Web Page Recommendation Using Efficient Weight Based Prediction System

Sheetal Kumrawat[1], **Pramod S. Nair**[2]

[1, 2]Department of Computer, Science & Engineering, RGPV, Indore, M.P., India

Abstract: ***Demands for more and more extraction of knowledge from the web data is increasing these days. This web data is sometimes available and sometimes hidden from the normal web user. In this regard, we studied the web usage pattern analysis. Usage mining is a kind of hidden web pattern analysis. Web usage data analysis is used in various applications such as user data recommendations, web mastering, web administrating, web pre-fetching and caching. The primary focus of this thesis work is to develop a web recommendation system. The proposed web recommendation system focuses on some relevant concepts such as behavioral analysis of user access patterns, personalization of data and predictive modeling. Proposed work gives the result obtained after study of a traditional recommendation system which is updated with some new additional parameter such as web usages data personalization, user navigational frequency analysis, session wise data analysis and time stamp data analysis. Additionally present work has also been concerned with predicting rarely accessed patterns. For this new weighted prediction technique is proposed in order to get more accurate user patterns. This weighted prediction technique includes isolation of the user personalization, reduced complexity and less resource consumption. This is observed that as the current access pattern or sequence increases, the prediction becomes more accurate and closer to actual page prediction. This prediction mechanism will greatly enhance performance and efficiency of proposed system.***

Keywords*:* Data Mining; Web Usages Mining; Web Recommender System; Web Page Prediction; Web Log Pre-processing; URL Weight Analysis

1. Introduction

World Wide Web contains a huge amount of information and data, some of the information is distributed using the contents of web pages and some of the information is not directly gained from the web directly. Web mining is a process to recover the meaningful and essential knowledge from the web. Web mining utilizes this web information and data mining techniques to extract fruitful patterns from the web. Web mining is divided in three main key domains. The content of web pages are analysed using content mining, web access log are analysed under web usages mining and the connectivity of the web pages are analysed under the structure mining.

Recommendation system is an essential application of web mining. That is frequently used with a number of applications which not only helps to recommend a web page in a recommendation engine but also contributes to enhance the performance of the prediction system and web servers. These techniques are also used for enhancing the predictive modeling, user navigational behavior analysis and web pre-fetching and caching. In recent years a number of different kinds of recommendation systems are developed. These systems are promised to provide more accurate prediction for the user behavior but there is need of such kind of the system which also introduces the recommendations of such web pages those are not regularly visited by the users. This approach of recommendation is newer and a little amount of work is available in literature. Thus the proposed work is dedicated to find an optimum solution for the recommendations on those URLs which are not frequently visited by the user.

2. Literature Survey

Qingtian Han [2] web Mining is the technique of useful patterns analysis and utilization of information to enhance the current web systems, finding applications need and trends .Web usage mining based recommendation systems try to find out user's behaviour of web access and recommend pages to user by finding best match of the similar browsing behaviour available on historical web access data.The user interests and needs change with time. Identifying these changes and adapting to them is a key goal of personalization [4]. Web personalization is the process of customizing the content and structure of a Web site to the specific needs of each user by taking advantage of the user's navigational behavior [5]. Web usage mining means analyzing the data generated by web surfer's sessions or behaviours [6]. A user may have a single or multiple sessions during a period of time. Presently sessions are identified either on Time based method or Navigation based method [7].Recommender systems help users in the effective identification of items suiting their wishes, needs or preference [12].Web site personalization can be defined as the process of customizing the content and structure of Web site to the specific and individual needs of each user taking advantage of the user's navigational behaviour. The steps of a Web personalization process includes the collection of Web data, the modelling and categorization of these data (pre-processing phase),the analysis of the collected data, and the determination of the actions that should be performed[7].There are different web usage data mining techniques and algorithms that can be adopted for pattern discovery and recommendation, which include path analysis, clustering, and associate rule [8].The requirement for predicting user needs in order to improve the usability and user retention of a Web site can be addressed by personalizing it [3].Rana Forsatiis[1]proposed effective and scalable technique to deliver the web page as recommendation. They use distributed learning automata to learn users' behaviour and usages of clustering for pattern learning. The key contribution of the work is demonstrated by dealing with unvisited or newly arrived pages. They also recommend rarely visited or newly added pages as recommendation. Thus they introduced a novel Weighted Association Rule mining algorithm, with the HITS algorithm to extend the seed of recommendation. The proposed system evaluation shows that it improves the quality of web recommendations. This work is required to be extended and improve their parameters by which the application accuracy and performance becomes more effective.

