Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.209 – 212

REVIEW ARTICLE

Prediction of Web Users Browsing Behavior: A Review

Neha V. Patil[1], Dr. Hitendra D. Patil[2]

[1]Master Student, Computer Engineering, SSVPS'S B.S.Deore College of Engineering, India

[2]Professor and Head, Computer Engineering, SSVPS'S B.S.Deore College of Engineering, India

[1] nehavpatil@gmail.com; [2] hitendradpatil@gmail.com

Abstract- In Web Usage Mining (WUM), Classification technique is useful in web user prediction, where user's browsing behavior is predicted. In Web prediction, classification is done first to predict the next Web pages set that a user may visit based on history of previously visited pages. Predicting user's behavior effectively is important in various applications like recommendation systems, smart phones etc. Currently various prediction systems are available based on Markov model, Association Rule Mining (ARM), classification techniques, etc. In Markov prediction model it cannot predict for a session that was not observed previously in the training set and prediction time is also constant. In ARM, Apriori algorithm requires more time to generate item sets. This paper presents review of various prediction models used to predict web users browsing behavior.

Key Words— WUM, classification, Browsing behavior, prediction, ARM, session

I. INTRODUCTION

The World Wide Web (WWW) continues to grow at an increased speed in both the volume of traffic and the size and complexity of Web sites. It is important to extract useful knowledge from this web data to capture tastes of users. Application of data mining techniques is a web mining used to extract useful knowledge from web data, consisting documents, hyperlinks between documents, web sites access logs, etc. Web mining is classified as Web Content Mining, Web Structure Mining and Web Usage Mining. Web content mining is the mining of required data, information and knowledge from content of the web page. In web structure mining graph theory is used to analyze the web site's node and connection structure. Web usage mining is performed to discover interesting usage patterns from web usage data which fulfils the needs of web-based applications [2]. Web users' identity and their surfing behavior are captured by web usage data. This web usage data is the web user access logs collected at the web server. Some of the typical data collected at a Web server include IP addresses, domain name, page references, and access time of the users etc. Web usage mining consists of three steps, namely pre-processing, pattern discovery, and pattern analysis [3]. In preprocessing raw data is cleaned to get required data, then patterns, rules and statistics are determined so that interesting rules and patterns are determined. These interesting patterns are used to predict web users browsing behavior.

It is important to predict the next web page accessed by the user in various web applications whether they are search engines or e-commerce or marketing web sites. In Web Prediction the knowledge of web pages history from user surfing behavior is maintained at the web server in form of logs which is used to predict future set of pages likely to be visited by user [1]. Such knowledge of user's navigation history within a period of time is called as a session. These sessions are the main source of data for training, obtained from the Web server logs, and they contain sequences of web pages that users have gone through along with date, time and duration etc. Predicting a user's web browsing behavior on a web-site can be beneficial to improve the web cache performance, recommendation of related pages, improve search engines, understand and influence buying patterns, and personalizes the web browsing experience. Improving the prediction process can lower the user's access times while surfing and it can be very useful to reduce network traffic by avoiding visiting unnecessary pages. There are various prediction models which are based on Markov models, classification techniques, ARM etc.

II. ARCHITECTURE OF PREDICTION MODEL

Figure 2.1 shows the general architecture of prediction model where initially client requests are maintained as logs at web server, but this data may be noisy so data preprocessing is carried out to get the clean data. Data preprocessing consists various steps like cleaning of data in which different records of web log are removed like records of videos and graphics and failed http requests, user identification and session identification in which different users and sessions are identified and path completion. Preprocessed data is used by prediction method. Prediction method is used to identify web users browsing behavior.

