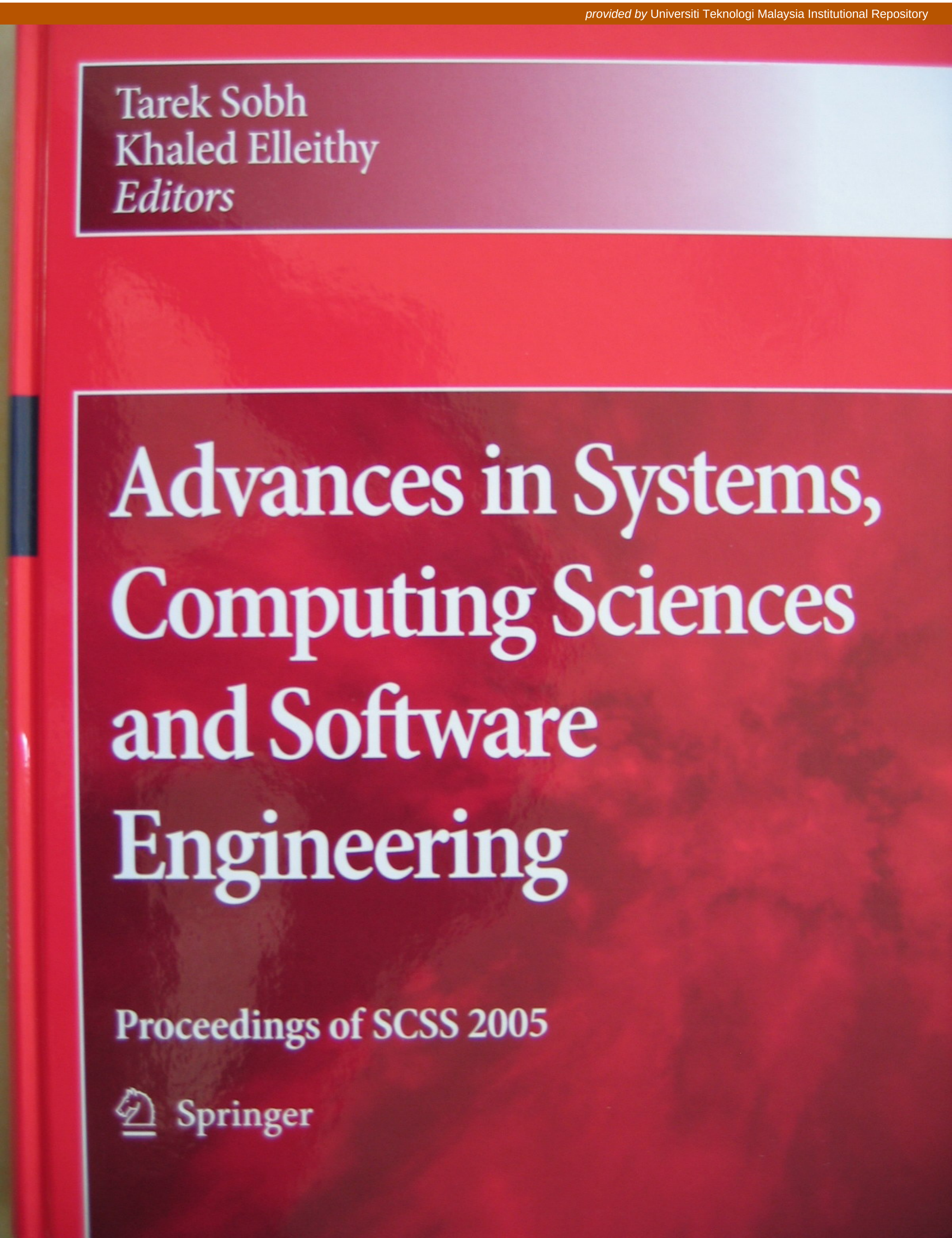

Tarek Sobh
Khaled Elleithy
Editors

Advances in Systems, Computing Sciences and Software Engineering

Proceedings of SCSS 2005

Springer

Using Markov Model and Association Rules for Web Access Prediction

Siriporn Chimphlee [1] **Naomie Salim** [2] **Mohd Salihin Bin Ngadiman** [3] **Witcha Chimphlee** [4]