3. Proposed Data Model

In this paper the primary focus is to develop a web recommendation system. The proposed web recommendation system focuses on some relevant concepts such as behavioural analysis of user access patterns, personalization of data and predictive modelling. Proposed work gives the result obtained after study of a traditional recommendation system which is updated with some new additional parameter such as web usages data personalization, user navigational frequency analysis, session wise data analysis and time stamp data analysis. Additionally present work has also been concerned with predicting rarely accessed patterns. For this new weighted prediction technique is proposed in order to get more accurate user patterns.

This weighted prediction technique includes isolation of the user personalization, reduced complexity and less resource consumption. It is observed that as the current access pattern or sequences increases, the prediction becomes more accurate and closer to actual page prediction. This prediction mechanism will greatly enhance performance and efficiency of proposed system.

A) System Architecture

The proposed recommendation system of the next user web page prediction is demonstrated using figure 1.

a) Method

Step 1: Select a web log which is already stored in the system.
Step 2: The web access log is processed to remove the unwanted data attributes from the web access log.
Step 3: Find out the targeted attributes from the web log and preserve them separately in the temporary data base.
Step 4: Now calculate the parameters to do this, the temporary database information is used to evaluate the clustering mechanism of data for extracting web log pattern information.
Step 5: Next calculate the time domain data analysis parameter to do this, the temporary database information is used to evaluate the amount of time spent on a particular URL in the entire log file.
Step 6: Next calculate the frequency based data parameter to do this, the temporary database information is used to evaluate the navigational frequency count of each URL accessed in entire log file.
Step 7: Next calculate the session based data analysis parameter to do this, the temporary database information is used to evaluate morning, afternoon and evening accessed sessions of URL in the entire log file.
Step 8:The parameters contributes to generate the weight, the system predicts the most optimum URL as prediction or recommendation.
Step 9: On the basis of weight computation of individual URLs, the system predicts the most optimum URL as prediction or recommendation. Finally the performance factors are stored in the temporary data base.

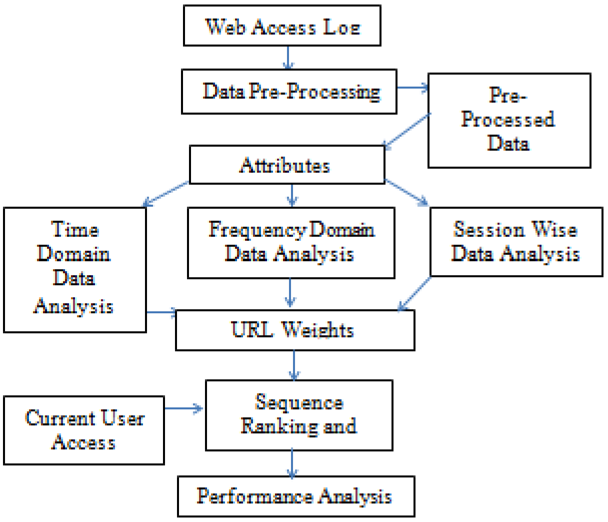

Figure 1: System Architecture

B) Web Access Log

A server log is also known as the web access log file created and maintained by web server to manage the activities. A typical web access log maintains a record of page requests made by clients or end user. The W3C is a standard and default web log format for server log. The new entries are appended at end of the file. This contains information about, client unique address, date and time stamp of request, request method, response code, request protocol, HTTP code, operating system, operating system version, browser and browser bytes sent, bytes received, user agent, and referrer are typically added.

