An Improved Markov Model Approach to Predict Web Page Caching

Meenu Brala  
Student, JMIT, Radaur  
meenubrala@gmail.com

Mrs. Mamta Dhanda  
Asstt. Prof, CSE, JMIT Radaur  
mantanain@gmail.com

Abstract

Optimization of Web page search and the web page access is always the major issue for a web user, because of this there is always some scope to improve the web page access based on user requirement. One of such approach is given by web page pre-fetching. The pre-fetching concept is about to avail the web page to user before the user demand. It means as the user is visiting some page the next page that he visit will be copied to the user cache. In this present work we are performing an intelligent web page prediction approach based on history of web page visit. The proposed approach is three level approach in which we have combined markov model along with association mining and clustered approach. As the next web page will be predicted it will perform the efficient web page access.

Keywords : Prediction, Markov, Clustering, Association Mining, Prediction.

1. Introduction

The World Wide Web is a huge information repository. When so many users access this information repository, it is easy to find certain patterns in the way they access web resources. Web request prediction has been implemented in the past, primarily for static content. Increasing web content and Internet traffic is making web prediction models very popular[1].

The objective of a prediction model is to identify the subsequent requests of a user, given the current request that a user has made. This way the server can pre-fetch it and cache these pages or it can pre-send this information to the client. The idea is to control the load on the server and thus reduce the access time. Careful implementation of this technique can reduce access time and latency, making optimal usage of the server’s computing power and the network bandwidth.

Markov model is a machine learning technique and is different from the approach that data mining does with web logs. Data mining approach identifies the classes of users using their attributes and predicting future actions without considering interactivity and immediate implications. There are other techniques like prediction by partial matching and information retrieval that may be used in conjunction with Markov modeling, to enhance performance and accuracy[2].

A web prediction model unlike other prediction models are particularly challenging because of the many states that it has to hold and the dynamic nature of the web in terms of user actions and continuously changing content. We therefore use the Markov probabilistic idea to design a prediction model. It is a stochastic counterpart to deterministic process in probability theory[3].

1.1. Web Mining

Web mining consists of a set operations defined on data residing on WWW data servers defines web mining as“…the discovery and analysis of useful information from the World Wide Web”. Such data can be the content presented to users of the web sites such as hyper text markup language (HTML) files, images, text, audio or video. Also the psychical structure of the websites or the server logs that keep track of user accesses to the resources mentioned above can be targets of web mining techniques[4]. Web mining as a sub category of data mining is fairly recent compared too their areas since the introduction of internet and its widespread usage itself is also
recent. However, the incentive to mine the data available on the internet is quite strong. Both the number of users around the world accessing online data and the volume of the data itself motivate the stakeholders of the web sites to consider analyzing the data and user behavior. Web mining is mainly categorized into two subsets namely web content mining and web usage mining[5]. While the content mining approaches focus on the content of single web pages, web usage mining uses server logs that detail the past accesses to the web site data made available to public. Usually the physical structure of the web site itself which is a graph representation of all web pages in the web site is used as a part of either method. However recent approaches that appoint more focus on the physical link structure of the web site have introduced web structure mining as a separate concept[6].

1.2. Web Usage Mining

The main topic of this paper is the web usage mining. Usage mining as the name implies focus on how the users of websites interact with web site, the web pages visited, the order of visit, timestamps of visits and durations of them. The main source of data for the web usage mining is the server logs which log each visit to each web page with possibly IP, referrer, time, browser and accessed page link. Although many areas and applications can be cited where usage mining is useful, it can be said the main idea behind web usage mining is to let users of a web site to use it with ease efficiently, predict and recommend parts of the web site to user based on their and previous users actions on the web site[7].