[1,4]*Faculty of Science and Technology*
Suan Dusit Rajabhat University, 295 Rajasrima Rd, Dusit, Bangkok, Thailand
Tel: (+66)-2445675, Fax: (+66) 2445675, Email: [1]siriporn_chi@dusit.ac.th, [4]witcha_chi@dusit.ac.th

[2,3]*Faculty of Computer Science and Information Systems,*
University Technology of Malaysia, 81310 Skudai, Johor, Malaysia
Tel: (607) - 5532070, Fax: (607) 5565044, Email: [2]naomie@fsksm.utm.my , [3]msn@fsksm.utm.my,

Abstract

Mining user patterns of log file can provide significant and useful informative knowledge. A large amount of research has been done on trying to predict correctly the pages a user will request. This task requires the development of models that can predicts a user's next request to a web server. In this paper, we propose a method for constructing first-order and second-order Markov models of Web site access prediction based on past visitor behavior and compare it association rules technique. This algorithm has been used to cluster similar transition behaviors for efficient used to further improve the efficiency of prediction. From this comparison we propose a best overall method and empirically test the proposed model on real web logs.

Keywords: Association rules, Marov Model, prediction

1. 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 better serve the web users and increase the value of enterprises. One important data source for this study is the web-server log data that trace the user's web browsing actions. The web log data consist of sequences of URLs requested by different clients bearing different IP Addresses. Association rules can be used to decide the next likely web page requests based on significant statistical correlations. The result of accurate prediction can be used for recommending products to the customers, suggesting useful links, as well as pre-sending, pre-fetching and caching of web pages for reducing access latency [1]. The work by Liu et al. [2] and Wang et al. [3] considered using association rules for prediction by selecting rules based on confidence measures, but they did not consider the sequential classifiers [1]. It has been observed that user tend to repeat the trails they have followed once [4]. So, better prediction of a user's next request could be made on the data pertaining to that particular user, not all the users. However, this would require reliable user identification and tracking users among sessions. This is usually achieved by sending cookies to a client browser, or by registering users. Both require user cooperation and might discourage some of potential site visitors. As a result, many web sites choose not to use these means of user tracking. Also, building prediction models on individual data would require that users have accessed enough pages to make a prediction, which is not usually the case for a university website that has many casual users [5].

In the network system area, Markov chain models have been proposed for capturing browsing paths that occur frequently [4, 6]. However, researchers in this area did not study the prediction models in the context of association rules, and they did not perform any comparison with other potential prediction models in a systematic way. As a result, it remains an open question how to construct the best association rule based prediction models for web log data [1].

This paper is organized as follows. In section 2, we discuss the background and review the past works in related research. In section 3, we present the experimental design. In section 4, we discuss the experimental result. We conclude our work in section 5.

2. Background of study

2.1 Web Mining

Web mining is the term of applying data mining techniques to discover automatically and extract useful information from the World Wide Web documents and services [7]. The web mining system uses information determined from the history of the investigated web system. When valuable hidden knowledge about the system of interest has been discovered, this information can be incorporated into a decision support system to improve the performance of the system. These rules, patterns typically reflecting real world phenomena, are mined from the web server logs, proxy server logs, user' profiles, registration data etc. Three major web mining methods are web content mining, web structure mining and web usage mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining aims to generate structural summaries about web sites and web pages [8]. Web usage mining is to discover usage patterns from web data, in order to understand and better serve the needs of web-based application. It is an essential step to understand the users' navigation preferences in the study of the quality of an electronic commerce site. In fact, understanding the most likely access patterns of users allows the service provider to personalize and adapt the site's interface for the individual user, and to improve the site's structure.

