IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 8, Issue 5 (Jan. - Feb. 2013), PP 70-74
www.iosrjournals.org

Performance Analysis Of Web Page Prediction With Markov Model, Association Rule Mining(Arm) And Association Rule Mining With Statistical Features(Arm-Sf)

[1]Sampath P., and [2]Ramya D.
[1(]*Research Scholar,Department of Computer Science and Engineering,Bannari Amman Institute of Technology, Sathyamangalam,*
[2 (]*PG Scholar,Department of Computer Science and Engineering,Bannari Amman Institute of Technology, Sathyamangalam,*

Abstract: *Web prediction is a classification problem in which we have to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. Predicting user's behavior can be applied effectively in various critical applications in the internet environment.Such application has traditional tradeoffs between modeling complexity and prediction accuracy. The web usage mining techniques are used to analyze the web usage patterns for a web site. The user access log is used to fetch the user access patterns. The access patterns are used in the prediction process. Markov model and all-Kth Markov model are used in Web prediction. A Markov model is proposed to alleviate the issue of scalability in the number of paths. The framework can improve the prediction time without compromising prediction accuracy.*

The proposed system is to compare the prediction accuracy with the markov model, ARM, ARM-SF and Boosting and Bagging model. The system improves the accuracy with scalability considerations. Finally the result will shows which would have better prediction accuracy.

Keywords: *Association rule mining (ARM),Association rule mining with statistical features(ARM-SF),Markov model.*

I. Introduction

The interests and tastes of browsers are captured according to the previously visited categorized Web pages. So the web prediction problem (WPP) can be generalized and applied in many essential industrial applications such as search engines, caching systems, recommendation systems, and wireless applications[5].Thus while a prediction model for a certain Web site is available, the search engine can utilize it to cache the next set of pages that the users might visit. Such caching mitigates the latency problem of viewing Web documents particularly during Internet traffic congestion periods. In Web prediction, we face challenges in both preprocessing and prediction. Preprocessing challenges include handling large amount of data that cannot fit in the computer memory, choosing optimum sliding window size, identifying sessions, and seeking/ extracting domain knowledge. Prediction challenges include long training/ prediction time, low prediction accuracy, and memory limitation.

Additionally, some models, such as association rule mining (ARM) and SVM, do not scale well with large data sets. Furthermore, some models, such as SVM and ANN, do not handle the multiclass problem efficiently because of the large number of labels/classes involved in the WPP. In this paper, we are analyzing the prediction accuracy range comparable to markov model and modified markov model with the association rule mining.

Figure 1.Web page prediction from different links.

Our contributions in this paper can be summarized as follows.

1)We present an analysis study for Markov model and all-Kth model in the WPP utilizing different N-grams. Specifically, we show how accuracy is affected when using different N-grams.

2)We propose a new modified Markov model that handles the excess memory requirements in case of large data sets by reducing the number of paths during the training and testing phases.

3)We conduct extensive experiments on three benchmark data sets to study different aspects of the WPP using Markov model, ARM,ARM-SF and all-Kth Markov model. Our analysis and results show that higher order Markov model produces better prediction accuracy.

Moreover, the results demonstrate the positive effect of our proposed modified Markov model in reducing the size of the prediction models without compromising the prediction accuracy.

II. Related Works

Prediction should be classified into two types as path-based prediction and point-based prediction. Path-based prediction is based on user's previous and historic path data, while point-based prediction is based on currently observed actions. Accuracy of point-based models is low due to the relatively small amount of information that could be extracted from each session to build the prediction model.

Researchers have used various prediction models including k-nearest neighbor (kNN), ANNs [7],fuzzy inference [6] SVMs, Bayesian model, Markov model and others.

Mobasher et al. use the ARM technique in WPP and propose the frequent item set graph to match an active user session with frequent item sets and predict the next page that user is likely to visit. However, ARM suffers from well-known limitations including scalability and efficiency.

In the context of adaptive learning, Anderson et al. use dynamic links that provide shorter path to reach the final destination.Perkowitz and Etzioni utilize adaptive Web sites based on the browsing activities. Su et al. have proposed the N-gram prediction model and applied the all-N-gram prediction model in which several N-grams are built and used in prediction.

Hassan et al. use Bayesian model to focus on certain patterns such as short and long sessions, page categories, range of page views, and rank of page categories.

III. Markov Model, Arm And Arm-Sf

3.1 Markov Model

The concept of Markov model is to predict the next action depending on the result of previous actions. In Web prediction, the next action corresponds to predicting the next page to be visited. The previous actions correspond to the previous pages that have already been visited. In Web prediction, the K^{th}-order Markov model is the probability that a user will visit the k^{th} page provided that she has visited the ordered $k - 1$ pages. For example, in the second-order Markov model, prediction of the next Web page is computed based only on the two Web pages previously visited.

In this section, we propose to compare Markov model,ARM,ARM-SF for finding better prediction accuracy model for web page prediction. Recall that, in Markov model, we consider lists in building the model, for example, user sessions S1 = (P1, P2) and S2 = (P2, P1) are two different sessions; hence, each session can have different prediction probability. On the other hand, in ARM, S1 and S2 are the same item set.

