An Integrated Approach to Forecast the Future Requests of User by Weblog Mining

Chintankumar S. Maisuriya
Research Scholar, CSE Department
Parul Institute of Engineering & Technology
Vadodara, Gujarat, India

Vaibhav Gandhi
Assistant Professor, CSE Department
Parul Institute of Engineering & Technology
Vadodara, Gujarat, India

ABSTRACT

With the wide availability of information over the internet and amount of its’ potential users’ makes it challenging for service provider to make such relevant information available to users in a fast and personalized manner nowadays, resulting in decrease in the web performance. One way to tackle with this challenging issue is to use prediction or recommendation technique, which can make users to discover future offerings. Web Usage Mining is a discipline, where the navigational access behavior of users’ over the web is tracked and scrutinized. By collecting and analyzing this behavior of user activities, websites owner can easily identify the web access patterns of its users’. The mining of user access logs could help to improve the performance of search engines, since users have a specific goal when searching for information. In this paper, an approach is proposed to get more accurate prediction results by using both bi-clustering and greedy search method. The predictions of users’ future access requests will be made by this can improve the accuracy of results and will helps in order to reduce the search time.

General Terms

Data Mining

Keywords

Future Requests Prediction, Pattern Discovery, Users’ Navigational Behavior, Web Usage Mining.

1. INTRODUCTION

Today in the era of information, Data is very crucial because it dominates the world more than any time before and needed to be ubiquitous. Thus, World Wide Web (WWW) has turned to be the significant and largest source of information retrieval. Web data might be either web data pages or data describing the activity of users [13]. It is very significant to examine the usage of the websites by its users’. By collecting the browsed information from the weblogs could give significant insight into, for example- what are the browsing patterns of the users’? Which will help to predict, for example- what will be the next requests of the users? Thus, there has been huge interest towards web mining. In web mining, data mining techniques are used to automatically discover and extract information from web data and services [14]. Web mining can be divided into three general categories: web content mining, web structure mining and web usage mining [13].

Web usage mining is the process of analyzing and automatically discovering the useful knowledge from the data collected in the web log files. The task of modeling and user’s navigational behavior on web can be useful in quite many web applications such as web caching, web page recommendation, web search engines and personalization [13]. The collected web log file and pattern analysis click streamed knowledge helpful to web usage mining which can recommend a set of objects to the active user, possibly consisting of links, ads, text or products, tailored to user perceived preference [6].

User’s future request prediction is a technique of weblog mining for predicting the next page access requests of user. The significance of prediction is to increase the browsing speed efficiently, to decrease the user latency as well as possibly reducing the loading of web server [11]. Over the last few years, lots of research work has been done in this field. The main motivation behind this work is to know what kinds of research has been done in the field of web usage mining for future requests prediction, thus by extracting and analyzing the browsing patterns of users will helps to find out what will be the next webpage requests from the user.

2. BACKGROUND

Recently, several weblog mining systems have been proposed for predicting users’ navigational behavior and their preferences. Identification of web browsing strategies is a crucial step in web designing and evaluation, which requires approaches that cater information on both the extent of particular type of user behavior and incentives for such behavior. Thus, pattern discovery from web data is the fundamental element of web mining and it assemble algorithms and techniques from several research areas.

In 1996, a new concept has been proposed that was called “Web mining” [14]. The data mining techniques has used to automatic discovery and extract information from abundant data on WWW. The in-depth study and research to all the methods of web usage mining has been carried out by Cooley et al. They have categorized data pre-processing task into subtasks and discussed the methods to pre-process log data [1]. The ultimate outcome of pre-processing stage should be data that allows identification of a particular user’s browsing pattern in the form of page views, sessions, and click streams have noted by them. The approach has been proposed to pre-fetch web pages into local cache in the background before an explicit request is made for them to reduce the delay in making future request to web objects by users because of its significance in reducing user perceived latency present in every web based application [2].

Moreover, a proxy based framework for pre-fetching web pages has been used to present user behavior by sequences of consecutive web page accesses, derived from the access log of a proxy server [3]. A novel web-based recommendation approach called ‘WebPUM’ has been proposed by classifying user navigation patterns to predict users’ future intentions based on the new graph partitioning algorithm for automatic personalization [6]. The fusion of web usage and web structure has been also very popular among researchers in web mining and has used both by collaborating information from web access log and website structure repository [8]. A Page Rank-like algorithm has also used for conducting web page access prediction [10].