2.2 Association rules

Agrawal and Srikant [9] were proposed to capture the co-occurrence of buying different items in a supermarket shopping. Given a set of transactions, where each transaction is a set of items, an association rule is an expression *X=>Y*, where *X* and *Y* are sets of items. The intuitive meaning of such a rule is that transactions in the databases which contain the items in *X* tend to contain also the items in *Y* [10]. For instance, 98% of customers who purchase tires and auto accessories also buy some automotive services; 98% is called the confidence of the rule. The support of the rule *X=>Y* is the percentages of transactions that contain both *X* and *Y*. Association rule generation can be used to relate pages that are most often referenced together in a single server session [11]. In the context of Web usage mining, association rules refer to set of pages that are accessed together with a support value exceeding some specified threshold. The association rules may also serves as a heuristic for prefetching documents in order to reduce user-perceived latency when loading a page from a remote site [11]. These rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. A transaction is a projection of a portion of the access log. In their work Liu et al. [2] and Wang et al. [3] considered using association rules for prediction without considering sequential classifiers. In contrast, Lan et al. [12] developed an association rules mining technique for the pre-fetching of web document from the server's disk into the server's cache. They constructed rules of the form $D_i \to D_j$, where D_i and D_j are documents (URLs). The intuitive interpretation of such rules are that document C is likely to be requested by the same user sometimes after document D_i has been requested and there is no other request between the requests for D_i and D_j , since it is usually the case according to the log. The counting of support is done differently than in Agrawal and Srikant [9] since the ordering of documents is considered [13]. They defined confidence is the ratio support $\left(D_i D_j\right)$/support$\left(D_i\right)$, support $\left(D_i\right)$ is the total number of occurrences of document D_i in the transactions over the total number of the transactions and support $\left(D_i D_j\right)$ is the total number of occurrences of a sequence $D_i D_j$ in the transactions over total number of transactions. Only consecutive subsequences inside a user transaction are supported. For instance, the user transaction ABCD supports the subsequences: AB, BC, and CD. Yang et al. [1] studied different association-rule based methods for web request prediction. Their analysis is based on a two dimensional structure and real web logs are used as training and testing data. First dimension is named antecedent of rules. They consist of five rules; called subset rule, subsequence rule, latest-subsequence rule, substring rule and latest-substring rules. These representations build the left-hand-side of association rules using non-empty subsets of URLs, non-empty sequences of URLs, non-empty sequences of URLs which end in the current time, non-empty adjacent sequences of URLs, and non-empty adjacent sequences of URLs which end in the current time. Second dimension is named criterion for selecting prediction rules. It represents as three prediction methods called longest match selection, most confident selection and pessimistic selection. The longest-match method selects the longest left-hand-side of the rule and matches an observed sequence from all rules with the minimum support rules. Most confident selection always chooses a rule with highest confidence and minimum support rules. Pessimistic selection combines the confidence and support for a rule to form an unified selection measure. The result showed that the latest substring rule coupled with the pessimistic-selection method gives the most precision prediction performance. In this work, we also used the latest-substring to represent the prediction rules.

2.3 Markov model

Markov models [14] have been used for studying and understanding stochastic processes, and were shown to be well-suited for modeling and predicting a user's browsing behavior on a web-site. Markov models have been widely used to model user navigation on the web and predicting the action a user will take next given the sequence of actions he or she has already performed. For this type of problems, Markov models are represented by three parameters < *A*, *S*, *T* >, where *A* is the set of all possible *actions* that can be performed by the user; S is the set of all possible states for which the Markov model is built; and T is a |S| × |A| Transition Probability Matrix (TPM), where each entry t_{ij} corresponds to the probability of performing the action *j* when the process is in state *i* [15]. In general, the input for these problems are the sequence of web-pages that were accessed by a user and the goal are to build Markov models that can be used to model and predict the web-page that the user will most likely access next. Padbanabham and Mogul [16] use N-hop Markov models predicted the next web page users will most likely access by matching the user's current access sequence with the user's historical web access sequences for improving pre-fetching strategies for web caches. Pirolli and Pitkow [4] predict the next web page by discovering the longest repeating subsequences in the web sessions, and then using a weighted scheme to match it against the test web sessions. Sarukkai [17] used first-order Markov models to model the sequence of pages requested by a user for predicting the next page accessed. Cadez et al. [18] clustered user behaviors by learning a mixture of first-order Markov models using the Expectation-Maximization algorithm. They then display the behavior of a random sample of users in each cluster along with the size of each cluster. They also applied to the visualization of web traffic on the msnbc.com site.