1.3. Markov Model

Markov Models have been widely used for predicting next Web-page from the users’ navigational behavior recorded in the Web-log. This usage-based technique can be combined with the structural properties of the Web-pages to achieve better prediction accuracy. This paper proposes one of the pre-fetching techniques relying both on Markov Model and Ranking which considers the structural properties of the Web. In this paper, prediction accuracy is realized as a linear function of transition probability of first order Markov Model and ranking of the Web-page. The chance of the predicted Web-page being the next Web-page would be higher if the prediction accuracy of the Web-prediction[8]. This research work proposes an improved hash mining association algorithm. To minimize the number of candidate sets while generating association rules by evaluating quantitative information associated with each item that occurs in a transaction, which was usually, discarded as traditional association rules focus just on qualitative correlations. The proposed approach reduces not only the number of item sets generated but also the overall execution time of the algorithm. Any valued attribute will be treated as quantitative and will be used to derive the quantitative association rules which usually increases the rules' information content. Transaction reduction is achieved by discarding the transactions that does not contain any frequent item set in subsequent scans which in turn reduces overall execution time. Dynamic item set counting is done by adding new candidate item sets only when all of their subsets are estimated to be frequent. The frequent item ranges are the basis for generating higher order item ranges using a Priory algorithm. During each iteration of the algorithm, use the frequent sets from the previous iteration to generate the candidate sets and check whether their support is above the threshold[9].

1.4. Clustering

Many clustering schemes have been proposed; that when applied along with Markov prediction techniques, achieve better accuracy. Proposes an unsupervised distance based partitioned clustering scheme. It is widely used in grouping web user sessions. It is also known as K-means clustering algorithm. Prediction techniques were applied using each cluster and using the whole data set. Results indicate that the clustering algorithms on the data set improve the accuracy of the prediction model. There are other clustering schemes like distance based hierarchal clustering and model based clustering, which are also known to improve predictive accuracy. In our implementation of the model, we have not used any advanced clustering scheme on the data set. However we grouped the URLs as accessed by the user classifying it using the IP address available in the log files. We then see if these URLs have been accessed within fifteen, thirty, forty-five minute or above interval. Our results revealed that the prediction was better for the fifteen and thirty minute scheme as compared to other intervals or having no session intervals at all. However we do believe that this conclusion may not be true for all web servers and largely depends on the content the web site hosts and the users visiting that website[11].
1.5. Association Mining

Association rule mining is a major pattern discovery technique. Association rule discovery on usage data results in finding group of items or pages that are commonly accessed and purchased together. The original goal of association rule mining is to solve the market basket problem. The application of association rule mining is far beyond market basket application & they have used for various domains including web mining. In web mining context, association rules help optimize the organization & structure of web site. Association rules are mainly defined by two matrices: support and confidence. The mining support requirement dictates the efficiency of association rule mining. Support corresponds to statistical significance while confidence is a measure of the rule strength[13].

2. Problem Definition

Caching popular objects close to the users provides an opportunity to combat this latency by allowing users to fetch data from a nearby cache rather than from a distant server. Web caching has been recognized as one of the effective schemes to alleviate the service bottleneck and reduce the network traffic, thereby minimize the user access latency, but it has the drawback that it stores the pages without any prior knowledge. Predictive caching becomes an attractive solution where in the forthcoming page likely to be requested soon are predicted based on user access logs information and prefetched , while the user is browsing the current display pages.

As web page prediction gained its importance, this thesis proposes a bracing approach for increasing web server performance by analyzing user behavior, in this prefetching and prediction is done by preprocessing the user access logs and integrating the three techniques i.e. Clustering Markov Model and association rules which achieves better web page access prediction accuracy.

In first part of our work we employed clustering technique. In this technique the preprocessed data is divided in to meaningful clusters then in the second part of our work, these clusters are used as training data while performing Markov Model techniques. We are also presenting the comparison study between these. As a result our work helps in better prediction, decreasing the access time of user and improving web performance.

3. Research Methodology

Our work improves the Web page access prediction accuracy by combining both Markov model and clustering techniques. It is based on dividing Web sessions into groups according to Web services and performing Markov model analysis using clusters of sessions instead of the whole data set. We are providing an approach with composition of two algorithms K-Means Clustering and the Hidden Markov Model.

The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. After the completion of K-Means Clustering Algorithm on our session data set, we will apply another algorithm i.e. Markov Model to find out the page which has highest occurrence among all the web pages visited by a web user under different sessions. These are some of the steps which follow in our complete process.

