Web Page Prediction Techniques: A Review

Sunil Kumar[#1], Ms. Mala Kalra[*2]

[#] Assistant Professor, Computer Science Department
MIT, Moradabad (Uttar Pradesh) (INDIA)

[*] Assistant Professor, Computer Science and Engineering Department
NITTTR, Chandigarh (INDIA)

Abstract— **this paper proposes a survey of Web Page Prediction Techniques. Prefetching of Web page has been widely used to reduce the access latency problem of the Web users. However, if Prefetching of Web page is not accurate and Prefetched web pages are not visited by the users in their accesses, the limited bandwidth of network and services of server will not be used efficiently and may face the access delay problem. Therefore, it is critical that we have an effective prediction method during prefetching. The Markov models have been widely used to predict and analyse user‘s navigational behaviour. All the activities of web users have been saved in web log files. The stored users' session is used to extract popular web navigation paths and predict current users' next web page visit.**

Keywords— **Web Usage Mining, Clustering, Markov Model, User Sessions, N-Grams**

I. Introduction

World Wide Web (WWW) [1] is a collection of data which can be accessed by Web Browser. The WWW is just a subset of Internet. The WWW is a conceptual but Internet is physical aspect like cable, router, switch etc. The Internet is the actual network of networks where all the information presents. The Hyper-Text Transfer Protocol (HTTP) and File Transfer Protocol (FTP) are the methods used to transfer Web pages over the Network. Hypertext is a text which contains an address of another file or data.

Web mining research works with other areas, including data mining, text mining, information retrieval, and Web retrieval. The classification is based on two aspects: the purpose and the data sources. Mining research concentrates on finding new information or knowledge in the data. On the basis of above information, Web mining can be divided into web structure mining, web content mining, and web usage mining as shown in Figure 1.

Web Content Mining [2] [3] [4] is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in web content mining.

Web usage mining [3][5] is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve needs of Web based applications. It consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. Web servers, proxies, and client applications can quite easily capture data about Web usage.

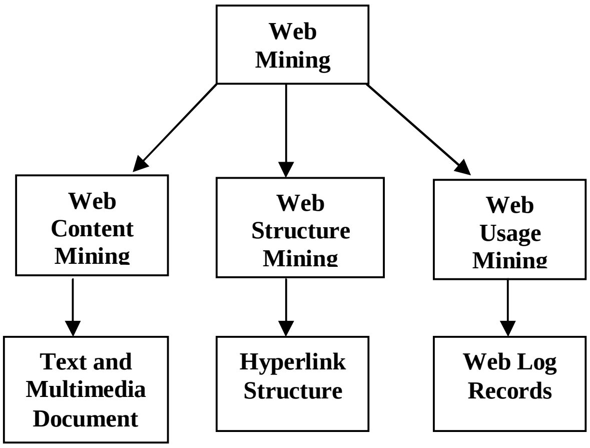

Figure 1: Types of Web Mining

The goal of web structure mining [6] is to generate structural summary about the website and web page. The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location.

II. MARKOV MODELS FOR PREDICTING USER’S ACTIONS

As discussed in the introduction, techniques derived from Markov models [7] have been extensively used for 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 < *X*; *Y*; *T* >, where *X* is the set of all possible *actions* that can be performed by the user; *Y* is the set of all possible states for which the Markov model is built; and *T* is a jY_j _ jX_j *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*. 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. A somewhat more complicated model computes the predictions by looking at the last two actions performed by the user. This is called the *second-order Markov model*, and its states correspond to all possible pairs of actions that can be performed in sequence.

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. For example, suppose the prediction of next page accessed by a user on a website is a problem. The input data for building Markov models consists of *web-sessions*, where each session consists of the sequence of the pages accessed by the user during his/her visit to the site. In this problem, the actions for the Markov model correspond to the different pages in the web site, and the states correspond to all consecutive pages of length *K* that were observed in the different sessions. In the case of first-order models, the states will correspond to single pages, in the case of second-order models, the states will correspond to all pairs of consecutive pages, and so on. Once the states of the Markov model have been identified, the transition probability matrix can then be computed. There are many ways in which the TPM can be built.

