ISSN:2231-0711

Shikha Yadav et al | IJCSET(www.ijcset.net) | June 2016 | Vol 6, Issue 6, 250-253

An Approach to Improve Web Performance via Markov Prediction Model

Shikha Yadav[1], Urvashi[2]

1. *Affliated to MDU Rohtak: M.Tech Computer Science and Engineering*
World College of Technology and Management
Gurgaon, Haryana, India

2. *Affliated to MDU Rohtak: Assistant Prof. Computer Science and Engineering*
World College of Technology and Management
Gurgaon, Haryana, India

Abstract- Web prefetching is a prominent solution to reduce the network resources consumed by Web services as well as the access latencies perceived by Web users. Unlike Web caching, which exploits the temporal locality, Web prefetching utilizes the spatial locality of Web objects. Specifically, Web prefetching fetches objects that are likely to be accessed in the near future and stores them in advance. In this context, a sophisticated combination of these two techniques may cause significant improvements on the performance of the Web infrastructure. Considering that there have been several caching policies proposed in the past, the challenge is to extend them by using data mining techniques. In this paper, we present a clustering-based prefetching scheme where a graph-based clustering algorithm identifies clusters of ''correlated'' Web pages based on the users' access patterns. This scheme can be integrated easily into a Web proxy server, improving its performance. Through a simulation environment, using a real data set, we show that the proposed integrated framework is robust and effective in improving the performance of the Web caching environment. Reducing the web latency is one of the primary concerns of Internet research. Web caching and web prefetching are two effective techniques to latency reduction. A key method for intelligent prefetching is to rank potential web documents based on prediction models that are trained on the past web server and proxy server log data, and to pre-fetch the highly ranked objects. The PROPOSED APPROACH is three level approach in which we have combined markov model along with association mining and clustered approach. For this method to work well, the prediction model must be updated constantly, and different queries must be answered efficiently.

Keywords: Markov, association mining, Clustering, Web Mining.

1. INTRODUCTION

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

1.1 Markov model

The Markov Model collected in this implementation clearly indicates that a third and higher order model has high success rate in terms of positive future predictions. One can therefore build variations of this model and use the one with the highest applicability and success rate or use a combination.

The different order model is directly associated with n-gram, used by the speech and language processing community. We may borrow this idea and consider an n-gram as a sequence of n consecutive request. To make a prediction, one should match the prefix of length n-1 of an n-gram and use the Markov model to predict the n^{th} request. However, the important thing here is, given a prefix of length n-1, there are numerous possibilities of the n^{th} request. How do we identify the n^{th} request appropriately?

We use the Markov model's idea of states. Each node and therefore each request represent a state transition. The transition from one state to another has some probability associated with it. Given a sequence of n-1 states, we pick the n^{th} state with the highest probability. How we calculate this probability is explained in the next section.

1.2 Markov based web prediction model

The Markov Model collected in this implementation clearly indicates that a third and higher order model has high success rate in terms of positive future predictions. One can therefore build variations of this model and use the one with the highest applicability and success rate or use a combination.

The different order model is directly associated with n-gram, used by the speech and language processing community. We may borrow this idea and consider an n-gram as a sequence of n consecutive request. To make a

prediction, one should match the prefix of length n-1 of an n-gram and use the Markov model to predict the n^{th} request. However, the important thing here is, given a prefix of length n-1, there are several possibilities of the n^{th} request. How do we identify the n^{th} request appropriately?

We use the Markov model's idea of states. Each node and therefore each request represent a state transition. The transition from one state to another has some probability related with it. Given a sequence of n-1 states, we pick the n^{th} state with the highest probability. How we calculate this probability is explained in the next section.

Note that the transition with the highest probability may not be the correct request always and hence we use the idea of *top-n* predictions. Here we not only consider the n^{th} request with the highest probability but more than one request with high probabilities. We could establish a minimum threshold to achieve higher accuracy. The change between different states is known as transition and the probability with which this occurs is known as transition probability.

So when we have a request that gets a web page or a resource, it will be considered as the current state of the system. Given this prediction model (associated transition probabilities) and a set of request sequence, our goal is to predict the future state of the request.

1.3 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 minutes 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.

Clustering Web sites can be achieved through page clustering or user clustering. Web page clustering is performed by grouping pages having similar content. Page clustering can be simple if the Web site is structured hierarchically. In this case, clustering is obtained by choosing a higher level of the tree structure of the Web site. On the other hand, clustering user sessions involves selecting an appropriate data abstraction for a user session and defining the similarity between two sessions. This process can get complex due to the number of features that exist in each session. These features are service request, navigation pattern and resource usage.

1.4 Association Rules Mining

Association rules mining is a major pattern discovery technique. Association rules discovery on usage data results in finding group of items or pages that are commonly accessed and purchased together. **The original goal of association rules mining is to solve the market basket problem. the application of association rules mining is far beyond market bucket 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.