The use of web usage mining techniques to identify similar access patterns from web log and to form highly dense clusters by pair-wise nearest neighbor (PNN) based clustering using only k- neighbors have explored by [7]. User navigation patterns are also identified and retrieved using Clustering as well as Classification techniques from web log data and predictions of users future movements is done based on it [9].Apart from classical clustering, fuzzy clustering approach is also well-known and used for predicting user future access request by using fuzzy clustering methods as fuzzy c-means (FCM) and kernelized fuzzy c-means (KFCM) [11]. In [12], a hybrid approach has proposed by combining two existing approaches and it focuses on the combination of Association Rule, Frequent Pattern (FP), FP tree for the pattern discovery and pattern.

In summary, all of these works attempt to predict the users' future webpage access requests to improve the quality of prediction systems, but alas all of these existing approach works well for particularly small sized dataset and the results are still below the satisfaction. For large sized data the accuracy of the resulting page predictions becomes the major issue. Thus, users' might be getting improper results during web surfing. In our work we advance the existing methodology and propose an integrated approach to predict user intention in the near future request.

3. PROPOSED METHODOLOGY

The flow diagram of the proposed methodology is given in the Figure 1.

3.1 Data Pre-Processing

Every request is made by the users to view particular web pages or any other documents are stored in weblog files which are automatically created and can be obtained from server. The clickstream data is defined as a sequence of Uniform Resource Locators (URLs) browsed by users within a particular time period. Therefore, it requires to be pre-processed to remove irrelevant entries from it.

Only the entries with the extension .html, .asp, .php and .jsp are accepted because it directly refers to actual webpage. Entries containing the extension like .jpg, .jpeg, .bmp, .png, .gif, .css, etc. are removed because it does not show the actual behavior of users. Users are identified by IP address and sessions are identified based on the time threshold.

3.2 User Access Matrix

To recognize 'how many pages accessed by particular user' and 'how many users accessed particular page', there is a need to generate web access matrix of users. The retrieved data pattern from previous stage is converted into web user access matrix U_{AM} in which rows represent users and columns represent pages of website. It is used to describe relations between web users and pages who accessed by users. The element a_{ij} of U_{AM} represents the frequency of the user u_i of u visit the page p_j of p during a given period of time.

$$a_{ij} = \begin{cases} hits(u_i, p_j), if\ p_j is\ visited\ by\ u_i \\ 0, othervise\ take\ it\ as\ zero \end{cases} \quad (1)$$

Where a hit (ui, pj) is the count of user u_i accesses the page p_j during a given period of time.

3.3 Bi-Clustering

Basically, clustering approach is used to discover interesting patterns of group of users based on their relevance for visiting particular website. Traditional clustering typically partitions users according to their similarities of searching behavior under all webpages. However, it might be possible that some users behave uniformly only on a subset of pages and their behavior inconsistent over rest of pages. In such case, traditional clustering methods fail to recognize such user groups. Thus, to overcome such problems bi-clustering approach is used.

Bi-clustering approach involves simultaneous clustering of rows and columns of a data matrix. It is being used to cluster web users and pages simultaneously based on similar interest under specific subset of webpages. Here k-means clustering method is applied on user access matrix U_{AM} *(U, P)* along both dimensions separately to generate C_u user_clusters and C_p page_clusters. Then combine the results to obtain small co-regulated sub-matrices *(C_u X C_p)* referred bi-clusters.

A measure called Average Correlation Value (ACV) is used to measure the degree of coherence of the bi-cluster. It is used to evaluate the homogeneity of a bi-cluster. It suggests high similarities among the rows or columns and used to find correlated/coherent bi-cluster.

3.4 Greedy Search Approach

A greedy search approach repeatedly executes a search procedure to maximize the bi-cluster based on scrutinizing local conditions. It is intended to maximize the volume of the bi-cluster seed without degrading the quality of measure. Here ACV is used as merit function to grow the seeds Therefore, it is used to refine the clusters and to make optimal solution to users with the hope that outcome will lead to a desired outcome for global problem and based on that prediction of users' future accesses is being made.