The most commonly used approach is to use a *training* set of action-sequences and estimate each t_{ji} entry based on the frequency of the event that action a_i follows the state s_j. For example consider the *web-session WS*2 (*P*3; *P*5; *P*2; *P*1; *P*4) shown in Figure 2. If they are using *first-order Markov model* then each state is made up of a single page, so the first page *P*3 corresponds to the state *s*3. Since page *p*5 follows the state s_3 the entry t_{35} in the TPM will be updated. Similarly, the next state will be s_5 and the entry t_{52} will be updated in the TPM. In the case of higher-order model each state will be made up of more than one actions, so for a second-order model the first state for the *web-session WS*2 consists of pages {*P*3; *P*5} and since the page *P*2 follows the state {*P*3; *P*5} in the web-session the TPM entry corresponding to the state {*P*3; *P*5} and page *P*2 will be updated. Once the transition probability matrix is built making prediction for web sessions is straight forward. For example, consider a user that has accessed pages {*P*1; *P*5; *P*4}. If they want to predict the page that will be accessed by the user next, using a first-order model, we will first identify the state *s*4 that is associated with page *P*4 and look up the TPM to find the page *pi* that has the highest probability and predict it. In the case of our example the prediction would be page *P*5.

Web Sessions:

*WS*1: {*P*3; *P*2; *P*1}
*WS*2: {*P*3; *P*5; *P*2; *P*1; *P*4}
*WS*3: {*P*4; *P*5; *P*2; *P*1; *P*5; *P*4}
*WS*4: {*P*3; *P*4; *P*5; *P*2; *P*1}
*WS*5: {*P*1; *P*4; *P*2; *P*5; *P*4}

1st Order	P1	P2	P3	P4	P5
S1={P1}	0	0	0	2	1
S2={P2}	4	0	0	0	1
S3={P3}	0	1	0	1	1
S4={P4}	0	1	0	0	2
S5={P5}	0	3	0	2	0

2nd Order	P1	P2	P3	P4	P5
{P1;P4}	0	1	0	0	0
{P1;P5}	0	0	0	1	0
{P2;P1}	0	0	0	1	1
{P2;P5}	0	0	0	1	0
{P3;P2}	1	0	0	0	0

2nd Order	P1	P2	P3	P4	P5
{P2;P5]	0	1	0	0	0
{P2;P4}	0	0	0	0	1
{P4;P5}	0	2	0	0	0
{P5;P2}	3	0	0	0	0
{P3;P4}	0	0	0	0	1

Figure 2: Sample web sessions with the corresponding 1st & 2nd order Transition Probability Matrices.

III. CLUSTERING

Clustering [8] is the most important *unsupervised learning* problem. So the simple definition of Clustering could be “the process of organizing objects into groups whose members are similar in some way”. A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. We can show this with a simple example:

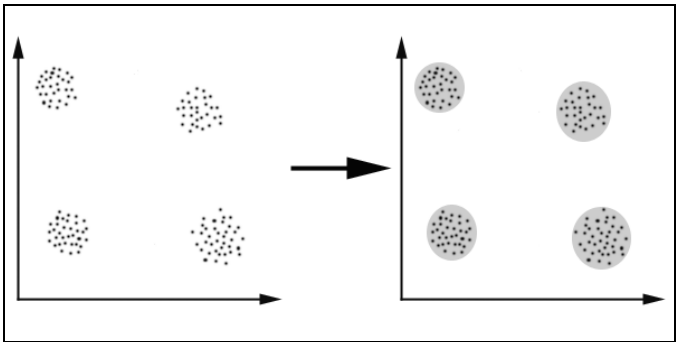

Figure 1.2 Clustering

IV. RELATED WORK

Recommendation systems are one of the early applications of Web prediction. Joachims *et al.* [9] proposed the Web Watcher which is a path-based recommender model based on *k*NN and reinforcement learning. The system contains some properties like (a) WebWatcher provides several types of assistance but most importantly highlights interesting hyperlinks as it accompanies the user. (b) It learns from experience. (c) WebWatcher runs as a centralized server so that it can assist any Web user running any type of Web browser as well as combine training data from thousands of different users.

Su et al. [10] 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. Basically N-gram is a collection of N visited web pages by user. Their work is aimed at showing that using simple n-gram models for n greater than two will result in significant gain in prediction accuracy while maintaining reasonable applicability. They proposed path-based model for web page prediction. Their path-based model is built on a web-server log file L. They consider L to be reprocessed into a collection of user sessions, such that each session is indexed by a unique user id and starting time. Each session is a sequence of requests where each request corresponds to a visit to a web page (an URL). The log L then consists of a set of sessions. Their algorithm builds an n-gram prediction model based on the occurrence frequency. Each sub-string of length n is an n-gram. These sub-strings serve as the indices of a count table T. During its operation, the algorithm scans through all sub-strings exactly once, recording occurrence frequencies of the next click immediately after the substring in all sessions. The maximum occurred request is used as the prediction for the sub-string.

