ISSN:2231-0711

Shikha Yadav et al | IJCSET(www.ijcset.net) | June 2016 | Vol 6, Issue 6, 254-259

An Approach to Improve Web Performance via Markov and C-Means Algorithm

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- 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 will visit will be copied to the user cache. Web prefetching is an attractive 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. 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, C-means, 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 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 presented the content to users of the web sites such as hypertext 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. Web mining as a sub category of data mining is fairly recent compared to 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**. 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.

1.2 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.

1.3 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.

1.4 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 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.

1.5 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. LITERATURE REVIEW

2.1 Review

Lot of work is already done in the area of web page prediction and web caching. In this section, the work done by the earlier researchers in this area is presented and discussed. In this thesis, we defined the optimization process to reduce the web information access and to reduce the error. we optimized the search mechanism along with encoded search. we improved the quality of the search algorithm with the reduction of integration error. Another work on the improvement of web page access was defined by us. we presented the prediction analysis approach to improve the web page caching. The work proposed considered a realistic prefetching architecture using real and representative traces. We implemented the work in real web environment and the obtained results shows the significant improvement over the existing approaches. 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. A performance evaluation is presented using real Web logs. In this thesis, preliminary work in the area of Web page prediction is presented. The designed and implemented prototype offers personalized interaction by predicting the user's behavior from previous Web browsing history. Those predictions are afterwards used to simplify the user's future interactions. Rather simple and feasible prototype enhancements are offered and discussed. Its simplicity and effectiveness makes it potentially useful for widespread application. In this thesis, we presented an improvement over the caching scheme so that the page access consistency will be improved.

2.2 Literature Reference

[1] Arwen Twinkle Lettkeman,[2006] performed a work**, "Predicting Task-Specific Webpages for Revisiting".** Author address how Author can better support web users who want to return to information on a webpage that they have previously visited by building more useful history lists. The paper reports on a combination technique that semi-automatically segments the webpage browsing history list into tasks, applies heuristics to remove webpages that carry no intrinsic revisit value, and uses a learning model, sensitive to individual users and tasks, that predicts which webpages are likely to be revisited again.

Author present results from an empirical evaluation that report the likely revisit need of users and that show that adequate overall prediction accuracy can be achieved.

[2] Banu Deniz GUNEL,[2010] performed a work, **"Investigating the Effect of Duration, Page Size and Frequency on Next Page Recommendation with Page Rank Algorithm".** In this paper, Author extend the use of page rank algorithm for next page prediction with several navigational attributes, which are size of the page, duration time of the page and duration of transition (two page visits sequentially), frequency of page and transition. In Presented model, Author define popularity of transitions and pages by using duration information and use it in relation to page size and visit frequency factors. By using the popularity value of pages Author bias conventional Page Rank algorithm and model a next page prediction system that produces page predictions under given top-n value. Actually Author devise Duration Based Rank (DPR), which focuses on page duration with size proportion and Popularity Based Page Rank (PPR) ranking model.

[3] Ben Taskar,[2010] performed a work**, "Link Prediction in Relational Data".** This paper focuses on predicting the existence and the type of links between entities in such domains. Author apply the *relational Markov network* framework of Taskar *et al.* to define a joint probabilistic model over the entire link graph—entity attributes and links. The application of the RMN algorithm to this task requires the definition of probabilistic patterns over subgraph structures. Author apply this method to two new relational datasets, one involving university webpages, and the other a social network. Author show that the collective classification approach of RMNs, and the introduction of subgraph patterns over link labels, provide significant improvements in accuracy over flat classification, which attempts to predict each link in isolation.

[4] Brigitte Kaltenbacher,[2009] performed a work**, "From Prediction to Emergence Usability Criteria in Flux".** This paper aims to discuss the position of the traditional usability model in the context of current technical interaction, and in particular in internet interaction. The traditional usability model was developed in the context of software development.

Yet it is relevant to IA for two reasons: firstly, on the internet information design and retrieval (IR) benefits from its application just as much as software development did, due its vast user base. Secondly, large parts of the internet are application or software driven by now. At the same time, the interplay of information and applications on the internet has produced new ways of interaction, and new demands towards the quality of interaction. Consequently,

the traditional usability model needs to be expanded beyond an entirely functional focus, to accommodate the richer notion of the user experience.

[5] Chennupati.R.Prasanna,[2010] performed a work, **"An Approach for Restructuring of Web Pages".** There are generally three tasks in Web Usage Mining: Preprocessing, Pattern analysis and Knowledge discovery .Preprocessing cleans log file of server by removing log entries such as error or failure and repeated request for the same URL from the same host etc. The main task of Pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. The information collected from the log file can help to discover the knowledge. This knowledge collected can be used to take decision on various factors like Class1, Class 2, users and Eminent, Average and Delicate web pages based on hit counts of the web page in the website.

[6] Chun-Jung Lin,[2005] performed a work, **"Using Hidden Markov Model to Predict the Surfing User's Intention of Cyber Purchase on the Web.**This paper is the first one applying the Hidden Markov Model, the stochastic tool used in information extraction, in predicting the behavior of the users on the web. Author collect the log of web servers, clean the data and patch the paths that the users pass by. Based on the HMM, Author construct a specific model for the web browsing that can predict whether the users have the intention to purchase in real time. The related measures, such as speeding up the operation, kindly guide and other comfortable operations, can take effects when a user is in a purchasing mode.

