A Combined Model of Supervised and Unsupervised Learning for Reliable Web Page Prediction

Mayank Khattar
Department of Computer Science and Engineering
Manav Institute Of technology And Management,
Hissar, Haryana, India

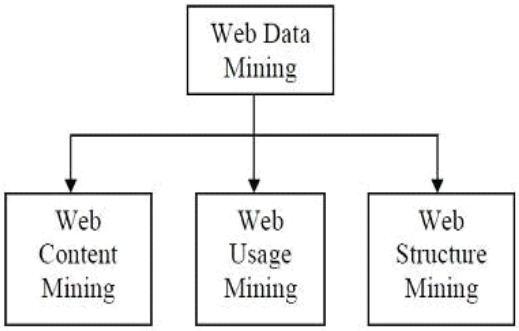
Vikas Malik
Department of Computer Science and Engineering
Manav Institute Of technology And Management,
Hissar, Haryana, India

Abstract - Web usage mining is one of the important aspects of web mining that not gives the benefit the service providers to get information about user interest in their contents or the web site, but also it helps a user to identify the best services provider. The web usage mining is about to analyze the web links respective to user interest and the frequency of use. In this work we have defined a work on estimation of prefetching contents based on usage mining. Prefetching is the term used to retrieve a web page before the user demand. This actually a prediction system, in which user web visit history is analyzed to identify the most required and visited contents over the web. Once these contents are obtained the next is to perform the clustering to identify the most visited and least visited web pages. Now the cluster showing the higher frequency interest is been processed under the markov model to predict the next page. At the final stage the neural network is implementing to obtain the more accurate prediction of the web page. In this paper, the proposed architecture of the defined work is presented. The results obtained from the system shows the effective prediction of web page.

Keywords: Similarity, Predictive, Web mining, Fetching, Optimizing.

I. INTRODUCTION

To speed up the process of web access, the concept of web caching and the pre fetching is defined. Web Caching helps to speed up the access of frequently accessed pages or user interest web pages. Web caching is the concept to improve the performance of Web servers. Caching actually stores the Webpages in a central place and only update them based on the user regular visit. They make the pages appear to load faster, while sacrificing timeliness. Web Caching is about to maintain the server page information or the contents on client side so that as the user perform request on server side to retrieve the webpage, client side avail the information with the request itself. The objective of web caching is to improve the boost up the page access over the web[1][2][5][6].



Here figure 1 is showing different categories of web mining. Web caching is the approach that enhances the web browsing in an effective way. When a user visits a site, say www.yahoo.com, and the information is updated to the client cache. Along with this it perform the history analysis to observe the visit of the particular web page. Once the visit count is obtained, the association of this particular web page is compared with all other web pages [7][8][9].

The web page that provides the highest association is then perfected and stored in the cache before the user demand to that particular web page. Web prefetching actually defines a prediction system in which a page is loaded in the cache before it is requested by the user. The prediction of web page basically provides the effective utilization of the available resources and also reduces the consumption of these resources. Web prefetching provides the access of the web page from cache table or from client side so that each time downloading of the web page will be avoided. The accuracy of these kind of system is based on the estimation of the requirement of the web user in near future, once the estimation is predicted, web page is loaded in the cache as the next visiting page. These kind of systems are based on the web architecture or web infrastructure so that maximum benefit will be derived from the caching policies. In this paper, we are presenting a hybrid page prediction architecture that will collect the three main approaches of web page access called clustering approach, Markova model and the neural network. Here the clustering is basically used to optimize the results when we have a large dataset in the form of session. The clustering will divide the available history in terms of sessions and provide the effective access of web pages. In the second stage, markov model is defined to predict the web page on a web session analysis. The markov model is basically an intelligent supervised learning approach that can be implemented on a smaller dataset. To deal with large dataset, an enhanced neural based approach is suggested in this work. The neural will use the association mining as the final stage to identify the page that is needed to put in cache memory [10][11]

II. WEB USAGE MINING

Web usage mining basically defines the analysis over the usages of web pages. The web usage in simple term can be defined as a dataset of visited web pages by the user. Web usage mining is effective in many ways. One aspect of web usage mining is to track the web users so that relative interest can be identified and based on which some notifications can be sent to the users. These kinds of systems are used by the promotional sites that capture the user interest and based on the analysis offer them the products. Another user of web usage mining is to save the system from some attackers. If the usage history of a user is available, it is easy to identify the intruders over the web. The risk vector associated with a particular site can also be identified easily. Another use of web usage mining is to perform the web page prediction. The challenges associated with web usage mining are the preprocessing and the filtration stage. Preprocessing is about to analyze the web page history as it is generally having a huge collection of data. The preprocessing stage remove the non-required pages and categorize them either respective to web servers or web sessions. Once the preprocessing is done, the next stage is to filter the dataset and remove the unwanted pages [12][13].