C) Pre-Processing

The input raw web access log file is produced as input to this phase. In this phase the data is processed and is filtered to obtain the cleaned and transformed attributes. Therefore the individual attributes are extracted from web access file and processed or extracted data is arranged on a relational database. This pre-processed data is used further for finding the patterns over the database.

D) Attributes Selection

During the pre-processing of log files the selected or targeted attributes are extracted and preserved in a database. These attributes are used for computing the different parameters on which the prediction of next web page is performed. Therefore three key steps are described for further data processing and management. Thus the data is grouped according to the IP address available on web log files and their associated accessed data is used.

E) Time Domain Analysis

Using the filtered log data or pre-processed data the time stamp is analysed for a single or targeted user. In this evaluation the first a user accessed URL and the amount of time is computed to know the amount of time consumed with the URL. That can be understood by the given table 1examples.Suppose a user accessed the following five pages:

Table 1: Example URLs

IP	*URL*
192.168.1.4	www.rediffmail.com
192.168.1.4	www.yahoo.com
192.168.1.4	www.facebook.com
192.168.1.4	www.msn.com
192.168.1.4	www.google.com

And for accessing a web page after accessing next web page user consumes a specific amount of time. For example the table 2 contains the amount of time consumed with a navigated URL.

Table 2: Example TimeStamp Computation

IP	*URL*	*Accessed Time Stamp*
192.168.1.4	www.rediffmail.com	4.7 sec
192.168.1.4	www.yahoo.com	18.5 sec
192.168.1.4	www.facebook.com	80.48 sec
192.168.1.4	www.msn.com	199.83 sec
192.168.1.4	www.google.com	392.3 sec

According to the given table a user consumes time on a URL is listed in table's third column. That can be computed using the following formula:

$$\mathbf{User_{At} = Time\ of\ request_2 - Time\ of\ request_1}$$

Where,
$User_{At}$ = User Access Time.

Now the amount of time which is accessed by the different user for the same URLs are also evaluated to prepare the combine time domain analysis.

$$\mathbf{URL_{At} = Time\ of\ request_2 - Time\ of\ request_1}$$

Where,
URL_{At}= URL Access Time.

F) Frequency Domain Analysis

Now the pre-processed data is again utilized for finding the frequency of a web page access. That is also computed for both the ends first for individual user and also for the entire web access log. For example the table 3 contains the URLs for a single user and the amount of times the URL navigated by the target user.

Table 3: URL Frequency

IP	*URL*	*Frequency Count*
192.168.1.4	www.rediffmail.com	5
192.168.1.4	www.yahoo.com	8
192.168.1.4	www.facebook.com	16
192.168.1.4	www.msn.com	2
192.168.1.4	www.google.com	21

Similarly the entire log based navigation for the URLs navigated by a user is also obtained that is shown in a table 4.

Table 4: URL Frequency for Entire Log

URL	*Frequency Count for Entire log*
www.rediffmail.com	58
www.yahoo.com	109
www.facebook.com	298
www.msn.com	10
www.google.com	299

Here two other parameters are extracted first the URL frequency user wise and for entire log file that is denoted using $\boldsymbol{f_U}$ and $\boldsymbol{f_L}$ respectively.

G) Session Based Analysis

In this phase the session based navigational patterns and similar user browsing patterns are analysed. Therefore in this phase two different factors are used for evaluating the weights of data. Therefore the associated data from a single user is sub-divided into three observations these observations are demonstrated using the table 5.

Table 5: URL Observations

IP	**URL**	**Session 1**	**Session 2**	**Session 3**
192.168.1.4	www.rediffmail.com	2	1	4
192.168.1.4	www.yahoo.com	1	5	5
192.168.1.4	www.facebook.com	2	1	10
192.168.1.4	www.msn.com	1	0	1
192.168.1.4	www.google.com	3	7	15

And in similar ways the similar sessions which are more nearest to the navigated patterns of an individual user is demonstrated for entire web log data using table 6.

