Improving Web Page Prediction Using Default Rule Selection

Thanakorn Pamutha, Chom Kimpan
Faculty of Information Technology
Rangsit University, RSU
Phathumthani, THAILAND

Siriporn Chimplee
Faculty of Science and Technology
Suan Dusit Rajabhat University, SDU
Bangkok, THAILAND

Parinya Sanguansat
Faculty of Engineering and Technology
Panyapiwat Institute of Management, PIM
Nonthaburi, THAILAND

***Abstract*—Mining user patterns of web log files can provide significant and useful informative knowledge. A large amount of research has been done in trying to predict correctly the pages a user will most likely request next. Markov models are the most commonly used approaches for this type of web access prediction. Web page prediction requires the development of models that can predict a user's next access to a web server. Many researchers have proposed a novel approach that integrates Markov models, association rules and clustering in web site access predictability. The low order Markov models provide higher coverage, but these are couched in ambiguous rules. In this paper, we introduce the use of default rule in resolving web access ambiguous predictions. This method could provide better prediction than using the individual traditional models. The results have shown that the default rule increases the accuracy and model-accuracy of web page access predictions. It also applies to association rules and the other combined models.**

Keywords-web mining, web usage mining; user navigation session; Markov model; association rules; Web page prediction; rule-selection methods.

I. INTRODUCTION

The rapid expansion of the World Wide Web has created an unprecedented opportunity to disseminate and gather information online. There is an increasing need to study web-user behavior to improve the service to web users and increase their value of enterprise. One important data source for this study is the web-server log data that traces the user's web browsing actions [1]. Predicting web-users' behavior and their next movement has been recognized and discussed by many researchers lately. The need to predict the next Web page to be accessed by the users is apparent in most web applications today whether or not they are utilized as search engines, e-commerce solutions or mere marketing sites. Web applications today are designed to provide a more personalized experience for their users [2]. The result of accurate predictions can be used for recommending products to customers, suggesting useful links, as well as pre-sending, pre-fetching and caching of web pages in reducing access latency [1]. There are various ways that can help us make such a prediction, but the most common approaches are the Markov models and association rules. Markov models are used in identifying the next page to be accessed by the Web site user based on the sequence of their previously accessed pages. Association rules can be used to decide the next likely web page requests based on significant statistical correlations.

Yang, Li and Wang [1] have studied different association-rule based methods for web request predictions. In this work, they have examined two important dimensions in building prediction models, namely, the type of antecedents of rules and the criteria for selecting prediction rules. In one dimension, they have a spectrum of rule representation methods which are: subset rules, subsequence rules, latest subsequence rules, substring rules and latest substring rules. In the second dimension, they have rule-selection methods namely: longest-match, most-confident and pessimistic selection. They have concluded that the latest substring representation, coupled with the pessimistic-selection method, gives the best prediction performance. The authors [2-5] have applied the latest substring representation using the most-confident selection method to building association-rule based prediction models from web-log data using association rules. In Markov models, the transition probability matrix is built making and predictions for web sessions are straight forward. In Markov models, the target is to build prediction models to predict web pages that may be next requested, The consequence to this is that, only the highest condition probability is considered. Hence, a prediction rule with the highest condition probability is chosen.

In the past, researchers have proposed different methods in building association-rule based prediction models using web logs, but none had yielded significant results. In this paper, we propose the default rule in resolving ambiguous predictions. This method could provide better prediction than using the traditional models individually.

The rest of this paper is organized as follows:
Sec.2 Related Works
Sec.3 Markov Models
Sec.4 Ambiguous rules
Sec.5 Experimental Setup and Result
Sec. 6 Conclusion

II. Related Work

There are wide application areas in analyzing the user web navigation behaviors in web usage mining [6]. The analysis of user web navigation behavior can help improve the organization of web sites and web performance by pre-fetching and caching the most probable next web page access. Web Personalization and Adaptive web sites are some of the applications often used in web usage mining. Web usage mining can provide guidelines in improving e-commerce to handle business specific issues like customer satisfaction, brand loyalty and marketing awareness.

