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. 2, February 2015, pg.144 – 148

RESEARCH ARTICLE

Prediction of Link and Path for User's Web Browsing Using Markov Model

Anand Charpate, Chetan Bramhankar, Prashant Gaikawad, A.D.Londhe

Dept. of Computer Engg., Pune University, India

anand.charpate@yahoo.com chetanbramhankar1@gmail.com prashantpaddy1@gmail.com aoudumber.londhe@gmail.com

Abstract*-- World Wide Web is a huge storehouse of web pages and links. It offers large quantity of data for the Interne users. The growth of web is incredible as around one million pages are added per day. Users' accesses are recorded in weblogs. Web usage mining is a kind of mining techniques in logs[1].Because of the remarkable usage, the log files are growing at a faster rate and the size is becoming very large. This leads to the difficulty for mining the usage log according to the needs. This provides a vast field for the researchers to provide their suggestion to develop a better mining technique. In this paper, we analyze and study Markov model and all-Kth Markov model in Web prediction[2]. We propose a new modified Markov model to alleviate the issue of scalability in the number of paths. In addition, we present a new two-tier prediction framework that creates an example classifier EC, based on the training examples and the generated classifiers. We show that such framework can improve the prediction time without compromising prediction accuracy[5]. We have used standard benchmark data sets to analyze, compare, and demonstrate the effectiveness of our techniques using variations of Markov models and association rule mining. Our experiments show the effectiveness of our modified Markov model in reducing the number of paths without compromising accuracy. Additionally, the results support our analysis conclusions that accuracy improves with higher orders of all-Kth model.

***Key words*— All-Kth Markov model ,web mining, Markov model**

I. INTRODUCTION

World Wide Web (WWW) is very popular and interactive. It has become an important source of information and services. The Web is huge, diverse and dynamic. Extraction of interesting information from Web data has become more popular and as a result Web mining has attracted lot of attention in recent time [1]. Web mining can be defined roughly as data mining using data generated by the Web. Our study addresses two research questions: (i) when and to what extent are users link and path browsing on the Web? and (ii) what affects link and path browsing behavior during interaction with Web search results? To answer these questions, we analyzed browser logs, which describe natural user behaviors at scale. We collected these logs from a popular Web browser plug-in and used the data to analyze link and path browsing behavior through metrics such as pageviews, outclicks, and tab switches. We also study link and path browsing in search results to analyze user branching behavior. We conclude by discussing the implications of our findings for Web sites and browsers, search interfaces, and log analysis. Web prediction is a classification problem in which we attempt to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. Such knowledge of user's history of navigation within a period of time is referred to as a session. These sessions, which provide the source of data for training, are extracted from the logs of the Web servers, and they contain sequences of pages that users have visited along with the visit date and duration[5]. 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. Therefore, it is crucial to look for scalable and practical solutions that improve both training and prediction processes. Improving the prediction process can reduce the user's access times while browsing, and it can ease network traffic by avoiding visiting unnecessary pages. When a user is reading his current accessed page, the next predicted page is loaded into the user cache memory. It decreases the loading time for next page access at user end so that the web page retrieval efficiency will be improved[7]. The concept of web page prediction is the application comes under the web page mining along with data mining. When the page access is performed, it comes under the web content mining to locate and load the predicted page into the cache. When the history of the web server is collected in the form of user web usage history and presented in the form of web pages. Once the information database gets available, the next work is to perform the data mining operations to prediction. But generally, the size of this kind of datasets is quite large, because of this to reduce the dataset size, some clustering process is required. The clustering can be static session based clustering or an intelligent clustering using some analytical approach. Once the clustering is performed, the identification of the appropriate cluster is performed to that relates the user existence[4]. This identified cluster is selected as the working dataset based on which the prediction is performed. The prediction process is basically to identify the frequency of next visiting pages in relevancy to the current page. Once the prediction analysis is performed, the association identification is performed to identify most associated next page. This page is then selected as the next predicted web page. In this paper we did literature survey on "Users‟ future request prediction – Web Usage Mining". The various methods have been proposed on this work and this paper highlights about the various techniques advantages & limitations. The prediction process is basically to identify the frequency of next visiting pages in relevancy to the current page. Once the prediction analysis is performed, the association identification is performed to identify most associated next page. This page is then selected as the next predicted web page. The basic structural model of this working process is shown in figure 1.