Table 6: URL Observations for Entire Server Log

URL	**Session 1**	**Session 2**	**Session 3**
www.rediffmail.com	5	4	7
www.yahoo.com	6	9	10
www.facebook.com	12	9	18
www.msn.com	7	5	6
www.google.com	8	10	19

Using the table observation for a given user is concluded as $\boldsymbol{O_u}$ and by the table entire navigational observation is prepared as$\boldsymbol{O_e}$.

H) URL Weight Analysis

In previous section the computations of the parameters that are used for calculating the weights which helps to find the appropriate URL is computed. In this section the weights of each URL according to the recovered parameters are discussed.

The desired weight for the URLs which can be the next web page for user is computed using the following formula:

$$\mathbf{W = F_u * F_L + O_e * O_u + User_{At} * URL_{At}}$$

Where,
$User_{At}$= User Access Time.
URL_{At}= URL Access Time.

I) Sequence Ranking and Prediction

In the previous phase for the both the sides (user and similar to user) the aggregated URL weights are evaluated for the given system. And according to the obtained weights the URLs are sorted and the maximum weight is selected as prediction of recommendation system. The highest weight shows the higher probability of visiting a web page after the current navigated web page.

J) Current User Access

The end user access a number of web pages in an active session. User accessed first web page is the current user access of page. The system accepts the first web page and predicts the user's next web page. But this is observed that when the current access pattern or sequence increases the prediction becomes closer to the actual page prediction. After computing the user next web page for a user, the performance of the system is evaluated in terms of accuracy, time complexity and the space complexity.

K) Performance Analysis

After computing the user recommended web page, the performance of the system is evaluated in terms of recommendation time, accuracy, time complexity and the space complexity.

4. Results Analysis

The given section includes the performance analysis of the implemented algorithms for the recommendation systems. The performance of algorithms are evaluated and compared in this chapter.

A) Training Time

The amount of time consumed during the training of the system is termed here as the training time of the algorithm. During the experiment, the performance of the prediction system is evaluated. As the size of dataset increases best results are reported. Graph represents the training time of prediction system where X axis of graph shows the dataset size and the Y axis shows consuming time in milliseconds.

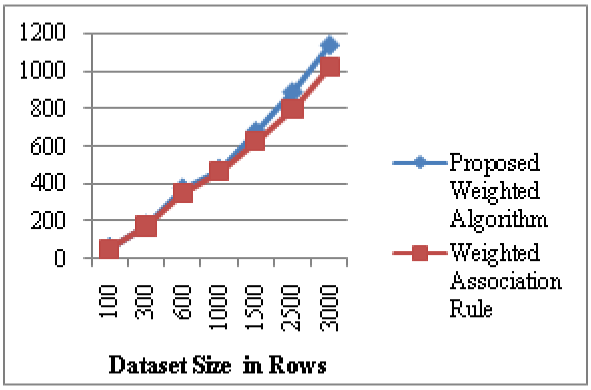

Figure 2: Training Time Consumption

According to the observations the training time consumption is dependent on the size of data. The proposed algorithm consumes additional time as compared to the traditional algorithm because for improving the accuracy in prediction the system needs to calculate the additional parameters as compared too traditional method thus the time complexity of proposed prediction is higher as compared to traditional prediction system.

B) Recommendation Time

The amount of time required to evaluate the URLs form making accurate prediction is termed here as the recommendation time. Graph represents the recommendation time taken by prediction system to generate recommendations. Where X axis of graph shows the dataset size and the Y axis shows recommendation time in milliseconds.

According to observations the amount of recommendation time taken by prediction system during prediction is not much fluctuated and not also affected by the amount of data to be process. The comparative results of the systems show the effectiveness of the proposed technique that consumes less time to compute URL as compared to traditional approach because the time based data clustering reduces the amount ofdata analysis which traditional system has to process.