The most widely used approach is web usage mining that includes many models like the Markov models, association rules and clustering [5]. 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. Many authors have proposed models for modeling the user web navigation sessions. Yang, Li and Wang [1] have studied five different representations of Association rules which are: subset rules, subsequent 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. Deshpande and Karypis [7] have developed techniques for intelligently combining different order Markov models resulting to lower state complexity, improved prediction accuracy, while retaining the coverage of the All-Kth-Order Markov model. Three approaches had been widely used namely frequency-pruning, error-pruning and support-pruning to reduce the state space complexity. Khalil, Li and Wang [2] have proposed a new framework in predicting the next web page access. They used lower all k-th Markov models to predict the next page to be accessed. If the Markov model is not able to predict the next page access, then the association rules are used to predict the next web page. They have also proposed, on the other hand, solutions for prediction ambiguities. Chimphlee [5] has proposed a hybrid prediction model (HyMFM) that integrates Markov model, Association rules and Fuzzy Adaptive Resonance Theory (Fuzzy ART) clustering all together. These three approaches are integrated to maximize their strengths. This model could provide better prediction than using an individual approach. In fact, Khalil, Li and Wang [4] had introduced the Integration Prediction Model (IPM) by combining the Markov model, association rules and clustering algorithm together. The prediction is performed on the cluster sets rather than the actual sessions. A. Anitha [8] has proposed to integrate a Markov model based sequential pattern mining with clustering. A variant of Markov model called the dynamic support, pruned all kth order in order to reduce state space complexity. The proposed model provides accurate recommendations with reduced state space complexity. Mayil [9] has proposed to model users' navigation records as inferred from log data, A Markov model and an algorithm scans the model first in order to find the higher probability trails which correspond to the users' preferred web navigation trails.

Predicting a user's next access on a website has attracted a lot of research work lately due to the positive impact of predictions on the different areas of web based applications [4]. First, many of the papers proposed using association rules or Markov models for next page predictions [1, 7, 9-11]. Second, many papers have addressed the uses of combining both methodologies [2-5]. Third, many of the papers have addressed the integration Markov model, Association rules and clustering method. This model could provide better predictions than using each approach individually [5]. It can be inferred that most researchers use Markov models for this type of prediction and is thus, mostly preferred.

In this paper, we study Markov models in predicting a user's next web request using default rule. The prediction models are based on web log data that correspond with users' behavior. They are used to make predictions for the general user more fit for a particular client. This prediction requires the discovery of a web users' sequential access patterns and using these patterns to make predictions of users' future access more accurate. We can then incorporate these predictions into web pre-fetching system in an attempt to enhance the performance.

III. Markov Models

Markov models are commonly used method for modeling scholastic sequences with underlying finite-state structures that are shown to be well-suited for modeling and predicting a user's browsing behavior on a web site [7]. The identification of the next web page to be accessed by a user is calculated based on the sequence of previously accessed pages.

In general, the input for this problem is the sequence of web pages that were accessed by a user. It is assumed that it discusses the Markov property. In such a process, the past is irrelevant in predicting the future given the knowledge of the present. Let $P = \{p_1, p_2, \ldots, p_m\}$ 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. Thus an equation may thus be deduced:

$$p_{l+1} = \text{argmax}_{p \in P}\{P(P_{l+1} = p/W)\}$$
$$= \text{argmax}_{p \in P}\{P(P_{l+1} = p/P_l, P_{l-1}, \ldots, P_1)\} \quad (1)$$

Essentially, this approach for each symbol P_i computes its probability of being accessed next which then selects the web page that has the highest probability of accessibility. The probability, $P(p_i/W)$, is estimated by using all W. Naturally, the longer l and the larger W are, the more accurate the results are $P(p_i/W)$. However, it is not feasible to accurately determine these conditional probabilities because the sequences may arbitrarily be longer (or longer l), and the size of the training set is often much smaller than that required to accurately estimate the various conditional probabilities for long sequences (or large W). For this reason, the various 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

$$p_{l+1} = \text{argmax}_{peP}\{P(P_{l+1} = {}_p/P_l,P_{l-1},\ldots,P_{l-(k-1)})\} \quad (2)$$

Where k denotes the number of the preceding pages as it identifies the order of Markov model. The resulting model of this equation is called the k^{th} –order Markov model. In order to use the k^{th} –order Markov model the learning of P_{l+1} is needed for each sequence of k web pages.