III. PREDICTION MODEL

The prediction is about to analyze the subsequent web requests performed by thousands of users for millions of pages. The objective of this model is identify the current web page requirement of the user and identify the next required pages. This estimation is done statistically as well as intelligently. These are number of supervised and unsupervised learning approaches to perform the prediction of web pages. Once the page is predicted, it is pre-fetched by the model. Prefetching is about to avail a web page itself, when a user is busy with some work on his current page. This page is available to the client side without any request performed by the user. This approach is effective to optimize the access time and to reduce the wait time to access the web page. In systems, with low bandwidth such kinds of models are effective to achieve the maximum utilization of web pages [14][15][16].

IV. EXISTING WORK

In year 2005, Seung Yeol Yoo defined a web page prefetching scheme based on user interest. he work has defined the retrieval of web pages based on the keyword analysis. Author defined Author defined the interest in the access of web pages under the segmentation and based on the relevance under the feedback system for web information retrieval scheme. Author defined the work based on the structural analysis and discovers the information based on statistical analysis [1]. Another work on the concept of prefetching was presented by Cairong Yan in same year. Author defined work to optimize the web page access process and to reduce the wait time. Author performs the analysis on prefetching process and defines the significant approach for web page access. Author proposed a markov model based prefetching model under the probabilistic analysis [2]. Another algorithmic work on web page access was proposed by Junchang Ma in year 2006. The defined work was proposed on a large dataset of web pages. The objective of author was to optimize the usage of network bandwidth. Author defined the work on a shared network system and provided the fragment based caching over the web[3].

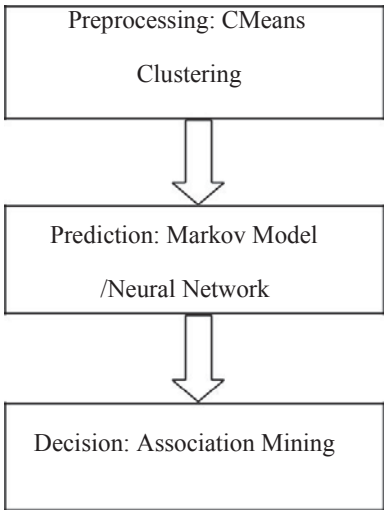
In year 2007, Ruma Dutta defined a probabilistic algorithmic approach based on PPM algorithm and markov tree model to perform the prefetching of web page. The markov tree based approach required large amount of memory and used the concept of automata system to generate all the possible combinations. Author also defined the page replacement approach to provide the effective utilization of resources [4]. In this paper, a page ranking based approach is defined for the prediction of web pages. Author defined a weighted analysis approach for page ranking. Author defined the analysis on the page visit frequency to obtain the accurate factorization of web pages [5]. In year 2008, Payal Gulati defined a Zipf law based approach for web page prediction. The work is probabilistic and based on it prefetching is done. The work was implemented to reduce the access latency [6].

Another work in same area to optimize the resources was done in year 2008 by Yong Zhen Guo. Author defined error identification and correcting approach to optimize the page access and to provide the effective access of pages [7]. Another work on web page prediction was proposed in year 2009 by Andrea Back. Author defined a behavior analysis approach for web browsing history. The work includes the future interaction analysis approach so that effective identification of next visiting page can be identified. Author defined graphical interface to achieve the effective prefetching [8]. In same year B.De La Ossa presented a study to describe the benefits of different prefetching technique. Author performed the theoretical and algorithmic analysis over the web. Web page prediction architecture was defined by the user based on the analysis and effective utilization of resources and the effective access of web pages is done [9].

V. PROPOSED WORK

In this paper, a hybrid model is presented with the integration of CMeans Clustering, Markov Model and the neural network. The work is presented achieve the different significance based on the size of dataset. For the smaller dataset, the markov model is here proposed to predict the web page and for the larger dataset, neural network is proposed in this work. The complete work is divided in three main stages shown in figure 2.