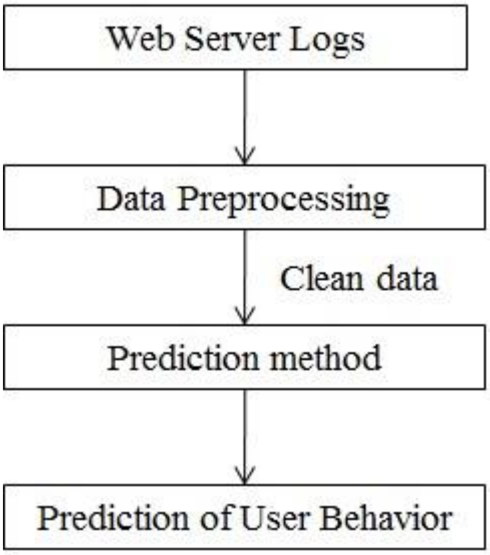

Fig 1: Working of Prediction model

The organization of this paper is as follows. In Section II the general architecture of prediction model is explained in detail. Section III will describe the related work of different prediction techniques based on Markov model, classification method, ARM etc. and we conclude the paper in Section IV.

III. RELATED WORK

There are various prediction techniques available to predict users web browsing behavior. This section describes these prediction methods. Prediction models are classified as point based and path based prediction models. In point based prediction model, prediction is done depending on currently observed user actions. In path based prediction model, prediction is done depending on user's previous path data. As path based models are able to look far in user's history of navigation, accuracy is more compare to point based models.

A. WebWatcher

WebWatcher is a recommender model presented by T. Joachims, D. Freitag, and T. Mitchell based on k Nearest Neighbor (kNN) classification technique and reinforcement learning. A WebWatcher model makes use of user's interest terms. It interactively suggests user where to go next. Hyperlinks suggested by WebWatcher knowledge base are highlighted by inserting eyeball icons around them. WebWatcher is implemented like a proxy server. When user visits a new hyperlink, it updates the log for the given search, retrieves the page if it is not prefetched already and returns copy to the user. Following target function [4] is required to do this task,

$$\text{LinkQuality}: \text{Page} \times \text{Interest} \times \text{Link} \rightarrow [0,1] \qquad (1)$$

Here link quality is described as the probability that user will choose the link given page and interest. In kNN approach, LinkQuality value for every hyperlink is estimated to be the average similarity of the k (generally 5) highest ranked keyword sets related with the hyperlink. The hyperlink is suggested only when LinkQuality value is greater than threshold value. In reinforcement learning, tours are discovered through the web such that the amount of related information is maximized to select optimal actions in certain settings [4].

B. Markov Model

Markov model is used for prediction. Three parameters < A, S, T > are used to represent Markov models, where A represents the set of all possible user actions that can be performed; S represents the set of all possible states; and T is a Transition Probability Matrix (TPM), where entry t_{ij} refers to the probability of executing the action j when the process state is i. In Web prediction, the next user action represents prediction of the next web page to be visited. The previous user actions refer to the previous web pages that have already been visited.

The basic Markov model predicts the next action by looking only at the last action of the user referred as the first-order Markov model. Second order Markov model does the predictions by looking at the last two user actions. This generalized approach is known as the K^{th} -order Markov model, which performs the predictions by considering the last K actions of user [5], [6].

C. All Kth Markov Model

Researchers found that first-order Markov models are inaccurate in predicting the user's browsing behavior in some applications, since these models are unable to look more into the past to correctly distinguish the different observed patterns [6], so higher-order models are referred. Higher-order Markov models provide high predictive accuracies but they consists more number of states which cause rise in their space and runtime requirements. To overcome some of these limitations, solution is to train different order Markov models and collectively use all of them during the prediction [1], which is provided in the All-K^{th} order Markov model. All K^{th} Markov model generates all orders of Markov models and use them collectively in prediction.

Algorithm: All K^{th} Prediction
Input: User session s of length K
Output: Next page to be visited, p

```
    1. p <-predict(s, mk)
    2. If p is not 0 then return p
    3. s <- remove first page ID from s
    4. K<- K-1
    5. If (K=0) return failure
    6. Goto step 1
    7. Stop
```

Here the function predict(s, mk) is used to predict the next web page visited of session s using the Kth-order Markov model mk. If the Markov model mk fails, then mk − 1 is consulted using a new session s' of length k − 1 where s' is calculated by removing the first page ID in s. This process continues until prediction is obtained or it fails [1]. For a given user session s = <P1, P2, P4>, all-Kth model prediction is carried out by referring third-order Markov model. If the third-order Markov model prediction fails, then the second-order Markov model is referred on the session s'= s − P1= <P2, P4> i.e. page ID P1 is removed from the session and new session is considered which consist of sequence of two pages P2 and P4. This process continues till the first-order Markov model.