Let S_j^k 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, $S_j^k = <p_{l-(k-1)},P_{l-(k-2)}, \ldots,P_1>$, by estimating the various conditional probability $P(P_{l+1} = p/P_l,P_{l-1},\ldots,P_{l-(k-1)})$. Using the maximum likelihood principle [5], the conditional probability $P(p_i/S_j^k)$ is computed by counting the number of times a sequence S_j^k occurs in the training set, and the number of times pi occurs immediately after S_j^k. The conditional probability in the ratio of these two frequencies, therefore,

$$P(p_i/S_j^k) = \frac{\text{Frequency}(<S_j^k,p_i>)}{\text{Frequency}(S_j^k)} (3)$$

IV. AMBIGUOUSRULES

The main problem in association rules in the application of large data item sets is the discovery of a large number of rules and the difficulty in identifying the area that leads to the correct prediction. As regards the non-Markov models, they lack web page prediction accuracy because they do not use enough history in web access whereas Markov models have a high state space complexity [2-5].

There is an apparent direct relationship between Markov models and association rules techniques. According to the Markov model pruning methods presented by Deshpande and Karypis [7] and the association rules selection methods presented by Yang, Li and Wang [1].

The researchers herein have proposed different methods for building association-rule based prediction models in using web logs, but ambiguous rules still exist. In order to solve this problem, we propose the use of default rule to keep both the low state complexity and high accuracy results. We use the following examples to show the idea creating default rules. Consider the set of user session in table 1. Note that the numbers are assigned to web page names. Table 1 examines the following 6 user session:

TABLE I. USER SESSIONS

S1	900, 586, 594, 618
S2	900, 868, 594
S3	868, 586, 594, 618
S4	594, 619, 618
S5	868, 594, 900,618
S6	868, 586, 618, 594, 619

Table 2 shows an example of counting the support of extracted web access sequences and the useful sequences are highlighted. The row represents the previously visited page and the column represents the next visited page. Each field in the matrix is produced by looking at the number of times the web page on the horizontal line followed by the web page on the vertical. In this example, web user's session, web page 586 and web page 594 co-occurred in session S1 and S3, and therefore web page (A,B) has the support of 2.

TABLE II. AN EXAMPLE OF EXTRACTED WEB ACCESS SEQUENCES AND THEIR SUPPORT COUNT OF FIRST-ORDER MARKOV MODELS

Second item in sequence	Support count				
First item in sequence	586	594	618	619	868
586		2	1		
594			2	2	
618		1			
619			1		
868	2	2			
900	1		1		1

The next step is to generate rules from these remaining sequences. All the remaining sequences construct the prediction rules using the maximum likelihood principle. The condition probabilities of all sequences are calculated and are ranked.

For example, given the antecedent that the web page is 586, the condition probability is that 586 -> 594 is 85.7%. This is calculated by dividing the support value of the sequence (586,594) with the support value of the web page (586); (2/3 = 66.7%). From the training if antecedent web page is 586, then the single-consequence can be 594 and 618 with their confidence levels at 66.7%) and 33.3%. In this case, 586->594 have the highest probability value. Then a prediction rule with the highest condition probability is chosen.

Applying the First-order Markov models to the above training user sessions, we notice that the most frequent state is <594> and it appeared 6 times as follows:

$$P_{l+1} = \text{argmax}\{P(618|594)\} = 618 \text{ OR } 619$$

Using Markov models, we can determine that there is a 50/50 chance that the next page to be accessed by the user after accessing page 594 could be either 618 or 619.

Obviously, this information alone does not provide us with correct predictions on the next page to be accessed by the user as we may have obtained the highest condition probability for both pages, using the 618 and 619.

To break the tie and find out which page would lead to the most accurate prediction, we choose the rule whose right hand side (RHS) is the most popular page in the training web log (page 618 = 5, page 619 = 2). Thus, we choose rule 594->618. We refer to this rule as the default rule.

In this paper, we introduce the default rule in resolving ambiguous predictions. It can apply to all Markov models and Association rule. This method avoids the complexity of high order Markov model. This method also improves the efficiency of web access predictions. Algorithm is summarized as follows:

```
Training:
Generate rule using the First-order Markov model
Test:
FOR each session of Test set
  Latest substring from session (got LHS->RHS)
  Compare with appropriate rules
   IF the matching rule provides an non-
ambiguous prediction
     THEN
        The prediction is made by the state
ELSE  //The ambiguous occurs
         Select rule whose RHS is the most
frequency the training web log THEN make
prediction
  ENDIF
ENDFOR
```

In this work, we define an ambiguous prediction as one predictive page. This task could apply for an ambiguous prediction as two or more predictive pages that have the same conditional probabilities by a Markov model. The ambiguous prediction potentially has other definitions, for example, the certainty of a prediction is below a threshold. We, nonetheless, did not explore the other options in this paper [2].