Levene and Loizou [11] computed the information gain from the navigation trail to construct a Markov chain model to analyse the user navigation pattern through the Web. Navigation through the web, colloquially known as "surfing", is one of the main activities of users during web interaction. When users follow a navigation trail they often tend to get disoriented in terms of the goals of their original query and thus the discovery of typical user trails could be useful in providing navigation assistance. Herein they give a theoretical underpinning of user navigation in terms of the entropy of an underlying Markov chain modelling the web topology. They present a novel method for online incremental computation of the entropy and a large deviation result regarding the length of a trail to realise they said entropy. They provide an error analysis for our estimation of the entropy in terms of the divergence between the empirical and actual probabilities. They also provide an extension of our technique to higher-order Markov chains by a suitable reduction of a higher-order Markov chain model to a first-order.

M. Deshpande, G. Karypis [12], presented a class of Markov model-based prediction algorithms that are obtained by selectively eliminating a large fraction of the states of the All-Kth-Order Markov model. Their experiments on a variety of datasets have shown that the resulting Markov models have a very low state-space complexity and at the same time achieve substantially better accuracies than those obtained by the traditional algorithms.

M. Awad and L. Khan [13] [14] have successfully combined several effective prediction models along with domain knowledge exploitation to improve the prediction accuracy. However, the module endures expensive training and prediction overheads because of the large number of labels/classes involved in the WPP.

M. T. Hassan, K. N. Junejo, and A. Karim [15] presented Bayesian models for two things like learning and predicting key Web navigation patterns. Instead of modeling the general problem of Web navigation they focus on key navigation patterns that have practical value. Furthermore, instead of developing complex models they present intuitive probabilistic models for learning and prediction. The patterns that they consider are: short and long visit sessions, page categories visited in first N positions, range of page views per page category, and rank of page categories in first N positions. They learn and predict these patterns under four settings corresponding to what is known about the visit sessions (user ID and/or timestamp).

F.Khalil, J. Li, H. Wang [16] improved the Web page access prediction accuracy by integrating all three prediction models: Markov model, Clustering and association rules according to certain constraints. Their model, IMAC, integrates the three models using lower order Markov model. Clustering is used to group homogeneous user sessions. Low order Markov models are built on clustered sessions. Association rules are used when Markov models could not make clear predictions. The integrated model has been demonstrated to be more accurate than all three models implemented individually, as well as, other integrated models. The integrated model has less state space complexity and is more accurate than a higher order Markov model.

Bhawna Nigamand Dr. Suresh Jain [17] proposed three different Prefetching and Caching schemes i.e. Prefetching only, Prefetching with Caching and Prefetching from caching. Dynamic Nested Markov model is used for predicting next accessed web page. The Experimental result shows that the Prefetching with caching scheme will give good results. By applying these schemes, users' web access latency can be minimized and quality of service can be provided to the web user.

Mamoun A. Awad and Issa Khalil [18], analysed and studied Markov model and all- Kth Markov model in Web prediction. They proposed a new modified Markov model to alleviate the issue of scalability in the number of paths. They have used standard benchmark data sets to analyse, compare, and demonstrate the effectiveness of our techniques using variations of Markov models and association rule mining. Their experiments show the effectiveness of modified Markov model in reducing the number of paths without compromising accuracy. Additionally, the results support their analysis conclusions that accuracy improves with higher orders of all-Kth model.

Poornalatha G, Prakash S Raghavendra [19][20][21] presented a paper to solve the problem of predicting the next page to be accessed by the user based on the mining of web server logs that maintains the information of users who access the web site. The prediction of next page to be visited by the user may be pre fetched by the browser which in turn reduces the latency for user. Thus analysing user's past behavior to predict the future web pages to be navigated by the user is of great importance.