Figure 2 : Major Stages of Proposed Work



Initially all the pages are representing in the form of a Pool After this some filtration is performed to separate all the pages in the form of session. Session pages representations the categorization of pages. Once the categorization performed, the next step is to perform the initial pruning. The pruning is used to remove all the pages that are rarely being accessed by the user. After this the C-Means clustering is performed on these session pages. The clustering will divide all the pages in separate clustered. We can see in the graphs that the pages that are used more frequently are placed in one cluster and that are less frequently being used are placed in other cluster. The clustering is here performed to categorize the web pages based on web sessions. The pruning is here been performed between the sessions. It means, the sessions that are more likely based on the user page visit are considered and rest sessions are pruned from the dataset. Here CMeans clustering is been defined to perform the clustering of web pages. The basic Algorithm of CMeans Clustering is defined here

VI. CMEANS CLUSTERING

1. Initially decide the total number of clusters.
2. Analyze the available page set and divide them to different session based on the visit time history. These sessions are represented as separate partitions.
3. Build a unique page list used in any of the session.
4. Obtain the frequency of the each webpage.
5. Derive the center of the session page frequency.
6. Calculate the Euclidean distance for each based from the center.
7. Decide the area vector for each cluster.
8. Identify the pages that come within the area and present it as the cluster page.
9. Repeat the process from step 6 for all pages.

Once the clustering performed one more time the pruning is implemented. This time it will remove all the pages that are being used very less frequently. Now we can clearly see in the graph the set of pages that are frequently used by the user and these all pages are placed in different clusters.

VII. MARKOV MODEL

Markov model is an intelligent prediction based learning approach. This approach is based on the statistical analysis as well as more effective than other data mining operations respective to the future aspect analysis quality. This model performs the analysis on web logs and provides the prediction based on information retrieval and partial matching process. In this model, the decision is taken based on the frequency analysis and the pruning is performed at each stage. Complete work of markov model is divided in number of levels. With each level, the combinational analysis is performed. At very first level, the single page analysis is performed and based on the mean value analysis, the pruning is performed. In next level, the page pair frequency analysis is performed. In same way with each level, a new combination and frequency analysis is performed. The basic process of markov model is shown here under

1. Perform the frequency analysis on distinct web pages of a specific cluster.
2. Identify the threshold frequency level based on pruning rule such as mean value.
3. Remove the value having less value than the mean value obtained.
4. Now perform level 2 Markov Model and find the appearance of group 2 pages like AB, AC.
5. Again eliminate the value having value less than the average.
6. Find the after and before visited page from the group.
7. Perform the strength calculation between the associated values with pair groups.

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

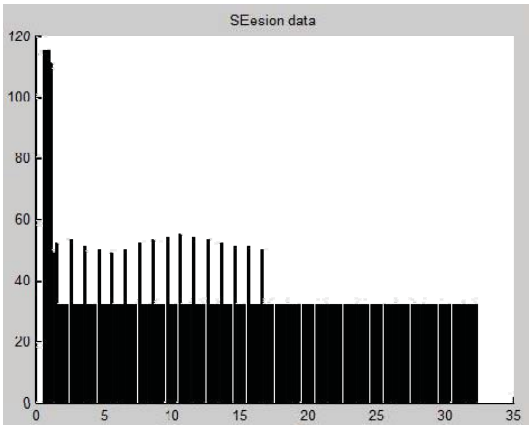
VIII. NEURAL NETWORK

Neural network is a unsupervised classification and computational analysis approach to identify the similarity between the data items and to identify the relation between them. The architecture defined here is based on a layered approach. In the simplest form, it requires three stage architecture called input layer, hidden layer and output layer. The input is here taken in the form of session pages with the frequency definition. In the hidden layer, the weights are assigned to these pages based on the level analysis. Based on these weights, the true positive level is defined for the pages. Once the rule is defined, the dataset is trained on neural network and the output is obtained. On this output, the association mining is performed to identify the next visiting page. The neural will be implemented when we have a large set of page history.

IX. RESULTS

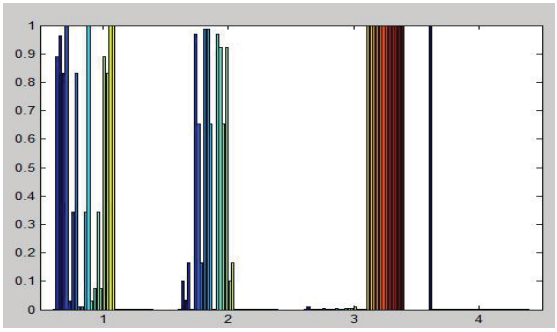
The presented work is implemented in Matlab environment on a sampled dataset. In this work, a sample of web pages is assumed with some marked naming such as page A,B, C etc. The dataset is taken in the form of matrices as well as in the form of CSV file. Once the input is taken, the algorithmic approach is implemented by using core programming and the neural network tool The results obtained from the system are listed as under.