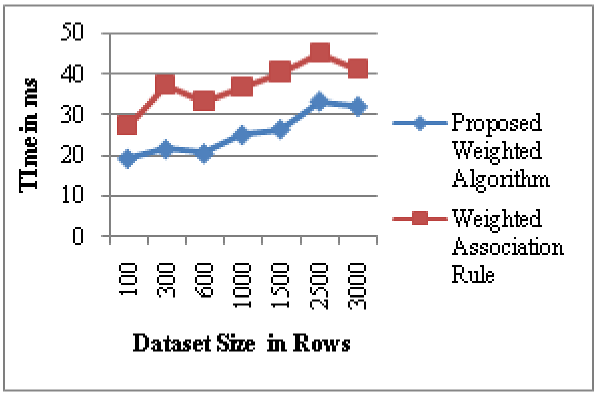

Figure 2: Recommendation Time Consumption taken by System

C) Memory used

The memory consumption shows the amount of main memory required to process the algorithm task. That is also known as the space complexity of algorithm. Graph represents the memory consumption to process the task. Where X axis of graph shows the dataset size and the Y axis shows memory consumption in kilobytes. According to the experimented results the amount of memory consumption is similar and not more fluctuating. But the respective proposed approach is more efficient than the traditional approach.

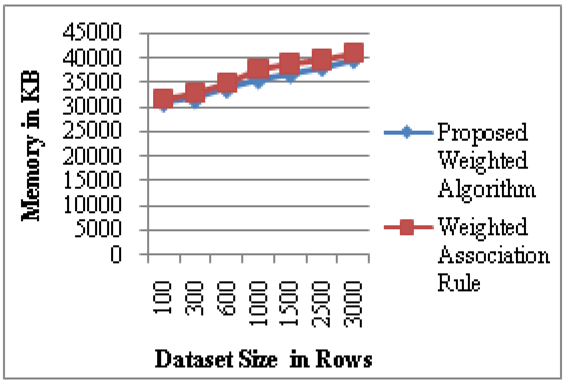

Figure 3: Memory Consumption to Process the Task

D) Accuracy

The accuracy of the predictive algorithm provides the amount of generated recommendation is similar to actual outcomes. It defines the amount of correctly recommendation patterns among the total samples produces to test. Graph represents the percentage of accurate URL recommendation. Where X axis of graph shows the dataset size and the Y axis shows accuracy in percentage.

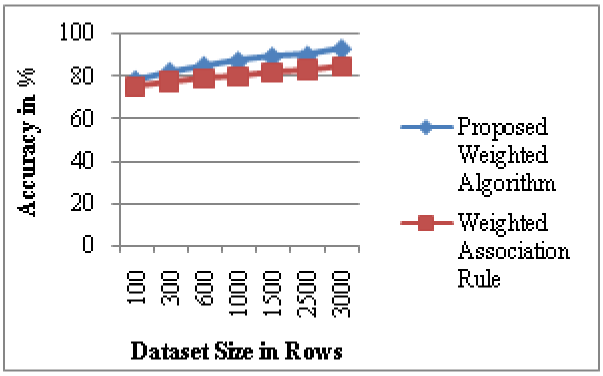

Figure 4: Percentage of Accurate URL Recommendation

According to the comparative results the performance of the proposed algorithms remains much consistent and improved as compared to the traditional algorithm. In order to evaluate the accuracy of algorithm a fixed amount random URLs are extracted from database. According to the obtained results the proposed system delivers the most accurate and efficient results as compared to the traditional system.

The performances in terms of time consumption, memory consumption and recommendation time consumption are evaluated. Finally by the cross validation process the system performance in terms of accuracy is measured. According to the obtained results the proposed system delivers the most accurate and efficient results as compared to the traditional system.
The performance summaries of both the algorithms are given in the table 7.