The basic idea in the modified Markov model is to consider a set of pages in building the prediction model to reduce its size. For example, we consider all the sessions (P1, P2, P3), (P1, P3, P2), (P2, P1, P3), (P2, P3, P1), (P3, P1, P2), and (P3, P2, P1) as one set P1, P2, P3. Our motivation is that a task on the Web can be done using different paths regardless of the ordering that the users choose. In addition, we reduce the size of prediction model by discarding sessions that have repeated pages. These sessions might result when the user accidentally clicks on a link and hits the back button.

Tablei:Two-Tier Framework Training Process

Input : M is the set of prediction model of size N: T is a set of training examples
Output : A set of trained classifiers and an example Classifier EC.
1)For each classifier model m in M train m on T
2)For each training example e in T and a classifier model m in M Do if m predicts the target of e correctly then map e to m and record the confidence of m in prediction.
3)For each example e in T, if e is mapped to more than one model then filter the labels so that only one label is kept.
4)Train EC on the training set T', where T' is a training set that has all examples in t and each example is mapped to one model in M

The K^{th} order of modified Markov model computes the probability that a user will visit the k^{th} page given that she has visited the $k - 1$ pages in any order as a toy example. Note that the last page of the session is assumed to be the final destination and it is separated from the sessions.

3.2 ARM

ARM is a data mining technique that has been applied successfully to discover related transactions. Specifically, ARM focuses on associations among frequent itemsets. For example, in a supermarket store, ARM helps uncover items purchased together which can be utilized for shelving and ordering processes. In the following, we briefly present how we apply ARM in WPP.

In WPP, prediction is conducted according to the association rules that satisfy certain support and confidence as follows. For each rule, $R = X \rightarrow Y$,of the implication, X is the user session and Y denotes the target destination page. Prediction is resolved as follows:

Note that the cardinality of Y can be greater than one, i.e., prediction can resolve to more than one page. Moreover, setting the minimum support plays an important role in deciding a prediction. In order to mitigate the problem of no support for $X \cup Y$, we can compute prediction ($X' \rightarrow Y$), where X' is the item set of the original session after trimming the first page in the session. This process is very similar to the all-K^{th} Markov model. However, unlike in the all-K^{th} Markov model, in ARM, we do not generate several models for each separate N-gram. In the following sections, we will refer to this process as all-K^{th} ARM model.

3.3 ARM-SF

To finding out the better prediction of web page by including association rule mining with statistical features.

IV. Performance Analysis

Table I presents our proposed two-tier framework. In step 1, we train the set of classifiers M on the training set T. Note that a subset of T is sufficient for training when T is large. In step 2, the labeling process is applied in which each example e of the training set is labeled with one of the M classifiers that successfully predicts the outcome of e. Filtering the examples that have more than one mapped label is done in step 3. In step 4, the EC is trained on the filtered mapped training examples.

In this paper, we generate N orders of Markov models, namely, first, second, . . Nth-order Markov models, by applying sliding windows on the training set T. These prediction models represent a repository that can be used in prediction.

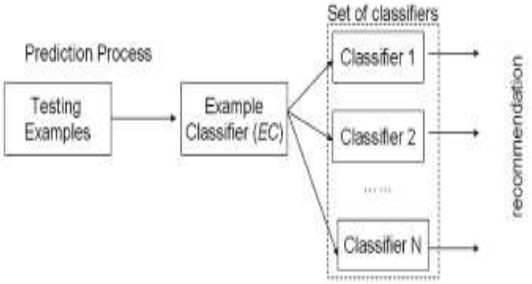

Figure.2 Prediction process in the two-tier model

Next, we map each training example in T to one or more orders of Markov models. For example, training example t3, (P1, P3, P5) is mapped to two classifier IDs, namely, C1 (first-order Markov model) and C2 (second-order Markov model). In prediction, a testing example x submits to two stages of prediction. First, x is fed to EC as input to predict the suitable classifier for x C_x. Next, x goes through the predicted classifier C_x to determine the final outcome.

This is a time-consuming process, particularly when prediction is required online. Next the same steps can be proceeded with the modified markov model. Finally we predict the set of pages and provide the better prediction accuracy model.

The user web browsing behavior identification is done with user access logs. The system is designed to perform pattern extraction and webpage prediction process. Log optimizer module is designed to perform access log preprocessing operations. Web user behaviors are identified using the Markov analysis module. The rule mining module is designed for pattern discovery and prediction process.

4.1 Log Optimizer

The web page requests are maintained under the access log files. Redundant page requests are removed from the log files. The page requests are grouped into sessions.

4.2 Markov Model

The Markov model is used to predict the next action depending on the result of previous actions.The K^{th} Markov model is prepared with request paths.

4.3 Rule Mining Process

The rule mining process is applied to extract the patterns. The apriori algorithm is used in the rule mining process. The patterns are identified from the item set collections. Support and confidence ratio are used in the prediction process.

4.4 Rule mining with statistical features

Statistical association rules are used to select the most relevant features to discriminate the web pages and eliminating noisy features that influence negatively the query results, making the whole process more efficient.