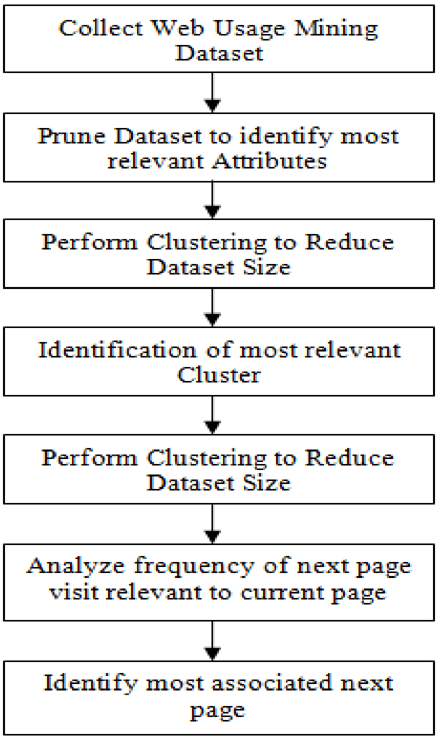

Collect Web Usage Mining Dataset
Prune Dataset to identify most relevant Attributes
Perform Clustering to Reduce Dataset Size
Identification of most relevant Cluster
Perform Clustering to Reduce Dataset Size
Analyze frequency of next page visit relevant to current page
Identify most associated next page

Figure 1 : Basic Structure of Web Page Prediction

In this paper, an improved web page prediction model is presented. The presented work is the improved with the association of three main concepts: markov model, vague rules and the association mining. Markov model will work as the intelligent prediction approach that will be filtered at two different levels using vague rules. Vague will define the intelligent rule set by performing the dataset analysis. At the later stage, the association mining will be implemented to perform the web page prediction for the caching.

II. LITERATURE SURVEY

The main focus of literature survey is to study and contrast the prediction models to predict the user's future web page requests. Prediction models are used addressing web prediction problem. The main aim is to study different prediction models to reduce user's access times and improving personalization while browsing the services of web. In addition, to reduce network traffic problems by avoiding pages visiting unintentionally and unnecessarily by users. Various prediction model like Markov model, artificial neural network's (ANN), k nearest neighbor (kNN), support vector machine (SVM), fuzzy inference, Bayesian model are proposed by researchers to predict user future request of page. Prediction models can be classified into two categories named as point-based and path based prediction models. When user's previous and historic path data are predicted then it is referred as path based prediction. Point-based prediction is based on user's current actions. Markov model by means of the anticipation-Maximization algorithm where they detachment locate punter by means of a replica-based bunch move toward. They displayed the paths for users with each cluster after partitioning the users into clusters, Our work is not a model based but space based and we worn Markov replica for forecast rather than clustering. In another paper the authors construct Markov models from log files and they use co-citation and coupling similarities for measuring the conceptual relationships between Web pages that coalesce two Markov replica and cluster process methodology for mesh page connection forecast. To Cluster conceptually related pages Citation Cluster algorithm is then proposed.

III. MOTIVATION AND RELATED WORK

Millions of users access web sites in all over the world. When they access a websites, a large amount of data generated in log files which is very important because many times user repeatedly access the same type of web pages and the record is maintained in log files[1]. These series can be considered as a web access pattern which is helpful to find out the user behavior. Through this behavior information, we can find out the accurate user next request prediction that can reduce the browsing time of web pages. In recent years, there has been an increasing number of research works done with regard to web usage mining „ Future request prediction‟. The main motivation of this survey is to know the what research has been done on Web usage mining in future request prediction. The wide usage of the Internet in various fields has increased the automatic extraction of the log data from the web sites. The usage of data mining techniques on the data collected from the web helps us pattern selection, which acts as a traditional way of decision-making tools. Web usage mining is the application of the data mining techniques on the web-collected data, which is already present in the form of various patterns. Web usage mining is presented on secondary data such as (user name, ip address, date and time, their type of browsers used, type of URL used to view the site etc.) which is deduced from the interactions of the users in between the web sessions.