3. Experimental design

In this paper, we study association rules and Markov models for predicting a user's next web requests. The prediction models that we build are based on web log data that correspond with users' behavior. They are used to make prediction for the general user and are not based on the data 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. We will then incorporate these predictions into the web prefetching system in an attempt to enhance the performance.

The experiment on used web data, collected from www.dusit.ac.th web server (see example in Figure 1) during 1st December 2004 – 31st December 2004. The total number of web pages with unique URLs is equal to 314 URLs, and there are 13062. These records are used to construct the user access sequences (Figure 2). The user sessions are split into training dataset and testing dataset. The training dataset is mined in order to extract rules, while the testing dataset is considered to evaluate the predictions made based on these rules. We experimentally evaluated the performance of the proposed approach: first-order markov model, second-order markov model, and association rule mining and construct the predictive model.

```
1102801060.863  1897600  172.16.1.98  TCP_IMS_HIT/304 203
GET http://asclub.net/images/main_r4_c11.jpg - NONE/- image/jpeg
1102801060.863  1933449  172.16.1.183  TCP_MISS/404 526
 GET  http://apl1.sci.kmitl.ac.th/robots.txt DIRECT/161.246.13.86
text/html
1102801060.863  1933449   172.16.1.183  TCP_REFRESH_HIT/200 3565
GET http://apl1.sci.kmitl.ac.th/wichitweb/spibigled/spibigled.html -
 DIRECT/161.246.13.86 text/html
```

Figure 1. Web log data

3.1 Web log preprocessing

Web log files contain a large amount of erroneous, misleading, and incomplete information. This step is to filter out irrelevant data and noisy log entries. Elimination of the items deemed irrelevant by checking the suffix of the URL name such as gif, jpeg, GIF, JPEG, jpg, JPG. Since every time a Web browser downloads a HTML document on the Internet, several log entries such as graphics and script are downloaded too. In general, a user does not explicitly request all the graphics that are in the web page, they are automatically down-loaded due to the HTML tags. Since web usage mining is interested in studying the user's behavior, it does not make sense to include file requests that a user does not explicitly request. The HTTP status code returned in unsuccessful requests because there may be bad links, missing or temporality inaccessible pages, or unauthorized request etc: 3xx, 4xx, and 5xx. Executions of CGI script, Applet, and other script codes are also eliminated. This is due to the fact that there is not enough Meta data to map these requests into semantically meaningful actions, as these records are often too dynamic and contain insufficient information that makes sense to decision makers.

3.2 Session identification

After the preprocessing, the log data are partitioned into user sessions based on IP and duration. Most users visit the web site more than once. The goal of session identification is to divide the page accesses of each user into individual sessions. The individual pages are grouped into semantically similar groups. A user session is defined as a relatively independent sequence of web requests accessed by the same user [19]. Fu et al. [20] identify a session by using a threshold idles time. If a user stays inactive for a period longer than the *max_idle_time*, subsequent page requests are considered to be in another episode, thus identificated as another session. Most researchers use heuristic methods to identify the Web access sessions [21] based on IP address and time-out does not exceeding 30 minutes for the same IP Address. A new session is created when a new IP address is encountered after a timeout. Catledge and Pitkow [22] established a timeout of 25.5 minutes based on empirical data. In this research, we use IP address, time-out of 30 minutes, to generate a new user (Figure 2).

Session 1 : 900, 586, 594, 618
Session 2 : 900, 868, 586
Session 3 : 868, 586, 594, 618
Session 4 : 594, 618, 619
Session 5 : 868, 586, 618, 900

Figure 2. User session from data set

We assume the access pattern of a certain type of user can be characterized by a certain a minimum length of a user's transaction, and that the corresponding future access path is not only related to the last accessed URL. Therefore, users with relatively short transactions (e.g. 2-3 accesses per transaction) should be handled in a different way from users with long transactions (e.g. 10-15 accesses per transaction) [23]. In this study, we proposed a case definition design based on the transaction length. User transactions with lengths of less than three are removed because it is too short to provide sufficient information for access path prediction [23].