Figure 3 : Session Data



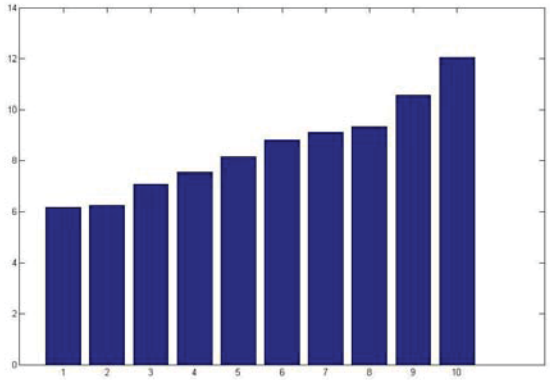
As we can see, in figure 3, the session pages are defined based on frequency analysis. Here x axis represnts the pages and y axis represents the frequency of the pages.

Figure 4: Clustered Results



Here figure 4 is showing the categorization of these web pages in 4 clusters. All the pages in a clsuter are forming a group shown in the figure. Here x axis represents the clusters and the y axis represnts the probabilistic analysis of web pages.

Figure 5: Prediction Results



Here figure 5 is showing the predictive results driven from the model. As we can see, the model has shown the incremental change in the web usage respective to the web pages.

X. CONCLUSION

In this present work we are presenting an improved model that will perform the similarity measure based on user web history. The analysis will find the next possible page and available it to the user before the user demand. The system will improve the web access efficiency.

REFERENCES

[1] Alexandros Nanopoulos," A Data Mining Algorithm for Generalized Web Prefetching", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 1041-4347/03 2003 IEEE

[2] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma, "Improving pseudo- relevance feedback in web information retrieval using web page segmentation", In Proceedings of the Twelfth International World Wide Web Conference, WWW2003, pp. 11-18, Budapest, Hungary, May 20-24, 2003.

[3] Cairong Yan," PARALLEL WEB PREFETCHING ON CLUSTER SERVER", CCECE/CCGEI, Saskatoon 0-7803-8886-0/05©2005 IEEE

[4] M. Junchang, G. Zhimin, "Finding Shared Fragments in Large Collection of Web Pages for Fragment-based Web Caching", Fifth IEEE International Symposium on Network Computing and Applications (NCA'06) 2006.

[5] Ruma Dutta," Offering Memory Efficiency utilizing Cellular Automata for Markov Tree based Web-page Prediction Model", 10th International Conference on Information Technology 0-7695-3068-0/07© 2007 IEEE

[6] Yong Zhen Guo," Personalized PageRank for Web Page Prediction Based on Access Time-Length and Frequency", 2008 IEEE/WIC/ACM International Conference on Web Intelligence 0-7695-3026-5/07© 2008 IEEE

[7] Payal Gulati," A Novel Approach for Determining Next Page Access", First International Conference on Emerging Trends in Engineering and Technology 978-0-7695-3267-7/08© 2008 IEEE

[8] Yong Zhen Guo," Error Correcting Output Coding-based Conditional Random Fields forWeb Page Prediction", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 978-0-7695-3496-1/08© 2008 IEEE

[9] B. de la Ossa," An Empirical Study on Maximum Latency Saving in Web Prefetching", E/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops 978-0-7695-3801-3/09© 2009 IEEE

[10] Alborz moghaddam," Dynamic and memory efficient web page prediction model using LZ78 and LZW algorithms", Proceedings of the 14th International CSI Computer Conference (CSICC'09) 978-1-4244-4262-1/09©2009 IEEE

[11] Andrea Bacic," Intelligent Interaction: A Case Study of Web Page Prediction", Proceedings of the *ITI 2009 31st Int. Conf. on Information Technology Interfaces*

[12] Yanjun Liu, "Strong Cache Consistency on World Wide Web", 2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)

[13] Bhawna Nigam," ANALYSIS OF MARKOV MODEL ON DIFFERENT WEB PREFETCHING AND CACHING SCHEMES", 978-1-4244-5967-4/10©2010 IEEE

[14] R.Khanchana," An Efficient Web Page Prediction Based on Access Time-Length and Frequency", 978-1-4244-8679-3/11©2011 IEEE

[15] Naved Ahmed, "Reducing User Latency in Web Prefetching Using Integrated Techniques", In proceeding of: IEEE, 2011

[16] Yaser Alosefer," Predicting Client-side Attacks via Behaviour Analysis using Honeypot Data", 978-1-4577-1127-5/11@2011 IEEE