There are four types of sequential association rules presented by:

1. **Subsequence rule**: They represent the sequential association rules where the items are listed in order.
2. **Latest subsequence rule**: They take into consideration the order of the items and most recent items in the set.
3. **Substring rule**: They take into consideration the order and the adjacency of the items.
4. **Latest substring rule**: They take into consideration the order of the items, the most recent items in the set as well as the adjacency of the items.

Support are defined as the discovery of frequent item set i.e. item sets which fulfill a minimum support threshold) and confidence is defined as the discovery of association rules from these frequent item set.

Association rules mining allows businesses to infer useful information on customer purchase patterns, shelving criterion in retail chains, stock trends etc. The basket data essentially consists of a large number of individual records called transactions and each transaction is a list of items that participated in the transaction. **The goal of association rules mining is to discover rules it is likely to contain a specific item.** A formal definition of association rule mining is presented sampling techniques for association rule mining in massive databases. Sampling has been used quite effectively for solving numerous problems in databases and data mining.

Association rules mining, the task of finding correlations between items in a dataset, Initial research was largely motivated by the analysis of market basket data, the results of which allowed companies to more fully understand purchasing behavior and, as a result, improved target market audiences.

2. PROPOSED WORK

2.1 Integration Markov Model with Association Rules

Integration association rule with markov model in order to improve the prediction accuracy. But both have been used individual for prediction purpose, but each of them has its own limitation when it occurs to be web page prediction

accuracy & state space complexity. The main advantage of markov model is that they can generate navigation paths that could be used for automatically for prediction, without any extra processing and thus they very useful for web personalization.

The original goal for association rules mining are used for predict the next page to be accessed by the web users. The more frequently the pages are accessed the higher the probability of the user accessing the next pagr.it involves dealing too many rules & it is easy to find the suitable subset of rule to make accurate and reliable prediction.The integration model profit from the decrease the state space complexity of the lower markov model by using association mining in case of ambiguity.the integration model also provides the complexity of the association rules since the rules are generated only in special cases.

In brief, the new integration model results in an increase the accuracy and a decrease in state & rule complexity.

2.2 Integration Markov with clustering Model (IMCM)

The web page prediction anticipated the next page to be accessed by the user or the link the web user. Clustering is able to personalize user according to their browsing experience. this is very significant since the markov model for a subgroup, that is assumed to be more homogenous than the whole data set, has an higher quality than the markov model of the whole data set. Markov model archives the higher accuracy but the association with the higher state space complexity with the clustering. Although the clustering technique have been used for the personalization purpose by discovering web site structure, and extracting useful pattern. Proper clustering groups user's sessions with similar browsing and facilities classifications.

The markov model and clustering based on the state space complexity association with the higher –order. Web sessions are categories into the numbers of categories: K-means clustering is based on the web sessions identified and is carried out according to some distance matrices.

2.3 The integration Markov Model with Clustering and Association Rules (IMCA)

Combine the Markov Model, association Rule and clustering to improve the next page prediction accuracy and combining Markov Model with clustering technique has been proved to improve the Prediction accuracy to a greater extend. It helps to reduce the I/O overhead association with large data base by making only one pass over the database when learning association rules. They combine the rules with Markov Model is novel to our knowledge and only few of past research's combined all three modules.It improves the performance of Markov Model, sequential association rules, and clustering by combining all these model together. It improver the prediction accuracy as opposed to other combines that prove to improve the prediction coverage and complexity.

Therefore, better clustering means the better Markov Model Prediction accuracy because the Markov Model Prediction will be based on more meaningfully grouped data.

It also improves the state space complexity because Markov Model Predict will be carried out on one particular cluster as opposed to the whole dataset. The Integration Model then computer Markov Model Predict on the resulting clusters. Association rules only examined in the case where the Prediction results are based on the state.

