WEB PAGE ANTICIPATION SYSTEM USING MARKOV MODEL

Shivangi Sorout  
Department of computer science & Application  
Maharishi Dayanand University (M.D.U)  
Rohtak, Haryana

Dr. Balkishan (Assistant Professor)  
Department of computer science & Application  
Maharishi Dayanand University (M.D.U)  
Rohtak Haryana

Abstract: In the era of digital world, Information can be accessed with the help of Internet on the World Web. Due to the enormous growth of digital information, the limited bandwidth of the network is not utilized in an efficient way. With the help of this “Web Page Anticipation System”, we are trying to get rid of this problem. It provides us framework to incorporate the usage of pre-fetching mechanism, Clustering, Markov Model and Prediction Architecture. This framework allows us to anticipate the web page in advance with the help of user’s currently accessed web page.

Keywords: Web page anticipation, Pre-fetching, Clustering, Markov model, Users sessions

1. INTRODUCTION

Owing to the enormous growth of World Wide Web, congestion and overloading of server occurs. Due to problems incurred by the server in managing large information, latency of communication channel is substantially reduced. Various techniques for latency reduction are web catching, pre-fetching and preopening. Need of this web page anticipation system is required in the era of e-commerce digital world where every transaction is dependent only on the efficiency of how fast we are able to access the required web page within the particular time slot. Researchers use different kind of techniques comprising Markov Model for next web page anticipation, clustering and prediction architecture. For implementing this web page anticipation model, navigational behavior of the current users is stored in the web log files.

After the identification of the navigational behavior of current users, clustering is performed. Clustering is the main ingredient used in the exploratory data mining and commonly used in the statistical data analysis. Main task of clustering is to segregate the group of objects into different groups so that all objects in the one particular group are of similar nature. Scope of clustering is wide and used in the many fields like machine learning, image analysis, information retrieval, bioinformatics, data compression and computer graphics. Clustering models used in the Clustering are connectivity models, centroid based model (k-means algorithm), distribution model expectation-maximization model) and density model.

When the Clusters are formed we have to use the prediction algorithm to predict the next possible states. After that Markov model is applied on the clustering sets etc. Markov model is a stochastic model which is used in the designing the continuously changing system in variable time at different time slots. In this model future state is dependent only on the current state but also on the previous states covered. In the Markov model we have to train the model by estimating the transition probability which is denoted by:

\[ A_{ij} = P(Q(t+1)=Si | Qt=Sj) \]

Where, \( A_{ij} \) is the probability of going to the new state \( Si \) at time \( t+1 \) from state \( S_j \) at time \( t \).

![Markov Model States and their Transition Probability](image)

The first order Markov model provides us a simple way of to accumulate sequential dependence, however it does not take the aspect of long term memory in the navigation path. But with the increase of order of the Markov model subsequently there will be exponential increase in the complexity of state space. In turn we require the huge amount of data. A system with high number of states increase, systems which need to predict fast the accuracy is plummeted to large extent. So the need of the hour is to have that kind of system which predict fast with enhanced accuracy.

2. RELATED WORK

The review of past investigation serves as a guide to the researchers as it avoids duplications in the field. The knowledge of what has already been done in the field of investigation regarding the methods used for data is important. Research in any field implies a step ahead in exploration of the unknown. Any researcher to be able to take this step should be adequately prepared for it. One such preparation is gathering of
knowledge of much has already been done in the given field. A step towards unkno wn can only be taken after the review of literature an d research done in that ar ea. Any research without s uch a review is likely to be a building w ithout foundation. Thus, the review o f related literature is an indispensable step in research.

Yang et al. (2004) studied different association rule-based methods for web request prediction. The association rules for web access prediction involve dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions. Five different representations of association rules are: Subsitute rules, Subsequence rules, Lower subsequence rules, Substring rules and Lastest substring rules. The author concerned the p-revision of these five as association rules representations using different selection methods, the latest substring rules were proven to have the highest precision with decreased number of rules [1].

Liu et al. (1998) introduced a customized marketing based on the web approach using a combination of clustering and association rules. The author collected information about customers using forms, W eb server logs and cookies. It categorized customers according to the information collected. Since k-means clustering algorithm works only with numerical data, the authors used PAM (Partitioning around Round Models) algorithm to cluster data using categorical scales. Then perform association rule techniques on each cluster [2].