V. Experimental Setup and Result

The experiment used on web data, as collected from the web server at the NASA Kennedy space Centre form 00:00:00 Aug 1, 1995 through 23:59:59 Aug 31, 1995, yielded a total of 31 days. During this period there are totally 1,569,898 requests recorded by the log file (see example in Fig.1).

Before doing the experiments, we removed all the dynamically generated content such as CGI scripts. We also filtered out requests with unsuccessful HTTP response codes [1, 12]. See Web log pre-processing in [12].

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400] "GET
/history/apollo/ HTTP/1.0" 200 6245
```

Figure 1. A fragment from the server logs file

The next step that we had used was to identify the user sessions. The session identification splits all the pages accessed by the IP address which is a unique identity and a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user has started a new session.

A 30 minute default timeout is considered. Figure 2 shows a fragment from this session identification result. There were a total of 71,242unique visiting IP addresses, 935 unique pages and 130,976 sessions. Also, the frequency of each page visited by the user was calculated. The page access frequency is shown in Figure 3 which reveals that page number 295 is the most frequent page accessed at and it was accessed 41109 times.

```
S1: 634, 391, 396, 408
S2: 393, 392, 400, 398, 396, 408, 37, 53
S3: 91, 124, 206, 101, 42, 287, 277
S4: 634, 391, 396, 631
S5: 124, 125, 127, 123, 130, 126, 131, 128, 129, 83
S6: 391, 634, 633, 295
S7: 295, 277, 91, 919
S8: 935, 391, 631, 634
S9: 755, 295, 810
S10: 637, 391, 918
```

Figure 2. A fragment from session identification result

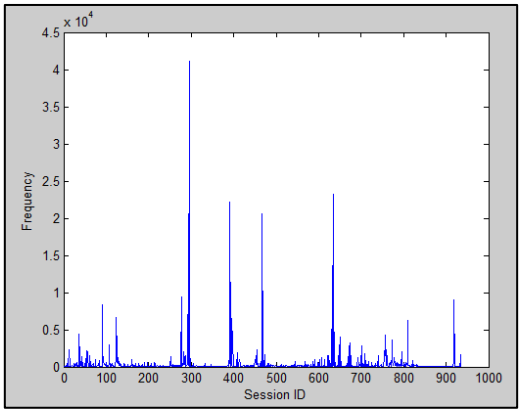

Figure 3. A frequency chart for the frequency visited sessions

All experiments were conducted on P i-7 2.20 GHz PC with 8 GB DDR RAM, running Windows 7 Home Premium. The algorithms were implemented using MATLAB.

This model will then be evaluated for accuracy on a previously unseen set of web sessions. These are called test set. When applying the trained model on the testing sequence, this is done by hiding the last symbol in each of the test sessions, and using the model to make a prediction of this trimmed session. During the testing step the model is given a trimmed sequence for prediction in which the last symbol of the sequence to compute the accuracy of the model is made. If the prediction model was unable to make a prediction for a particular web session, it was calculated as the wrong prediction. In the experiment, the proposed measure is compared with prediction model based evaluation that measures accuracy, coverage, and model accuracy [5]. Accuracy is defined as the ratio of the number of sequence for which the model is able to correctly predict the hidden symbol to the total number of sequence in the test set. The second is the coverage. It is defined as the ratio of the number of sequences whose required number for making a prediction was found in the model to the total number of sequences in the test set. The Third one is the model accuracy; it is calculated only on the web user sessions upon which the prediction was made. If the prediction model was unable to make a prediction for a particular web session, it is ignored in the model accuracy calculations.

