Fuzzy ART control structure. A web user session was represented into the transition matrix representation, referred to as session matrix, which is constructed based on a transition matrix of a first Hybrid Markov model. Both elements fit well into the design of this thesis and the clustering task which the web user sessions are treated as order sets of accesses. Consequently, the new similarity measures were developed to enable the application of Fuzzy ART clustering. This study defined two new similarity measures: Matrix norm similarity and Matrix distance similarity. These measures alleviate the overestimation problem in Fuzzy ART algorithm which use the city-block distance metric as the similarity between input and prototypes. Thus, the web user sessions were clustered into groups with similar patterns in during the training phase and when it is confronted by a new input, it produces a response that indicates which cluster the pattern belongs to and then HyMM applied to each cluster.

HyMFM-2 model proved to outperform all two models implemented individually, as well as, a HyMM and IPM model when it comes to accuracy and coverage. The advantages of a HyMFM-2 model include that it can solve arbitrarily complex clustering problem, it converges quickly to a solution (within a few presentations of the list of input/output patterns belonging to the training set), it has the ability to recognize novelty in the input patterns presented to it, it can operate in an on-line fashion (new input/output patterns can be learned by the system without re-training with the old input/output patterns) and lastly, it produces answers that can be explained with relative ease.

ACKNOWLEDGMENT

This study was supported by Suan Dusit Rajabhat University, Thailand and Universiti Teknologi Malaysia.

REFERENCES

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, May 25-28, New York, USA., pp: 207-216.

Alexandros, N., N. Ros, D. Katsaros and M. Yannis, 2003. A data mining algorithm for generalized web prefetching. IEEE Trans. Knowledge Data Eng., 15: 1155-1169.

Borges, J. and M. Levene, 2005. A clustering based approach for modeling user navigation with increased accuracy. Proceedings of the Second International Workshop on Knowledge Discovery from Data Streams (IWKDDS) in conjunction with PKDD 2005, Porto, Portugal, Outubro.

Cadez, I.V., D. Heckerman, C. Meek, P. Smyth and S. White, 2003. Model based clustering and visualization of navigation patterns on a web site. Data Mining Knowledge Discovery, 7: 399-424.

Carpenter, G.A., S. Grossberg and D.B. Rosen, 1991. Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks.

De, S.K. and P.R. Krishna, 2004. Clustering web transactions using rough approximation. Fuzzy Sets Syst., 148: 131-138.

Deshpande, M. and G. Karypis, 2004. Selective markov models for predicting web page accesses. ACM Transact. Internet Technol., 4: 163-184.

Khalil, F., J. Li and H. Wang, 2006. A framework of combining markov model with association rules for predicting web page accesses. Proceedings of the 5th Australasian Conference on Data Mining and Analystics, (AusDM'06), Australian Computer Society, Inc., pp: 177-184.

Khalil, F., J. Li and H. Wang, 2007. Integrating markov model with clustering for predicting web page accesses. Proceedings of the 13th Australasian World Wide Web Conference (AusWeb 2007), June 30-July 4, Coffs Harbor, Australia, pp: 1-26.

Khalil, F., J. Li and H. Wang, 2008. Integrating recommendation models for improved web page prediction accuracy. Proceedings of the 31th Australasian Computer Science Conference, (ACSC'08), Wollongong, NSW, pp: 91-100.

Kim, D.H., I.L. Im and V. Atluri, 2005. A clickstream-based collaborative filtering recommendation model for e-commerce. Proceedings of the 7th IEEE International Conference on E-commerce Technology, July 19-22, IEEE Computer Society, Washington, DC., USA., pp: 84-91.

Li, T., 2001. Web-document prediction and presending using association rules sequential classifiers. Simon Fraser University.

Lu, L., M. Dunham and Y. Meng, 2005a. Discovery of significant usage patterns from clusters of clickstream data. Proceeding of WebKDD 2005: KDD Workshop on Web Mining and Web Usage Analysis, in Conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), Aug. 21-24, Chicago, IL., pp: 1-12.

Lu, L., M. Dunham and Y. Meng, 2005b. Mining significant usage patterns from clickstream data. Proceedings of the Advances in Web Mining and Web Usage Analysis 7th International Workshop on Knowledge Discovery on the Web, (WebKDD'05), Springer Berlin/Heidelberg, pp: 1-17.

Nasraoui, O. and C. Petenes, 2003. Combining web usage mining and fuzzy inference for website personalization. Proceedings of WebKDD 2003-KDD Workshop on Web Mining as a Premise to Effective and Intelligent Web Applications, August 2003, Washington DC. USA., pp: 37-46.

Nichele, C.M. and K. Becker, 2006. Clustering web sessions by levels of page similarity. Proceedings of the 10th Pacific-Asia Conference, (PAKDD'6), Springer Berlin/Heidelberg, pp: 346-350.

Pallis, G., L. Angelis and A. Vakali, 2007. Validation and interpretation of web users, sessions clusters. Int. J. Inform. Process. Manage., 43: 1348-1367.

Park, S., N.C. Suresh and B.K. Jeong, 2008. Sequence based clustering for web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. Data Knowledge Eng., 65: 512-543.

Rangarajan, S.K., V.V. Phoha, K.S. Alagarri, R.R. Selmic and S.S. Iyengar, 2004. Adaptive neural network clustering of web users. IEEE Comput., 7: 34-40.

Richard, O.D., 2000. Pattern Classification. 2nd Edn., Wiley Interscience, New York, ISBN: 978-0-471-05669-0.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorat., 1: 12-23.

Vakali, A., J. Pokorny and T. Dalamagas, 2004. An overview of web data clustering practices. Proceedings of the Workshops on Current Trends in Database Technology, (EDBT'04), Springer-Verlag Berlin Heidelberg, pp: 597-606.

Wang, W. and O.R. Zaïane, 2002. Clustering web sessions by sequence alignment. Proceedings of the 13th International Workshop on Database and Expert Systems Applications, (DEXA'02), IEEE Computer Society, pp: 394-398.

Yang, Q., J.Z. Huang and M. Ng, 2003. A data cube model for prediction based web prefetching. Intelligent Inform. Syst., 20: 11-30.

Zhu, J., J. Hong and J.G. Hughes, 2002. Using Markov Chains for Link Prediction in Adaptive Web Sites. Springer-Verlag, Berlin Heidelberg.