Kim et al. (2004) introduced combination of association clustering and link-based association rules. Improvement is the performance of Markov model, sequential association rules, as association rules are used. Sequential association rules cannot cover the state, sequential association rules are used. If sequential association rules cannot cover the state, association rules are revised. If association rules are improved, then the clustering algorithm is applied. The author’s work improved recall and it did not improve the Web page prediction accuracy [3].

Vakali et al. (2003) categorized web data clustering into two classes (I) users’ sessions-based and (II) link-based. The former uses the web log data and tries to group together a set of users’ navigation s sessions h aving similar characteristics. If web log data provide information about activities performed by a user, then the model enters a web site to the moment the same user leaves it. The records of users’ actions within a web site are stored in a log file. Each record in the log file contains the client’s IP address, the date and time the record was generated and the action performed. This extends the space of the model and clustering. Linear regression models can be exploited. All the methods evaluated using real data sets obtained from the Internet Archive, Wikipedia and a nd Yahoo! ranking lists. It also study the temporal robustness of the prediction framework. Overall the system constitutes a set of tools for high accuracy page rank prediction which can be used for efficient resource management by search engines [8].

Spiliopoulou et al. (1999) investigated web e-commerce site design is currently based on interests of web site visitors and assumptions about their exact behavior. Concrete knowledge on the way avatars navigate web pages is essential. Web site owners assume that important information is exactly where the user is looking for it. Web utilization miner tool can provide such knowledge. The general problem a addressed is given a number of pages to be ranked over the page hierarchy.
Mukhopadhyay et al. (2011) studied about pre-fetching models based on e decision t rees, M arkov chains, and d p ath analysis. The author described increase uses of dynamic pages, frequent changes in site structure and user access patterns have limited th e efficacy of these t atical techniques. O ne of the techniques that are used for improving user latency is Caching and another is W eb pre-fetching. A approaches that bank solely on caching offer limited performance improvement because it is d ifficult for caching to ha ndle t he large um ber of f increasingly di verse files. An agent based method is proposed here to cluster related pages into different categories based on the access patterns. Additionally page ranking is used to build up the prediction model at the initial stages when users are yet to invoke any page [10].

Kumar et al. (2011) presented web p rovides a corpus of f design e samples unpa rallel e in h um an hi story. L everaging existing designs to produce new pages is difficult. The author introduced the Bricolage algorithm f or automatically transferring design and content between W eb pages. Bricolage introduces a novel structured prediction technique that learns to create coherent mappings between pages by training on human-generated exemplars. The produced mappings can then be used to au tomatically t ransfer t he content f rom o ne page i nto t he style a nd layout of another. The author showed that Bricolage can learn to accurately reproduce human p age m appings, and t hat it provides a general, efficient, and automatic technique forargeting content between a variety of real web pages [11].

Dutta et al. (2011) studied web page pr ediction p lays an important role by predicting and fetching probable web page of next request in advance, resulting in reducing the user latency. The users surf the internet either by entering URL or search for some topic or through link of same topic. For searching and for link pr ediction, c lustering plays a ni mportant role. W eb page prediction model gives us significant importance t o the us er’s interest us ing t he c lustering t echnique and the navi gational behavior of t he us er through M arkov m odel. The c lustering technique is us ed f or the a ccumulation of f t he s imilar w eb pages. S imilar w eb pages of a me type r eside in the s ame cluster, the c luster c ontaining w eb pages h ave t he s imilarity w ith respect to topic of the s ession. The c lustering algorithms considered are K -means and K -medioids, K is determined by HITS algorithm. F inally, the predicted web pages are stored in form of f a s estructural au tomata to make t he s ystem more m or e m ory efficient [12].

Su et al. (2000) studied the rapid development of i nternet has resulted in more and more multimedia in web content. The author studied due to the limitation in the bandwidth and huge size of the multimedia data, users always suffer from long time waiting. The author describe that to predict the web object or page that the user most likely will view next while the user is viewing the current page, and pre-fetch the content. Then the perceived n etwork latency can be s ignificantly reduced. T he author introduced n-gram based model to utilize path profiles of users from very large web log to predict the users’ future requests. Model is based on a s imple e xtension o f e xisting point-based models for such predictions, but results show that by sacrificing the applicability somewhat one can gain a great deal in pr ediction time. T he r esults e nable an p otentially b e applied to a w ide range of applications on the web, including pre-fetching, enhancement of recommendation systems as well as web ca ching p olicies. T he e xperiments b ased o n t hree realistic web logs have proved the effectiveness of the proposed scheme [13].