3.3 Prediction using Association rules

To capture the sequential and time-limited nature of prediction, we define two windows. The first one is called *antecedent Window (W1),* which holds all visited pages within a given amount of user requests and up to a current instant in time. A second window, called the *consequent* *window (W2),* holds all future visited pages within amount of user requests from the current time instant.

The web log data are a sequence of entries recording which documents was requested by a user. We extracted a prediction model based on the occurrence frequency and find the last-substring [1] of the W1. The last-substrings are in fact the suffix of string in W1 window. We will refer Left-Hand-Side as *LHS* and Right-Hand-Side as *RHS*.These rules not only take into account the order and adjacency information, but also the newness information about the LHS string. We used only the substring ending in the current time (which corresponds to the end of window W1) qualifies to be the LHS of a rule [1]. We give a simple example to illustrate the prediction scheme of association rules clearly. The input data for training the model consists of web sessions, where each session consists of the sequence of the pages accessed by a user during his/her visit to the site. The training data of our example, shown in Figure 2, is from five user sessions.

Table 1-The latest-substring rules

W1	W2	The latest-substring rule
900, 586, 594	618	{594} -> 618
868, 586, 594	618	

From these rules, we extract sequential association rules of the form LHS->RHS from the session [1]. The support and confidence are defined as follows:

$$supp = \frac{count\,(LHS\text{-} > RHS)}{number\ of\ sessions}$$

$$conf = \frac{count\,(LHS\text{-} > RHS)}{count\,(LHS)}$$

Our goal is to output the best prediction on a class based on a given training set. Therefore, we need a way to select among all rules that apply. In a certain way, the rule-selection method compresses the rule set. If a rule is never applied, then it is removed from the rule set. In association rule mining, a major method to construct a classifier from a collection of association rules is the most-confident selection method [2]. The most confident selection method always chooses a rule with the highest confidence among all the applicable association rules, where support values are above the minimum support threshold. For example, suppose a testing set has a previous sequence of (A, B, C). Using the most-confident rule selection method, we can find three rules which can be applied to this example,

Rule 1: (A, B, C) →D with confidence 35%
Rule 2: (B, C) →E with confidence 60%
Rule 3: (C) →F with confidence 50%

In this case, the confidence values of *rule 1*, *rule 2* and *rule 3* are 35%, 60% and 50%, respectively. Since *Rule 2* has the highest confidence, the most-confident selection method will choose *Rule 2,* and predict E as the next page to be accessed.

3.4 Markov prediction model

The Markov model has achieved considerable success in the web prefetching field [4, 24, 25]. However the limit of this approach in web prefetching is that only requested pages are considered. The state-space of the Markov model depends on the number of previous actions used in predicting the next action. The simplest Markov model predicts the next action by only looking at the last action performed by the user. In this model, also known as the first-order Markov model, each action that can be performed by a user corresponds to a state in the model (Table 2). A somewhat more complicated model computes the predictions by looking at the last two actions performed by the user. It is called the second-order Markov model, and its states correspond to all possible pairs of actions that can be performed in sequence (table 3). This approach is generalized to the K^{th}-order Markov model, which computes the predictions by looking at the last *K* actions performed by the user, leading to a state-space that contains all possible sequences of *K* actions [15]. We also used training user sessions, shown in figure 2. In this example, we find first-order and second-order Markov Model and set the support threshold as 2. Based on this training set, the supports of different order sequences are counted. Prediction rules and their predictions confidence are shown in Table 4.

Table 2-Sample First-order Markov

1^{st} order	Support count					
Second item in sequence First item in sequence	586	594	618	619	868	900
586	0	2	1	0	0	0
594	0	0	3	0	0	0
618	0	0	0	0	1	1
619	0	0	0	0	0	0
868	3	0	0	0	0	0
900	1	0	0	0	0	1

Table 3- Sample Second-order Markov

2^{st} order	Support count					
Second item in sequence First item in sequence	586	594	618	619	868	900
586->594	0	0	2	0	0	1
594->619	0	0	0	1	0	0
868->586	0	1	1	0	0	0
900->868	1	1	0	0	0	0