D. ARM

ARM is a data mining concept that has been applied successfully to discover related transactions. ARM discovers relationships among item sets depending on their co-occurrence in the transactions. Association rule generation can be used to relate pages that are most often referenced together in a single server sessions. Association rules in web usage mining refer to the sets of pages that are accessed together with a support value greater than specified threshold. Here prediction is conducted according to the association rules that satisfy certain support and confidence as follows. For every rule of the implication, R = X → Y, X is the session of the user and Y is the target destination page. Prediction is given as follows [1]:

$$\text{Prediction}(X \rightarrow Y) = \arg\max_Y \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}, X \cap Y = \emptyset \qquad (2)$$

Here prediction can resolve to more than one page, so to keep the minimum value for support plays crucial role in performing prediction. It may be possible that there is no support for X ∪ Y, in that case prediction is computed as prediction (X' → Y), where X' is the item set of the user session after removing the first page ID in the original session. ARM consist scalability and efficiency problems. The scalability problem derived from generating item sets because with the number of item sets it takes exponential time [1].

E. *Modified Markov Model*

Modified Markov model proposed by Mamoun A. Awad and Issa Khalil in which they reduce number of paths so that it can fit in the memory and predicts faster. User sessions S1=<P3, P4> and S2=<P4, P3> in Markov model are two different sessions; therefore, each session can have different prediction probability so the prediction is resolved to different page depending on the probability. In this model all the sessions <P1, P2, P3>, <P1, P3, P2>, <P2, P3, P1>, <P2, P1, P3>, <P3, P2, P1>, and <P3, P1, P2> are considered as one set P1, P2, P3 rather than considering six different sessions also sessions having repeated pages are discarded. As this model uses set of pages instead of sequence of pages prediction will be quite different from original Markov model in which sequence of pages is considered and since each session have different probability. The Kth order of modified Markov model gives the probability such that a web user will visit the Kth page given that she has visited the $k - 1$ pages in any order [1].

IV. CONCLUSION

In this paper different prediction methods are reviewed for web prediction. Low order Markov models are not suitable sometimes since they cannot look far in the past history of user behavior and high order Markov model increases states complexity. Markov model is probabilistic in nature so it can predict for observed sessions only because a new session has zero probability. ARM possess scalability problem related to tem sets. Modified Markov model reduces the complexity in terms of size of the model compared to original Markov model.

REFERENCES

[1] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model," *IEEE Trans. on Systems, Man, and Cybernetics.*, vol. 42, no. 4, pp. 1131-1142, August 2012.

[2] Srivastava, T., Prasanna Desikan, and Vipin Kumar, "Web mining–concepts, applications and research directions," *Foundations and Advances in Data Mining. Springer Berlin Heidelberg*, pp. 275-307, 2005.

[3] Rajni Pamnani and Pramila Chawan, "Web Usage Mining: A Research Area in Web Mining," *in Proc. ISCET,* Jun. 2013.

[4] T. Joachims, D. Freitag, and T. Mitchell, "WebWatcher: A tour guide for the World Wide Web," *in Proc. I JCAI,* pp. 770–777, 1997.

[5] Nizar R. Mabroukeh and C. I. Ezeife, " Semantic-rich Markov Models for Web Prefetching," *in Proc. IEEE International Conf. on Data Mining Workshops,* pp. 200-207, Jun. 2009.

[6] Deshpande Mukund and George Karypis, "Selective Markov models for predicting Web page accesses." *ACM Transactions on Internet Technology (TOIT),* pp. 163-184, 2004.