IV. Conclusions

World Wide Web has necessitated the users to make use of automated tools to locate desired information resources and to follow and asses their usage pattern. Web page prefetching has been widely used to reduce the user access latency problem of the internet; its success mainly relies on the accuracy of web page prediction. Markov model is the most commonly used prediction model because of its high accuracy. Low order Markov models have higher accuracy and lower coverage. The higher order models have a number of limitations associated with i) Higher state complexity, ii) Reduced coverage, iii) Sometimes even worse prediction accuracy. Clustering is one of the best solutions for resolving the problem of worse prediction accuracy of Markov model. It is a powerful method for arranging users' session into clusters according to their similarity. We have discussed some of the techniques to overcome the issues of web page prediction. As the web is going to expand, web usage in web databases will become more and more. The above findings will become good guide in web page prediction effectively. In this paper, we have presented a comprehensive survey of up-to-date researchers of web page prediction. Besides, a brief introduction about web mining, clustering and web page prediction have also been presented. However, research of the web page prediction is just at its beginning and much deeper understanding needs to be gained.

V. Future Work

This survey paper will help to upcoming researchers in the field of web page prediction to know the available methods. This paper will also help researcher to perform their research in right direction. In future, researcher can work on Markov model to enhance the accuracy of web page prediction. First order Markov model is based on the assumption that the next state to be visited is only a function of the current one. The first-order Markov models (Markov Chains) provide a simple way to capture sequential dependence, but do not take into consideration the long-term memory aspects of web surfing behavior. Higher-order Markov models and hidden Markov models are more accurate for predicting navigational paths. Researcher can get better result if they will do the pre-processing phase effectively. Markov model and Clustering can work together and provide better prediction results without compromise with accuracy.

References

[1] "WEB (World Wide Web)". Available at http://compnetworking.about.com/cs/worldwideweb/g/bldef_www.htm. [Online]

[2] S.K.Madria, S.S.Bhowmick, W.K.Ng, and E.P.Lim. " Research issues in Web data mining". In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK 1999.

[3] Raymond Kosala, Hendrik Blockeel, "Web Mining Research": A Survey, ACM SIGKDD Explorations Newsletter, Volume 2 Issue 1, June 2000.

[4] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. "Web mining: knowledge discovery on the Web". In Proceedings of Systems, Man, and Cybernetics IEEE SMC Conference, 1999.

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan, "Web Usage Mining: Discovery and

Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations Newsletter, Volume 1 Issue January 2000.

[6] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A.Tomkins, D. Gibson, and J. Kleinberg, "Mining the Web's link structure". Computer, 32(8):60–67, 1999.

[7] M. Deshpande, G. Karypis, "Selective Markov Models for Predicting Web Page Accesses," ACM transactions on Internet Technology, volume 4, No.2, pp.163-184, May 2004

[8] Teknomo, Kardi. "K-Means Clustering". Conferences in Research and Practice in Information Technology, Volume 74. July 2007.

[9] T. Joachims, D. Freitag, and T. Mitchell, "WebWatcher: A tour guide for the World Wide Web", in Proceedings of IJCAI, pp. 770–777, 1999.

[10] Z. Su, Q. Yang, Y. Lu, and H. Zhang, "WhatNext: A prediction system for Web requests using n-gram sequence models", in Proceedings of 1st Int. Conference Web Inf. Syst. Eng. Conference, Hong Kong pp. 200–207, Jun. 2000.

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

[12] M. Deshpande, G. Karypis, "Selective Markov Models for Predicting Web Page Accesses," ACM transactions on Internet Technology, volume 4, No.2, pp.163-184, May 2004

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

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

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

[16] F.Khalil, J. Li, H. Wang, "An Integrated Model for Next Page Access Prediction", Inderscience Enterprises Ltd., 2009.

[17] Bhawna Nigamand Dr. Suresh Jain, "Analysis of Markov Model on Different Web Prefetching and Caching Schemes", 978-1-4244-5967-4/10/ 2010 IEEE

[18] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model," IEEE Trans. Syst., Man,Cybern. A, Syst., Humans, volume 42, no. 4, pp., Aug. 2012.

[19] Poornalatha G, Prakash S Raghavendra, "Web Page Prediction by Clustering and Integrate Distance Measures" IEEE/ ACM Trans. Syst., Man,Cybern. A, Syst., Humans, volume 44, no. 2, pp., Sep. 2012.

[20] J.Han and M.Camber, "Data Mining - Concepts and techniques ", University of Illinois at Urbana-Champaign, Second Edition-2000

[21] B. Hay, G. Wets, K. Vanhoof, "Mining Navigation Patterns Using a Sequence Alignment Method," Journal of Knowledge and Information Systems, Springer-Verlag, pp. 150-163, 2004