Zukerman et al. (2009) used Artificial Intelligence-related techniques to predict user requests. T he author implement a learning algorithm such as some variation of M arkov chains and use a previous access log in order to train it. This approach also relies on tracking user patterns. Furthermore, it does not handle newly introduced pages, or old pages that have changed substantially. T his approach also requires a r ether long sequence of clicks from a user to learn his/her access patterns [14].

Safra nov et al. (2010) introduced the Page Rank based pre-fetching technique which is a server-side approach and uses the information about the link structure of the pages and the current and past user access to drive pre-fetching. T he approach is effective for access t o w eb p ages c lusters, is computationally efficient and scalable, and can immediately sense and react to changes in the link structure of web pages. Furthermore, the underlying algorithm uses relatively simple matrix operations and is easily parallelizable, making it suitable for clustered server environments [15].

Padminabh an et al. (1995) investigated ways of optimizing retrieval latency. W eb ca ching has been r ecognized as an effective solution to minimize user access latency. A method of called pre-fetching introduced in which clients in collaboration with server pre-fetch web page that the user is likely to access soon, viewing the current page. The benefit of f pre-fetching is that it provides low retrieval latency for users, which can be explained as high hit ratio. This approach reduces web latency b y p re-fetching be tween caching, p roxies, and browsers. Web pre-fetching has involved the important issue of log file processing and the determination of user transactions (sessions). It provides various data mining a lgorithms for the path traversal patterns and how to efficiently mine the access patterns from the web logs [16].

3. PROPOSED MODEL

Web page is the integration of the different web page contains frames, graphics and other information. User cache is used in this model to cache the frequently accessed web page. In the proposed model, a request t he server then server will send the URL of that web page to the predictor. A fter that predictor check the required specific web page. If it is present then predictor gives that web page to the server and the client give required page to the client to meet its requirements. B ut the predictor while sending the page to the server (in case the predictor i s n ot able to access c h eck t he predicted page), also give the client’s requested web page to t he p redictor i n t he p rocess of w eb pa ge request is that it uses the data structure for storing the web pages. [17]
3.1 FLOW CHART OF REQUIRED MODEL

The following flow chart presents the required model and the requisite steps to implement the required Markov model, which in turn helps in web page anticipation. In the proposed model, first step is to give input with the help of preprocessing the web server log files, after which similar web sessions are allocated to appropriate categories. By using clustering, we decide the number of clusters and among these clusters web sessions are partitioned. The process of clustering gives us the clustered data which is used for the Markov model approach.

When the Markov model is applied we have to decide the prediction algorithm and then apply the hidden Markov model in prediction algorithm. After the determination of the prediction algorithm, the next web page for user access is available as output.

1. STEP1: web server log files are preprocessed in a way such that similar web sessions are allocated to appropriate categories.

2. STEP2: Convert the web pages into numeric form and store in web.dat file and then determine the number of clusters and then division of web sessions into clusters is done using clustering tool based on FCM algorithm on Matlab.

The FCM Algorithm uses the k-means clustering to choose the number of clusters k. In this clusters centers U1 to Uk are decided after that we could pick k data points and set cluster centers to the data points or they could be assigned randomly a sign for points to clusters and take means of these clusters. For each data point, we decide the cluster center it is closest to and assign the data point to this cluster.

After that with the help of the findcluster command on the command prompt on the matlab, the web.dat file is loaded on Matlab.

3. STEP3: Perform Markov m odel analysis on each of the clusters and then make scaled Transition probability matrix and squared Emission Probability Matrix.

Making of squared Transition Probability Matrix (Rows=Columns=Total number of unique web pages=9)

TRANSITION (I,J) is the probability of transition from state I to state J.

Making of squared Emission Probability Matrix (Rows=Columns=Total number of unique web pages=9)

EMISSION(K,L) is the probability of transition from state K to L.

4. STEP4: Design the prediction algorithm and use hidden Markov Model approach. This algorithm gives us information about the next page with the help of user’s currently accessed web page.

ALGORITHM:-

1. Set Tr=Transition square matrix, Tr(I,J)=Probability of transition from state I to state J. [Initialize Tr]

2. Set E=Transition square matrix, Tr(K,L)=Probability of transition from state K to state L. [Initialize E]

3. Set seq=sequence of user’s accessed web page. [Initialize seq]

4. numStates=number of states and size=size of any column Tr.