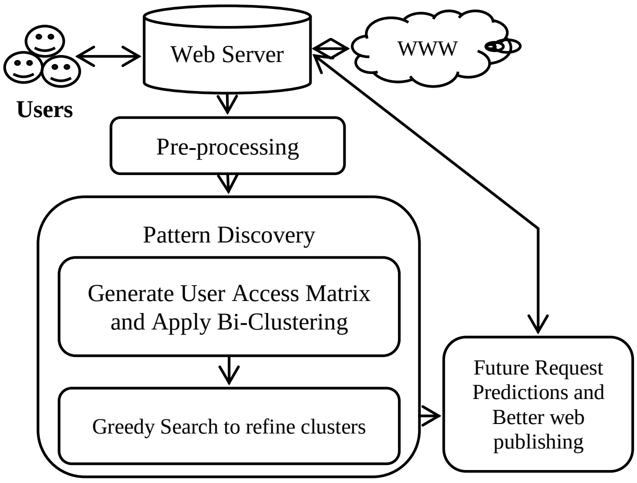
Users
Web Server
WWW
Pre-processing
Pattern Discovery
Generate User Access Matrix and Apply Bi-Clustering
Greedy Search to refine clusters
Future Request Predictions and Better web publishing

Fig 1: Process Flow of Proposed Prediction Approach

4. EXPERIMENTAL RESULTS

For our proposed approach, we have taken the msnbc dataset with 17 different page categories from msnbc.com portal that consist of Internet Information Server (IIS) logs for msnbc.com and news-related portion of msn.com. Initially the dataset consists of 989818 user records in the log file and after pre-processing 5561 user sessions are taken for the experimentation. The clickstream sequences are converted into user access matrix UAM (U, P) using equation (1).

Table 1. Results after Pre-processing

	Before Pre-processing	After Pre-processing
Total no. of transactions	989818	5561
Memory Used	12,287 KB	88.6 KB
Total no. of unique users = 5561		

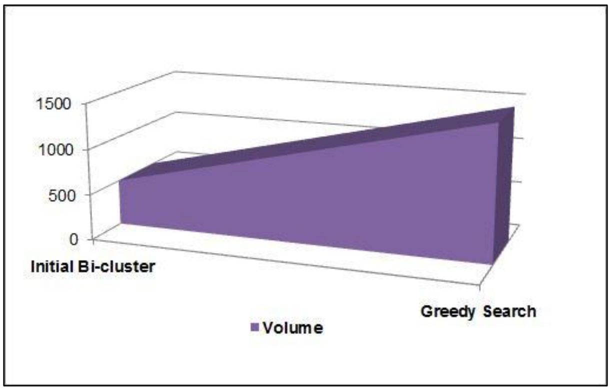

Fig 2: Average Volume of bi-clusters

Figure 2 shows the average volume of bi-clusters before and after applying the greedy search approach. The higher volume of bi-clusters may give the accurate results for predictions.

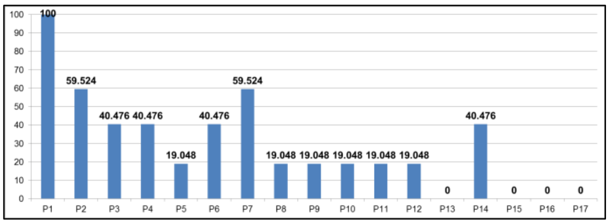

Fig 3: Average access frequency of webpages

Figure 3 shows the average access frequency of all the webpages of msnbc dataset. The webpage with higher the access frequency having the higher the probability for making predictions about users next access request which will be used to recommend the next pages to the users.

Figure 4 shows the comparison of prediction accuracy of existing and proposed approach. By using the proposed approach we can get approximately up to 88% prediction accuracy.