Table 8: Comparative Performance Summary

S. No.	*Parameters*	*Proposed system*	*Traditional system*
1	Memory consumption	Adoptable	Adoptable
2	Training time	High	Low
3	Recommendation time	Low	High
4	Accuracy	High	Low

The proposed recommendation system performances were evaluated. During the development and experiments the proposed system is found as more effective and accurate as compared to traditional methods.

5. Conclusion and Future Works

The web recommendation system is a concept by which the previous or historical user navigation data is analyzed. In order to find an accurate and efficient data model for web recommendation, a number of research articles are studied and a more promising model as given in [1] is selected for further study. The given model not only consumes the user accessed web pages that also evaluate the similar access patterns. This concept is used to predict the web pages that rarely accessed by the user.

Therefore to design more accurate model a weighted prediction method is proposed. The proposed recommendation system first find out the user accessed patterns form the web log and similarly the different users' accessed data also extracted using the K-mean clustering algorithm. Additionally the time based data clustering is also prepared to add more refinement on patterns analysis. After evaluation of all three parameters named as user navigational frequency, time based frequency and session wise data access pattern, a combined weight for all the URLs is evaluated. These weights are further sorted and by the rank of weights the most possible web page is predicted.

Proposed work focuses on some new additional parameters such as web usages data personalization, user navigational frequency analysis, session wise data analysis and time stamp data analysis to calculate weight of individual URL for web page prediction. We can work on the varieties of the different parameters to fulfil the efficiency and accuracy of recommendation system and can improving the training time of the proposed system.

References

[1] RanaForsati, Mohammad Reza Meybodi, AfsanehRahbar, “An Efficient Algorithm for Web Recommendation Systems”, IEEE, pp. 579-586, 2009.
[2] Qingtian Han, XiaoyanGao, Wenguo Wu, “Study on Web Mining Algorithm Based on Usage Mining”, Computer-Aided Industrial Design and Conceptual Design, November 2008.
[3] MagdaliniEirinaki and MichalisVazirigiannis, “Web Mining for Web Personalization”, ACM Transactions on Internet Technology, vol. 3, no. 1, pp.1-27, February 2003.
[4] H K Sawant, Shah Ashwini V., “An Evaluation of Techniques for Adaptive Search Web Mining Framework”, International Journal of Computer Technology and Electronics Engineering, vol. 1, no 2, pp. 51-57, 2010.
[5] PanagiotisGermanakos, ConstantinosMourlas, CharalIsaia, George Samaras, “An Optimization Review of Adaptive Hypermedia and Web Personalization – Sharing The Same Objective”, Available from George Samaras, September 2015.
[6] SnehaPrakash, “Web Personalization using web usage mining: applications, Pros and Cons, Future”, International Journal of Computing Science and Information Technology, vol.3, no.3, pp. 18-26, 2015.
[7] Neha Sharma &PawanMakhija, “Web usage Mining: A Novel Approach for Web user Session Construction”, Global Journal of Computer Science and Technology: E Network, Web & Security, vol. 15, no 3, 2015.
[8] D.A. Adeniyi, Z. Wei, Y. Yongquan, “Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method”, Saudi Computer Society, King Saud University, Applied Computing and Informatics, Elsevier, 2015.
[9] HaidongZhong, Shaozhong Zhang, Yanling Wang, ShifengWeng and YonggangShu, “Mining Users' Similarity from Moving Trajectories for Mobile Ecommerce Recommendation”, International Journal of Hybrid Information Technology, vol.7, no. 4, pp. 309-320, 2014

[10] “Web Data Mining Analysis”, http://shodhganga.inflibnet.ac.in/bitstream/10603/9532/5/05_chapter%203.pdf.

[11] K. Srinivas, P. V. S. Srinivas, A. Govardhan, V. ValliKumari, “Periodic Web Personalization for Meta Search Engine”, IJCST, vol. 2, no. 4, December 2011.

[12] A. Tejeda-Lorente, C. Porcel, E. Peisc, R. Sanz, E. Herrera-Viedma, “A quality based recommender system to disseminate information in a University Digital Library”, Information Sciences, October 2013.