IV. PROPOSED MODEL

In order to study web user navigational behaviour it will be important to clarify the system first. Web users are considered human entities that, by means of a web browser, access information resources in a hypermedia space called the World Wide Web (WWW). Common web users' objectives are information foraging (looking for information about something), social networking activities (e.g. Facebook), e-commerce transactions (e.g. Amazon Shopping), bank operations, etc. On the other hand, the hypermedia space is organized into web pages that can be described as perceived compact subunits called "web objects." The design of web pages is created by "web masters" that are in charge of a group of pages called a "web site." Therefore, the WWW consists of a vast repository of interconnected web sites for different purpose. While current approaches for studying the web user's browsing behavior are based on generic machine learning approaches, a rather different point of view is developed in this thesis. A model based on the neurophysiology theory of decision making is applied to the link selection process. This model has two stages, the training stage and the simulation stage. In the first, the model's parameters are adjusted to the user's data. In the second, the configured agents are simulated within a web structure for recovering the expected behaviour. The main difference with the machine learning approach consists in the model being independent of the structure and content of the web site. Furthermore, agents can be confronted with any page and decide which link to follow (or leave the web site). This important characteristic makes this model appropriate for heavily dynamic web sites. Another important difference is that the model has a strong theoretical basis built upon physical phenomenon. Traditional approaches are generic, but this proposal is based on a state-of-the-art theory of brain decision making. The proposal is based on the Markov's Model. The Markov's model simulates the artificial web user's session by estimating the user's page Sequences and furthermore by determining the time taken in selecting an action, such as leaving the site or proceeding to another web page. Experiments performed using artificial agents that behave in this way highlight the similarities between artificial results and a real web user mode of behavior. Furthermore, the performance of the artificial agents is reported to have similar statistical behavior to humans[6]. If the web site semantic does not change, the set of visitors remains the same. This principle enables the predicting of changes in the access pattern to web pages related to small changes in the web site that preserve the semantic. The web user's behavior could be predicted by simulation and then services could be optimized.