The evaluation was calculated using a 10-fold cross validation. The data was spit into ten equal sets. The first nine sets are considered as training data for rule constructions. Then, the second last set is considered as testing data for evaluation. The test set is continued moving upward until the first set was used for testing and the rest for training. The reported accuracy is the average of ten sets. The experiment results are the average of ten tests. We experimentally evaluated the performance of the proposed approach first-order Markov model and construct the predictive model.

Figure 4 shows the number of ambiguous rules (%). It shows that as the minimum support threshold has sequences varied (2-8), the number of ambiguous rules occurs at any support threshold.

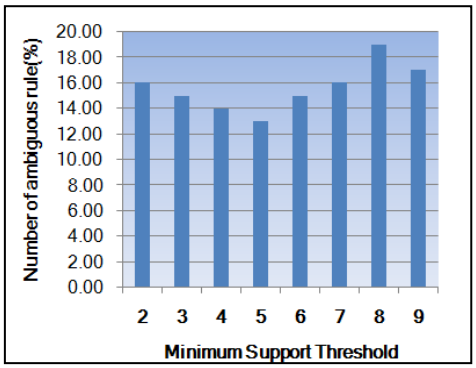

Figure 4. Number of ambiguous rules (%) occurred at difference minimum support thresholds.

The results are plotted in Figure 5-7. It shows that as the support threshold has increased, the coverage of the model decreased as accompanied by ad decrease in the accuracy. However, the model-accuracy of the model continues to increase.

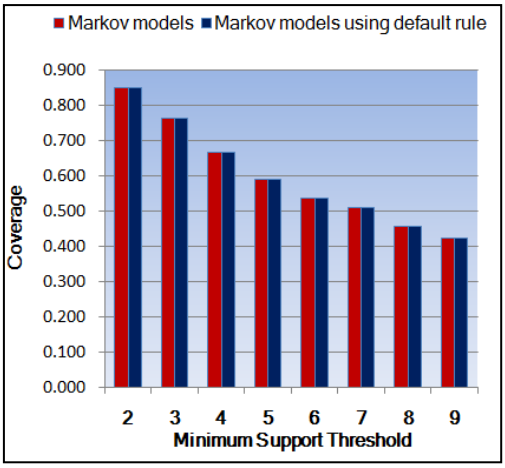

Figure 5. The comparison of coverage achieved by the Markov models and Markov models using default rules with difference minimum thresholds.

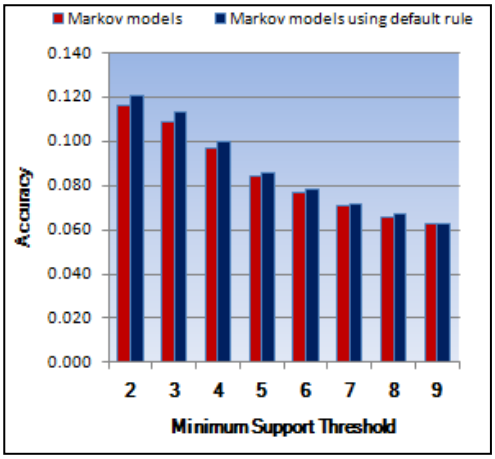

Figure 6. The comparison of accuracy achieved by the Markov models and Markov models using default rules with difference minimum support threshold.

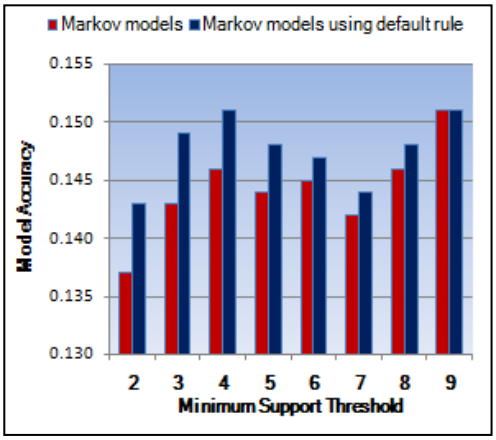

Figure 7. The comparison of Model-accuracy achieved by the Markov models and Markov models using default rule with difference minimum support threshold.

As it can be seen from Figures 5 – 7, the interesting observations can thus be summarized as the overall performances of Markov models using the default rule accuracy and model-accuracy while keeping their higher coverage abilities.

VI. CONCLUSION

A Markov models is a popular approach to predict what web page are likely to be accessed next. Many researchers have proposed to use the Markov models in predicting a web user's navigation session but there has been no proposal at present on how to solve the ambiguous problem rules. In this paper, we propose to use the default rules in resolving ambiguous prediction in the first order Markov models. This method could provide better web navigation prediction than merely using the individual traditional models individually. It can also apply to all Markov models, Association rule and the other combined Markov models.

ACKNOWLEDGMENT

The authors gratefully thank the anonymous referees and collaborators for their substantive suggestions. We also acknowledge research support from Rangsit University at Bangkok, Thailand.

REFERENCES

[1] Q. Yang, T. Li and K. Wang, "Building Association -Rules Based Sequential Classifiers for Web-Document Prediction, "Journal of Data Mining and Knowledge Discovery, Vol. 8, 2004.

[2] F. Khalil, J. Li and H. Wang, "A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses,"Conferences in Research and Practice in Information Tecnology (CRPIT), Vol. 61, 2006.

[3] S. Chimphlee, N. Salim, M.S.B Ngadiman, W. Chimphlee and S. Srinoy, "Predicting Next Page Access by Markov Models and Association Rules on Web Log Data,"The international Journal of Applied Management and Technology,Vol. 4, 2006.

[4] F. Khalil, J. Li and H. Wang, "Integrating recommendation models for improved web page prediction accuracy," Thirty-First Australasian Computer Science Conference (ACSC2008), 2008.