5. numStates=size

6. L=length of seq

7. Repeat for count 1 To L

8. Repeat for state 1 to numStates

9. Set bestVal=0 [initialize bestVal]

10. Set bestPTR=0 [initialize bestPTR]

11. Repeat for inner 1 to numStates

12. Val=Tr[innerstate]

13. If Val> bestVal

14. bestVal=Val

15. bestPTR[inner]=inner

[End of if structure]

[End of step 10 inner loop]

16. PTR[state,count]=bestPTR

17. v[state]=E[state,seq[count]]+ bestval

18. vOld=v

Figure 2. flow chart for web page anticipation
sessions were divided into varying clusters using k-means function in turn reduces the number of unique pages. The accesses webpage. This will give a next probable webpage for user's currently accessed page and Next webpage for user's currently accessed web page. The reported accuracy is the how many Next web pages is the number of clusters and equivalent to \(O(n^2 \log n)\) time complexity. This is because it requires space to store the data matrix. It is feasible to store the data matrix in a secondary memory and then the space complexity will become \(O(k)\). k-means algorithm is more time efficient than hierarchical clustering algorithms with \(O(n^2)\) space complexity.

All clustering runs have performed on a desktop PC with a Pentium IV Intel processor running at 2 GHz with 2 GB of RAM and 100 GB of hard disk memory. The runtime of the k-means algorithm, regardless of the distance measure used, is equivalent to \(O(nkl)\), where \(k\) is the number of clusters and \(l\) is the number of iterations taken by the algorithm to converge. For experiments, \(n\) and \(k\) are fixed, the algorithm has a linear time complexity in terms of the size of the data set. The k-means algorithm has a \(O(k + n)\) space complexity. The prediction algorithm for the determination of user's next probable page with the help of user’s currently accessed page is applied on the number of clusters. The results gives us the accuracy of the next page access and prediction by implementing the k-means clustering algorithm on the data set decided previously. More accuracy describes most accurate web page predicted with the help of prediction algorithm implementation as user want next time. Now user do not need to request the web page as user wants, because it has been available before the time user want it next time. In this way we are able to utilize the minimum bandwidth of the user and the use of pre-fetching and clustering reduces the request made by user and reduces request time. The prediction accuracy achieved is an improvement over the previous results. This paper in introduced the Prediction algorithm for automatically transferring Web pages. It demonstrated that it can learn to closely reproduce human mappings, and it takes a number of steps towards a powerful new paradigm for instance-web based de-sign a nd opens up ex citing ar eas for research. A t present, the algorithm employs only about thirty simple visual and semantic features. Expanding this set to include more complex and sophisticated properties, such as those based on computer vision, will lik ely improve the robustness of the machine learning algorithm. Additionally, this is an implementation that cannot handle idiosyncrasies of modern HTML. Extending Prediction to these technologies remains future work.

### 5. CONCLUSION

This paper describes the overall prediction accuracy by grouping the data set sessions into clusters and reduces the web latency time. The web pages in the user sessions are segregated into categories according to Web services that are functionally meaningful. Then k-means clustering algorithm is implemented using the most appropriate number of clusters. The Prediction algorithm for the determination of user's next probable page with the help of user's currently accessed page is applied on the number of clusters. The results gives us the accuracy of the next page access and prediction by implementing the k-means clustering algorithm on the data set decided previously. More accuracy describes most accurate web page predicted with the help of prediction algorithm implementation as user want next time. Now user do not need to request the web page as user wants, because it has been available before the time user want it next time. In this way we are able to utilize the minimum bandwidth of the user and the use of pre-fetching and clustering reduces the request made by user and reduces request time. The prediction accuracy achieved is an improvement over the previous results. This paper in introduced the Prediction algorithm for automatically transferring Web pages. It demonstrated that it can learn to closely reproduce human mappings, and it takes a number of steps towards a powerful new paradigm for instance-web based design a nd opens up exciting areas for research. At present, the algorithm employs only about thirty simple visual and semantic features. Expanding this set to include more complex and sophisticated properties, such as those based on computer vision, will likely improve the robustness of the machine learning algorithm. Additionally, this is an implementation that cannot handle idiosyncrasies of modern HTML. Extending Prediction to these technologies remains future work.

### 7. REFERENCES