4. Experimental results & discussions

The most commonly used evaluation metrics are accuracy, precision, recall and F-Score. Deshpande and Karypis [24]

Table 4- Prediction rule and confidence

Rule-selected	Prediction	confidence
586	594	2/3 = 67%
594	618	3/3 = 100%
868	586	3/3 = 100%
586->594	618	2/3 = 67%

used several measures to compare different Markov model-based techniques for solving the next-symbol prediction problem: accuracy, number of states, coverage and model-accuracy. Haruechaiyasak [26] and Zhu et al. [27] used precision and recall to evaluate the performance of method. The precision measure the accuracy of the predictive rule set when applied to the testing data set. The recall measures the coverage or the number of rules from the predictive rule set that matches the incoming request [26]. To evaluate classifiers used in this work, we apply precision and recall, which are calculated to understand the performance of the classification algorithms. Based on the confusion matrix computed from the test results, several common performance metrics can be as Table 5, where TN is the number of true negative samples; FP is false positive samples; FN is false negative samples; FP is true positive samples. Precision and recall can be as Table 5.

Table 5- Confusion Matrix

Actual	Predicted	
	positive	negative
Positive	TP	FN
Negative	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

4.1 Results

The results plotted in Figure 3 and show comparison of the different algorithms. We divided the web log as training data and testing data. As can be seen from the Figure 3, the first-order Markov model consistently gives the best prediction performance. The second-order worst when the recall less than 50% but the association rule is worst after the precision less than 50%.

4.2 Discussions

Web usage mining is the application of data mining techniques to usages logs of large Web data repositories to produce results that can be used in the design tasks. In this experiment, the three algorithms are not successful in correctly predicting the next request to be generated. The first-order Markov Model is best than other because it can extracted the sequence rules and chose the best rule for prediction and at the same time second-order decrease the coverage too. This is due to the fact that these models do not look far into the past to discriminate correctly the difference modes of the generative process.

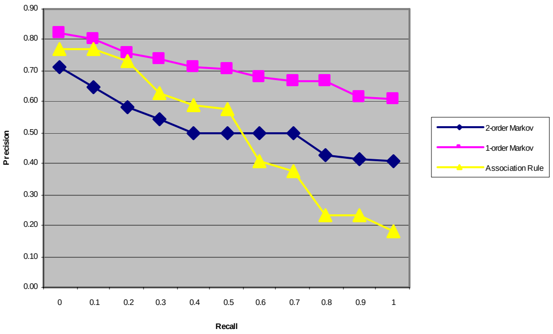

Figure 3. Result compare among three techniques

5. Conclusions and future work

Web servers keep track of web users' browsing behavior in web logs. From log file, one can builds statistical models that predict the users' next requests based on their current behavior. In this paper we studied different algorithm for web request prediction. Our analysis based on three algorithm and using real web logs as training and testing data and show that the first-order Markov model is the best prediction after compared to use. In the future, we plan to use rough sets for prefetching to extract sequence rules.

6. Reference

[1] Yang, Q., Li, T., Wang, K., "Building Association-Rules Based Sequential Classifiers for Web-Document Prediction", *Journal of Data Mining and Knowledge Discovery*, Netherland: Kluwer Academic Publisher, vol. 8, 2004, 253-273.

[2] Liu, B., Hsu, W., and Ma, Y., Integrating Classification and Association Mining, *Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining* (KDD-98), 1998.

[3] Wang, K., Zhou, S. Q. and He, Y., Growing Decision Trees on Association Rules, *Proc. of the International Conference of Knowledge Discovery in Databases*, 2000.

[4] Pitkow, J. and Pirolli, P., Mining Longest Repeating Subsequences to Predict World Wide Web Surfing, *Proc. USENIX Symp. On Internet Technologies and Systems*, 1999.

[5] Chakoula, O. Predicting Users' Next Access to the Web Server. Master diss., Dept. of Computer Science, University of British Columbia, 2000.

[6] Su, Z., Yang, Q., Lu, Y., and Zhang, H., What next: A Prediction System for Web Requests using N-gram Sequence Models, *Proc. of the First International Conference on Web Information Systems and Engineering Conference*, Hong Kong, 2000.