[5] S. Chimphlee, N. Salim, M.S.B. Ngadiman and W.Chimphlee, "Hybrid Web Page Prediction Model for Predicting a User's Next Access,"Information Technology Journal, Vol.9, No. 4, 2010.

[6] Bhawna Nigam, S.J.a.S.T., "Mining Association Rules from Web Logs by Incorporating Structural Knowledge of Website," International Journal of Computer Applications (0975-8887), 2012. Vol. 42(No. 11): p. pp. 17-23.

[7] M. Deshpande and G. Karypis,"Selective Markov Models for Predicting Web Page Accesses," In ACM Transactions on Internet Technology,Vol.4, No.2,2004.

[8] A. Anitha, "A New Web Usage Mining Approach for Next Page Access Prediction," International Journal of Computer Applications (0975-8887), Vol. 8, No. 11, 2010.

[9] V.V. Mayil, "Web Navigation Path Pattern Prediction using First Order Markov Model and Depth first Evaluation," International Journal of Computer Applications(0975-8887),Vol. 45, No. 16, 2012.

[10] S. Venkateswari and R.M. Suresh,"Association Rule Mining in E-commerce: A Survey,"International Journal of Engineering Science and Technology (IJEST), Vol. 3, No. 4, 2011.

[11] N.K. Tyagi and A.K. Solanki, "Prediction of Users Behavior through Correlation Rules,"International of Advanced Computer Science and Applications (IJACSA), Vol. 2, No. 9, 2011.

[12] T. Pamutha, S. Chimphlee, Ch. Kimpan and P. Sanguansat, "Data Preprocessing on Web Server Log Files for Mining Users Access Patterns,"International Journal of Research and Reviews in Wireless Communications (IJRRWC),Vol.2, No.2, 2012.

AUTHORS PROFILE

Thanakorn Pamutha received the M.Sc. degree in Computer Science from Prince of Songkla University, Songkla-Thailand in 2000. He has been an assistant professor in Yala Rajabhat University, Thailand since 2003. Now, he is a PhD student in Information Technology, Faculty of Information Technology, Rangsit University, Thailand. He is interested in database system, artificial intelligence, data mining, and web usage mining.

Siriporn Chimplee received the M.Sc. degree in Applied Statistics from National Institution Development of Administration (NIDA), Thailand, and Ph.D. degrees in Computer Science from University Technology Malaysia, Malaysia. She is a lecturer at the Computer Science Department, Faculty of Science and Technology, Suan Dusit Rajabhat University, Thailand. She is interested in data mining, web mining, web usage mining, statistical, and soft computing.

Chom Kimpan received his D.Eng in Electrical and Computer Engineering from King Mongkut's Institute of Technology Ladkrabang, M.Sc in Electrical Engineering from Nihon University, Japan, and bachelor's degree in Electrical Engineering from King Mongkut's Institute of Technology of Thailand. Now he is an Associate Profressor at the Department of Informatics, Faculty of Information Technology, Rangsit University, Thailand. He is interested in pattern recognition, image retrieval, speech recognition, and swarm intelligence.

Parinya Sanguansatreceived the B.Eng., M.Eng. and Ph.D. degrees in Electrical Engineering from the Chulalongkorn University, Thailand. He is an assistant professor in the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand in 2001, 2004 and 2007 respectively. His research areas are digital signal processing in pattern recognition including on-line handwritten recognition, face and automatic target recognition.