[7] David Ya,[2010] performed a work, **"Web-based Forecasting of potential Evapotranspiration for Improved Water Resource Management In California".** What to plant, how much and when to deliver irrigation water among the many decisions agricultural, water, and irrigation managers must make throughout California's Central Valley. If these planners and managers could have forecasts of water demand in the near (8-day) and long (90-day) term, the strong possibility exists for improved resource management decisions. To this end, a web-based tool that generates both 8-day forecasts and 90-outlooks of potential evapotranspiration (ETo) for use in physically based water management models has been developed. The forecasts and outlooks are made for California Irrigation Management Information System (CIMIS) sites throughout California, with the 8-day forecasts based on NCEP High Resolution Global Forecast System (1 degree GFS) model output with site specific bias-correction.

[8] Devanshu Dhyani,[2002] performed a work**, "Modelling and Predicting Web Page Accesses Using Burrell's Model".** In this paper, Author consider the application of patterns in browsing behavior of users for predicting access to Web documents. Author proposed two models for addressing Presented specification of the access prediction problem. The first lays out a preliminary statistical approach using observed distributions of interaccess times of individual documents in the collection.

[9] Elli Voudigari,[2010-2011] performed a work, **"A Framework for Web Page Rank Prediction".** Author propose a framework for predicting the ranking position of a Web page based on previous rankings. Assuming a set of successive top-k rankings, Author learn predictors based on different methodologies. The prediction quality is quantified as the similarity between the predicted and the actual rankings. Extensive experiments were performed on real world large scale datasets for global and query-based top-k rankings, using a variety of existing similarity measures for comparing top-k ranked lists, including a novel and more strict measure introduced in this paper.

[10] Faten Khalil,[2007] performed a work**, "Integrating Recommendation Models for Improved Web Page Prediction Accuracy".** Different Web usage mining frameworks have been implemented for this purpose specifically Association rules, clustering, and Markov model. Each of these frameworks has its own strengths and weaknesses and it has been proved that using each of these frameworks individually does not provide a suitable solution that answers today's Web page prediction needs. This paper endeavors to pro- vide an improved Web page prediction accuracy by using a novel approach that involves integrating clustering, association rules and Markov models according to some constraints.

[11] Osama Hamed,[2007] performed a work**, " Performance Prediction of Web Based Application Architectures Case Study: .NET vs. Java EE".** Efficient web application is a challenge that Author need to achieve when architecting web applications. This research follows a performance testing approach that aims to utilize load testing tools to give ideas about performance issues early in the development life cycle for applications implemented using Java Enterprise Edition (Java EE) or .NETplatform.

[12] Paramveer S. Dhillon,[2011] performed a work, **"Semi-supervised Multi-task Learning of Structured Prediction Models for Web Information Extraction".** Extracting information from web pages is an important problem; it has several applications such as providing improved search results and construction of databases to serve user queries. In this paper Author propose a novel structured prediction method to address two important aspects of the extraction problem: (1) labeled data is available only for a small number of sites and (2) a machine learned global model does not generalize adequately well across many websites. For this purpose, Author propose a weight space based graph regularization method. This method has several advantages. First, it can use unlabeled data to address the limited labeled data problem and falls in the class of graph regularization based semi-supervised learning approaches. Second, to address the generalization inadequacy of a global model, this method builds a local model for each website. Author demonstrate the efficacy of Presented method on several real-life data; experimental results show that significant performance improvement can be obtained by combining semi-supervised and multi-task learning in a single framework.

3. PROPOSED APPROACH

3.1 C-Means clustering

Partitional clustering is an important part of cluster analysis. Based on various theories, numerous clustering algorithms have been developed, and new clustering algorithms continue to appear in the literature. It is known that Occam's razor plays a pivotal role in data-based models, and partitioned clustering is categorized as a data-based model.

The three main contributions of this paper can be summarized as follows:

1) According to a novel definition of the mean, a unifying generative framework for partitioned clustering algorithms, called a general c-means clustering model (GCM), is presented and studied.

2) Based on the local optimality test of the GCM, the connection between Occam's razor and partitioned clustering is established for the first time. As its application, a comprehensive review of the existing objective function-based clustering algorithms is presented based on GCM

3) Under a common assumption about partitioned clustering, a theoretical guide for devising and implementing clustering algorithm is discovered. These conclusions are verified by numerical experimental results..

3.1.1 Distance measure

An important step in most clustering is to select a distance measure, which will determine how the *similarity* of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another. For example, in a 2-dimensional space, the distance between the point ($x = 1$, $y = 0$) and the origin ($x = 0$, $y = 0$) is always 1 according to the usual norms, but the distance between the point ($x = 1$, $y = 1$) and the origin can be 2, $\sqrt{2}$or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

3.2 C-Means Algorithm:

It minimize the index of quality defines as sum of squared distances for all points included in the cluster space to the center of the cluster

Algorithm:

1. Fix the number of cluster.
2. Randomly assign all training input vector to a cluster .this creates partition.
3. Calculate the cluster center as the mean of each vector component of all vectors assigned to that cluster. Repeat for all cluster
4. Compute all Euclidean distances between each cluster center and each input vector.
5. Update partitioned by assigning each input vector to its nearest cluster minimum Euclidean distance.
6. Stop if the center do not move any more otherwise loop to step, where is the calculation of a cluster center.

3.2.1 Quality of Result:

C-means depends on:

1. Amount of chosen cluster center.

2. Sequence of pattern survey.

3. Geometric properties of data.

3.3 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.

3.4 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.

4. 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 through 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-K^{th}-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 can be further improved as 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 to 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

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.

[11] **Faten Khalil,"** Integrating Recommendation Models for Improved Web Page Prediction Accuracy"**,**2007.

[12] **Osama Hamed,"** Performance Prediction of Web Based Application Architectures Case Study: .NET vs. Java EE",2007.