[7] Etzioni, O., "The World Wide Web: Quagmire or Gold Mine" *Communications of the ACM*, vol.39 (11), 1996, 65-68.

[8] Madria, S.K., Bhowmick, S.S., Ng, W.K.and Lim E., Research Issues in Web Data Mining, *Proc. First*

International Conference on Data Warehousing and Knowledge Discovery, Italy, Florence, 1999, 303-312.

[9] Agrawal, R. and Srikant, R., Fast algorithms for mining association rules", *Proc. of the 20th VLDB Conference ages*, Santiago, Chile, 1994.

[10] Srikant, R., Agrawal, R. "Mining Generalized Association Rules", *Proc. of the 21st Int'l Conference on Very Large Databases*, Zurich, Switzerland, Sep. 1995. Expanded version available as IBM Research Report RJ 9963, June 1995.

[11] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations*. vol.1 (2), 2000, 12-23.

[12] Lan, B., Bressan, S., Ooi, B.C. and Y. Tay, Making Web Servers Pushier, *Proc. Workshop Web Usage Analysis and User Profiling*, 1999.

[13] Nanolopoulos, A., Katsaros, D. and Manolopoulos, Y. "A Data Mining Algorithm for Generalized Web Prefetching", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15 (5), 2003, 1155-1169.

[14] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, NY: McGraw Hill, 1991.

[15] Deshpande, M. Prediction/Classification Technique for Sequence and Graphs, Ph.D. dissertion, University of Minnesota, January 2004.

[16] Padmanabhan, N. and Mogul, J.C. "Using predictive prefetching to improve World Wide Web latency", *ACM SIGCOMM Computer Communication Review*. vol. 26 (3), 1996, 22-36.

[17] Sarukkai, R.R., Link prediction and path analysis using Markov Chain, *Proc. of the 9th international World Wide Web conference on Computer networks: The international journal of computer and telecommunications networking*, Amsterdam, Netherlands, 2000.

[18] Cadez, I., Heckerman, D., Meek, C., Smyth, P.and White, S., "Visualization of Navigation Patterns on a Web Site Using Model Based Clustering", *Technical Report MSR-TR-00-18, Microsoft Research*, 2000.

[19] Cooley, R., Tan, P-N, Srivastava, J., "Discovery of Interesting Usage Patterns from Web Data", *Springer-Verlag LNCS/LNAI series*, 2000.

[20] Fu, Y., Sandhu, K. and Shih, M.Y., Clustering of Web Users Based on Access Patterns, *Proc. of the 5th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, San Diego: Springer, 1999.

[21] Pallis, G., Angelis, L. and Vakali, A., "Model-based cluster analysis for web users sessions", *Springer-Verlag Berlin Heideberg*, 2005, 219-227.

[22] Catledge, L. and Pitkow, J.E., "Characterizing Browsing Behaviors on The World Wide Web", *Computer networks and ISDN Systems*, vol.27 (6), 1995.

[23] Wong, C., Shiu, S. and Pal, S., Mining Fuzzy Association Rules for Web Access Case Adaptation", *Proc. of the workshop Programme at the fourth International Conference on Case-Based Reasoning*, Harbor Center in Vancouver, British Columbia, Canada, 2001.

[24] Deshpande, M. and Karypis, G., Selective Markov Models for Predicting Web Page Accesses, *In Workshop on Web Mining at the First SIAM International Conference on Data Mining*, 2001.

[25] Mobasher, B., Dai, H., Luo, T., Nakagawa, M., Using Sequential and Non-Sequential Patterns in Predictive Web Usage Mining Tasks, *(ICDM 2002)*, 2002.

[26] Haruechaiyasak, C. A Data Mining and semantic Web Framework for Building a Web-based Recommender System. PhD. diss., University of Miami, 2003.

[27] Zhu, J.,Hong, J. and Hughes, J.G., Using Markov Chains for Link Prediction in Adaptive Web Site, *Proc. of Soft-Ware 2002: First International Conference on Computing in an Imperfect World*, Lecture Notes in Computer Science, Springer, Belfast, 2002.