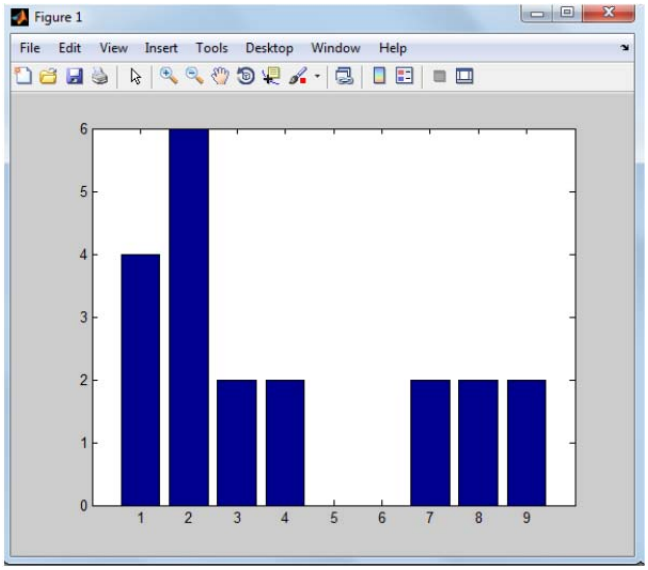
3. PROPOSED ALGORITHM

These are some of the steps which is follows in our complete process:-

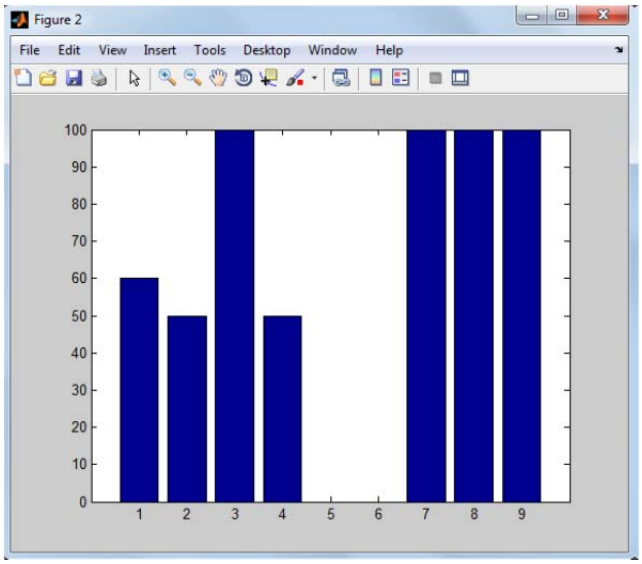
Step 1. Initialize the data set of visited web page on server.
Step 2. Define the number of clusters called K.
Step 3 . Determine the centroid of the data set.
Step 4. Take each sample in sequence and compute its distance from the centroid of each of the clusters and group these K clusters according 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 performs 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. CONCLUSION WITH RESULT

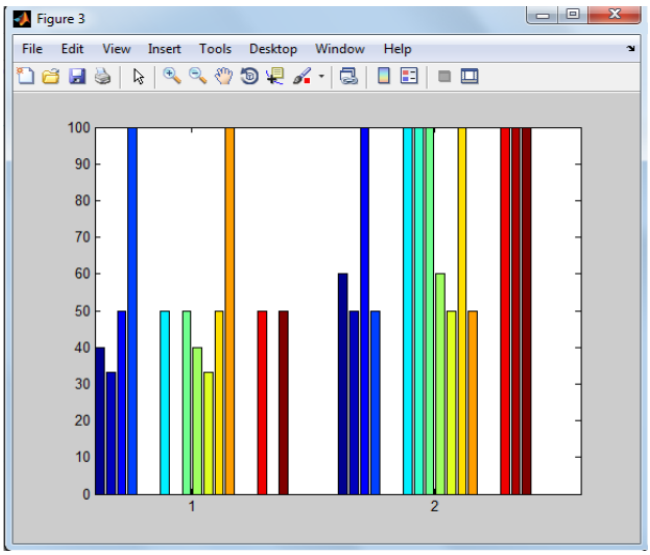
With the growth of Web based application, specifically electronic commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has led to a number of open issues in Web Usage Mining area. In many practical applications, 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.



Markov Model with web cache



Markov Model with Clustering



Markov Model with web Pages

REFERENCES

[1] **Andrea Bacic,"** Intelligent Interaction: A Case Study of Web Page Prediction" Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, June 22-25, 2009, Cavtat, Croatia.

[2] **Anna Gutowska**," A Comparison of Methods for Classification and Prediction of Web Access Patterns", COMP540 - Final Report - Spring 2010.

[3] **Arwen Twinkle Lettkeman,"** Predicting Task-Specific Webpages for Revisiting", Compilation copyright © 2006, American Association for Artificial Intelligence All rights reserved.

[4] **Banu Deniz GUNEL** performed a work,**"** Investigating the Effect of Duration, Page Size and Frequency on Next Page Recommendation with Page Rank Algorithm"2010.

[5] **Ben Taskar**," Link Prediction in Relational Data",2010.

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

[7] **Brigitte Kaltenbacher**" From Prediction to Emergence Usability Criteria in Flux",2009.

[8] **Chennupati.R.Prasanna**,**"** An Approach for Restructuring of Web Pages" IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.2, February 2010.

[9] **Chun-Jung Lin**," Using Hidden Markov Model to Predict the Surfing User's Intention of Cyber Purchase on the Web",2005.

[10] **David Ya**,**"** Web-Based Forecasting of Potential Evapotranspiration for Improved Water Resource Management in California**",** 2nd Joint Federal Interagency Conference, Las Vegas, NV, June 27 - July 1, 2010.