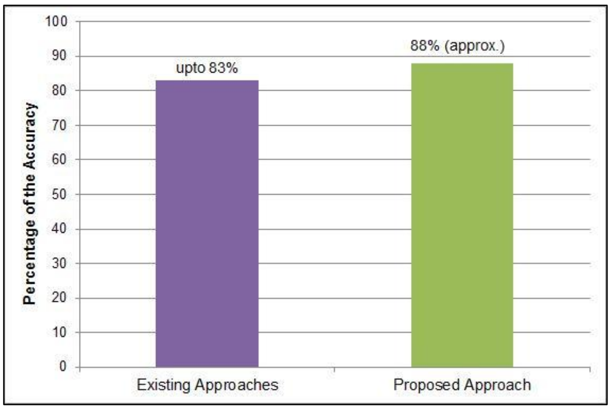

Fig 4: Comparison of prediction accuracy

5. CONCLUSION AND FUTURE WORK

As the information of page of web site is huge and develops rapidly, increasing the user's browsing speed efficiently as well as possible and reducing the loading of web server become very important issues. In this thesis, we proposed an integrated approach to recognize the frequent access patterns by analyzing past users access behavior and based on that the browsing behavior of the user will be analyzed which is useful to predict the next page access requests from the user. The proposed approach is used to improve the accuracy of predictions for users' future requests to better the web performance.

The MSNBC dataset was used in our experiment. The experimental results show that the proposed approach improved the quality of clustering for user access patterns and the quality of next request predictions and also retains more meaningful results for user. From the results we could conclude that by using this proposed approach we can get the prediction accuracy up to approximately 88%.

In the future, work could be extended by applying it over the different kinds of websites to assess its performance. The work could be also extended by implementing the proposed approach on the parallel as-well as on the cloud technology to evaluate its effectiveness.

6. REFERENCES

[1] J. Srivastava and R. Cooley, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web data", SIGKDD Explor. Newsl, New York, USA, Vol.1, pp 12-23, Jan-2000.

[2] Alexandros Nanopoulos, Dimitris Katsaros and Yannis Manolopoulos "Effective prediction of web-user accesses: A data mining approach," in Proc. Of the Workshop WEBKDD, 2001.

[3] Yi-Hung Wu and Arbee L. P. Chen, "Prediction of Web Page Accesses by Proxy Server Log" World Wide Web: Internet and Web Information Systems, 5, 67–88, 2002.

[4] Christos Makris, Yannis Panagis, Evangelos Theodoridis,and Athanasios Tsakalidis "A Web-Page Usage Prediction Scheme Using Weighted Suffix Trees" © Springer-Verlag Berlin Heidelberg 2007.

[5] Nien-yi jan and Nancy P. Lin, “Web User Behaviours Prediction System Using Trend Similarity“ Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization, Beijing, China, September 15-17, 2007.

[6] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat, “WebPUM: A Web-based recommendation system to predict user future movements” Expert Systems with Applications 37, 2010.

[7] A. Anitha, “A New Web Usage Mining Approach for Next Page Access Prediction”, International Journal of Computer Applications, Volume 8– No.11, October 2010.

[8] Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma, “A Novel Approach for Predicting User Behavior for Improving Web Performance”, International Journal on Computer Science and Engineering (IJCSE), Vol. 2, No. 04, 2010.

[9] V. Sujatha, Punithavalli, “Improved User Navigation Pattern Prediction Technique From Web Log Data”, Procedia Engineering 30, 2012.

[10] Phyu Thwe, “Proposed Approach For Web Page Access Prediction Using Popularity And Similarity Based Page Rank Algorithm”, International Journal of Scientific & Technology Research (IJSTR), Volume 2, Issue 3, March 2013.

[11] Dilpreet Kaur, A.P. Sukhpreet Kaur, “User Future Request Prediction Using KFCM in Web Usage Mining”, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 2, Issue 8, August 2013.

[12] Kaushal Kishor Sharma, Prof. Kiran Agrawal, “A Hybrid Approach for Predicting User’s Future Request”, Proceedings of Fourth International Conference on Communication System and Network Technologies, IEEE, 2014.

[13] C. Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis, A. Tsakalidis, “A Web page usage prediction scheme using sequence indexing and clustering techniques”, Data & Knowledge Engineering 69, 371-382, 2009.

[14] Etzioni, O. “The World-wide web: Quagmire or gold mine,” in Communication of the ACM, 1996, 65-68.