Step 1. Initialize the data set of visited web page on server.

Step 2. Define the number of clusters called K.

Step 3. Determine the centroid of the data set.

Step 4. Take each sample in sequence and compute its distance from the centroid of each of the clusters and group these K clusters according to the distance from centroid of the cluster.

Step 5. Check for some common data set in each group. If there is some such value, eliminate them.

Step 6. Repeat the process from Step 2 till all items are not collectively categorized.

Step 7. Now use these K clusters in the basic training set for the prediction algorithm.

Step 8. Now perform first level Markov Model to determine the occurrence of each page visited by user.

Step 9. Remove the value having less value than the average.

Step 10. Now perform level 2 Markov Model and find the appearance of group 2 pages like AB, AC.

Step 11. Again eliminate the value having value less than the average.
Step 12. Find the after and before visited page from the group.

Step 13. Perform the strength calculation between the associated values with pair groups.

Step 14. The value with highest strength will be represented as the highest calculated/strength value.

4. Results and Discussion

4.1 Result

Web users are facing the problems of information overload and drowning due to the significant and rapid growth in the amount of information and the number of users. The size of publically indexed World Wide Web has probably surpassed 24.39 billion pages in June 2010. As a result, delays in access to Web based Information continues to be a serious problem even with higher network bandwidth, due to overhead web latency has increased due to which web performance has decreased. User perceived latency from several sources such as bandwidth, speed, overhead, accessing the web page etc. In accordance of “Eight Second Rule”, it can be observed that web latency affects the user work and lots of efforts are taken to minimize the latency perceived by the user. Caching of web documents has been developed to reduce the latency but it has the drawback that it stores the pages without any prior knowledge i.e. the hit ratio is less. While predictive caching can optimize the www in many respects, this motivates our work.
4.2 General Discussion

The main objective of the paper is to help to achieve better prediction accuracy for web page access. Recommending a next page the web user will access is very important for various web applications web page prediction is addressed by many literature publications. The main topic implement for this Purpose is through using web page using Prediction model. In third dissertation we examine the three most important techniques as Markov Model, Association Rule Mining, Clustering and C-means. Through the Pattern discovery Models integrations, we exhausted their varied Positive impact on Web Page Prediction accuracy. By keeping the Modules limitations to a minimum on their advantages and by integrating the different modules according to different constraints, we will able to achieve more accurate prediction result.

5. Conclusion and Future works

5.1 Conclusion with Result

With the growth of Web based application, specifically electronic commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has lead to a number of open issues in Web Usage Mining area. In many practical applications, due to the introduction of stricter laws, privacy respect represents big challenge. In this survey paper, we briefly explored various applications of web usage mining suggested by authors. We also analyzed some problems and challenges of Web usage mining. Anyway we believe that the most interesting research area deals with the integration of semantics within Web site design so to improve the results of Web Usage Mining applications. Efforts in this direction are likely to be the most fruitful in the creation of much more effective Web Usage Mining and personalization systems that are consistent with emergence and proliferation of Semantic Web. Web prediction can go a long way in improving the experience of the users on the web. Web and web technologies are still evolving and the opportunities to incorporate such techniques are wide open.

5.2 Scope for future work

The above model can be modified to make further improvement in predictive accuracy. An important observation that I made while going though the literature on various web prediction models is that, not many researchers have made an effort to consolidate the different mechanism to improve accuracy of the model. These mechanisms may only give a small increase when applied singly. Though, when applied together, they may create a substantial difference. This involves using a model like an All-Kth-order Markov model in conjunction with clustering algorithms and pruning techniques. The above-mentioned reasons make it compelling for us to consider this proposition.

This model has a few drawbacks too. One of the drawbacks is that it does not track the probability of the next item that has never been seen. Using a variant of prediction by partial matching will help take care of this situation and should be considered in the work ahead. Another improvement that can be done is to perform statistical evaluation to establish pruning thresholds on the fly. These results are based on logs from a single web server. It is important to validate these observations over log files from different sources.

6. References