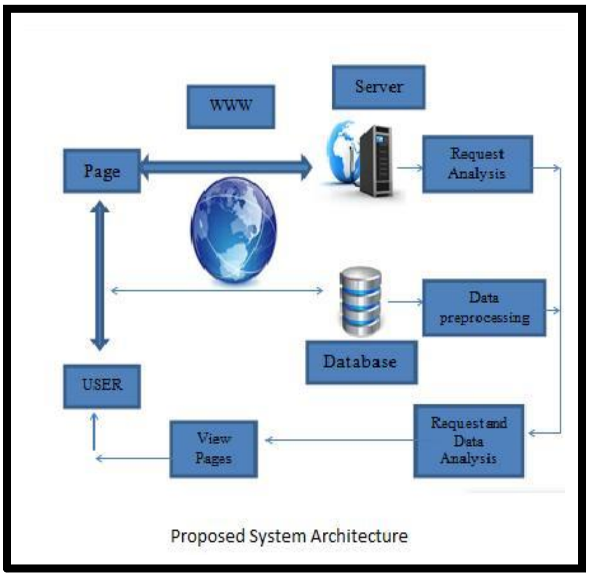


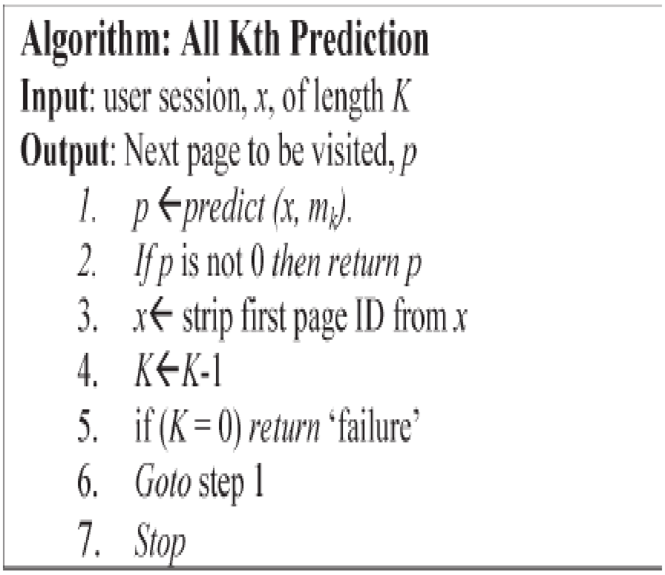
Figure 2 :Proposed System Architecture

V. Markov Model

The basic 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 Kth-order Markov model is the probability that a user will visit the kth 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. The main advantages of Markov model are its efficiency and performance in terms of model building and prediction time. It can be easily shown that building the kth order of Markov model is linear with the size of the training set . The key idea is to use an efficient data structure such as hash tables to build and keep track of each pattern along its probability. Prediction is performed in constant time because the running time of accessing an entry in a hash table is constant. Note that a specific order of Markov model cannot predict for a session that was not observed in the training set since such session will have zero probability.

All-Kth Markov Model

In all-Kth Markov model we generate all ordersof Markov models and utilize them collectively in prediction. Table I presents the steps of prediction using all-kth model. Note that the function predict(x,mk) is assumed to predict the next page visited of session x using the kth-order Markov model mk. If the mk fails, the mk−1 is considered using a new session x_ of length k − 1 where x_ is computed by stripping the first page ID in x. This process repeats until prediction is obtained or prediction fails. For example, given a user session x = _P1, P5, P6_, prediction of all-Kth model is performed by consulting third-order Markov model. If the prediction using third-order Markov model fails, then the second-order Markov model is consulted on the session x_ = x − P1 = _P5, P6_.This process repeats until reaching the first-order Markov model[1].



Algorithm: All Kth Prediction

Input: user session, x, of length K

Output: Next page to be visited, p

```
1. p ← predict (x, m_k).
2. If p is not 0 then return p
3. x ← strip first page ID from x
4. K ← K-1
5. if (K = 0) return 'failure'
6. Goto step 1
7. Stop
```

Figure 3 :ALL K-th Algorithm

VI. CONCLUSION

The paper gives a brief literature survey of research field in web user browsing prediction. The higher order markov model is suited and found to be best for methodology to implement. The frame work included the concept of variable length markov model and page rank , page rank concept may be used when the website is newly launched and the weblog is not sufficiently created so page rank may be used to predict the page and it may be also used when the ambiguity will arrived in the markov model.

ACKNOWLEDGEMENT

We would like to express our deep sense of gratitude to express our gratitude and sincere thanks to our Head of Department Prof. V. N. NANDGAONKAR and Project Guide Prof. A.D.LONDHE for their assistance in solving various difficulties. The acknowledgement will be incomplete if we do not record our sense of gratitude to our Principal, Dr. C. L. PRABHUNE who gave us necessary guidance, encouragement by providing us with all facility available to work on project.

REFERENCES

[1] Dembczynski, K., Kotłowski, W., Sydow, M.: Effective Prediction of Web User Behaviour with User-Level Models, 2007.

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

[3] Marcelo Maia, Jussara Almeida, and Virg ılioAlmeida . Identifying user behavior in online social networks. In Proceedings of the 1st Workshop on Social Network Systems ,Social Nets'08,pages1–6, NewYork, NY, USA, 2008. ACM.

[4] M. T. Hassan, K. N. Junejo, and A. Karim, "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] HAUGER, D., PARAMYTHIS, A., AND WEIBELZAHL, S. 2011. Using browser interaction data to determine page reading behavior. In Proceedings of the 19th International Conference on User Modelling, adaption, and Personalization (UMAP).147–158

[6] LEIVA, L. A. 2011. Mining the browsing context: Discovering interaction profiles via behavioral clustering. In Adjunct Proceedings of the 19th Conference on User Modeling, Adaptation, and Personalization (UMAP) . 31–33

[7] GUO,Q. AND A GICHTEIN , E. 2012. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behaviour. In Proceedings of the 21st International Conference on World Wide Web (WWW). 569–578.