V. Result

We present an analysis study for Markov model and all-*K*th model in the WPP utilizing different *N*-grams. Specifically, we show how accuracy is affected when using different *N*-grams. We propose a new modified Markov model that handles the excess memory requirements in case of large data sets by reducing the number of paths during the training and testing phases. The system also reduces the prediction time particularly when there are many prediction models to consult. The system is tested with different set of log transactions. The analysis shows that the ARM-SF improves the prediction accuracy 25% more than others.

Transactions	Markov	ARM	ARM-SF
100	0.573	0.607	0.652
200	0.594	0.638	0.698
300	0.608	0.672	0.745
400	0.626	0.719	0.796
500	0.649	0.751	0.848

Table 1.Prediction Accuracy Analysis -Markov Model, ARM and ARM-SF

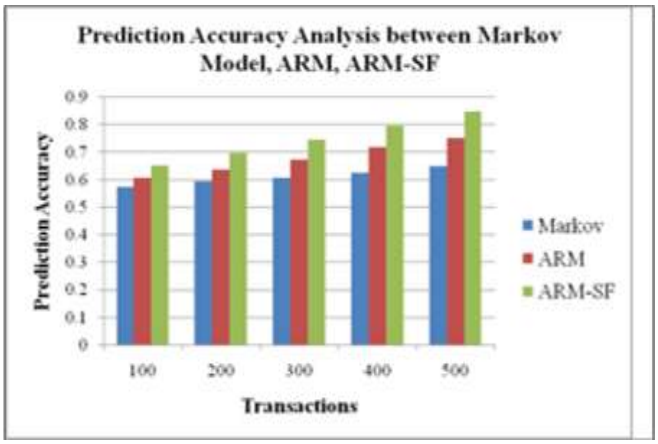

Figure 3. Prediction Accuracy Analysis -Markov Model, ARM and ARM-SF

VI. Conclusion And Future Enhancement

The web access pattern mining and prediction scheme is analyzed with different log files. The user access log files are collected from the web servers. The system is tested with markov model,ARM,ARM-SF. The prediction accuracy is used as the performance metric to evaluate the quality of the system. The system is designed to successfully improve prediction accuracy using simpler probabilistic models such as Markov model,ARM,ARM-SF. In future we extend our work with the comparision of markov model,ARM,ARM-SF and boosting and bagging model for finding better prediction accuracy.

References

[1] Awad.M,Khan.L and Thuraisingham.B, "Predicting WWW surfing using multiple evidence combination," VLDB J., vol.17, no.3, pp.401–417, May 2008.

[2] Awad.M and Khan.L,"Web navigation prediction using multiple evidence combination and domain knowledge," IEEE Trans. Syst.,Man,Cybern. A, Syst., Humans, vol. 37, no. 6, pp. 1054–1062, Nov. 2007.

[3] Fu..Y, Paul.H, and Shetty.N,"Improving mobile Web navigation using N-Gram prediction model," Int.J.Intell.Inf.Technol., vol.3, no.2, pp.51–64, 2007.

[4] Hassan.M.T, Junejo.K.N and Karim.A, "Learning and predicting key Web navigation patterns using Bayesian models," in Proc.Int.Conf.Comput.Sci.Appl.II, Seoul, Korea, 2009, pp. 877–887.

[5] Mamoun Awad.A and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model" IEEE Transactions on Systems, Man and Cybernetics-Part b:Cybernetics, vol.42, no.4, August 2012.

[6] Nasraoui.O and Petenes.C,"Combining Web usage mining and fuzzy inference for Website personalization," in Proc.WebKDD, 2003, pp.37–46.

[7] Nasraoui.O and Krishnapuram.R,"One step evolutionary mining of context sensitive associations and Web navigation patterns," in Proc.SIAM Int.Conf.Data Mining, Arlington, VA, Apr.2002, pp.531–547.

[8] Nasraoui.O and Krishnapuram.R, "An evolutionary approach to mining robust multi-resolution web profiles and context sensitive URL associations," Int.J.Comput.Intell.Appl.,vol.2, no.3, pp. 339–348, 2002.

[9] Levene.M and Loizou.G,"Computing the entropy of user navigation in the Web," Int. J.Inf.Technol.Decision Making, vol.2, no.3, pp.459–476, 2003.

Prof.P.Sampath received his M.E in Computer Science & Engineering and is now pursuing his Ph.D. in data mining at Anna University, Chennai. Currently, he is working as Associate Professor in the department Computer Science and Engineering,Bannari Amman Institute of Tech, Sathyamangalam, Erode, Tamil Nadu, India. He has 18 years of experience in teaching field. He has so far published 15 papers in National Conferences and he has presented 2 papers in International Conferences held at various reputed engineering colleges. He has also published two papers in international journals.

Ms.D.Ramya, received her B.E in Avinashilingam University for Women, Coimbatore and she is now pursuing her ME in Computer Science & Engineering at Bannari Amman Institute of Tech,Sathyamangalam,Erode,TamilNadu, India. She had published 1 Papers in International Conferences and 2 Papers in National Conference.