International Journal of Current Trends in Engineering & Technology
www.ijctet.org, ISSN: 2395-3152
Volume: 06, Issue: 06 (November - December, 2020)

Genetic Algorithm Based Web Page Prediction Using Weblog Feature

Rajesh Kumar Nigam[1], Dr. Chandikaditya Kumawat[2], Dr. Manish Shrivastava[3]
[1]Research Scholar, CSE Deptt., Mewar University, Rajasthan (India) (rajeshrewa37@gmail.com)
[2]Professor, CSE Deptt., Mewar University, Rajasthan (India) (chandikaditya@gmail.com)
[3]Professor, IT Deptt LNCT, Bhopal (India) (contct.manishshrivastav@gmail.com)

Abstract—*The website depends on user visits to increase user engagement, and on-site web portals are working on page prediction algorithms. This paper has proposed a web page prediction model using the weblog feature. The raw weblog feature was pre-processed by applying the Markov model with a hierarchical structure. Markov model gives web page pattern with support value. Genetic algorithm steps did user page prediction. In a genetic algorithm, the fitness function uses pattern support value. The experiment was done on a real weblog dataset. Results were compared with the existing model on different evaluation parameters.*

Keywords- Data Mining, Pattern Reorganization, Feature Extraction, Web mining, Clustering.

I. INTRODUCTION

Web usually occupies many data, which brings challenges related to the efficiency to load the web pages in the allotted bandwidth. It is essential to improve the users' experience so that they are encouraged to return to any website. Web usage mining deals with such problems and so have attracted many people in real-time application. It is essential to understand the web applications to improve the user's experience in browsing websites; it is also called QoE or quality of experience. Web browsing is a process in which the clicking of the users triggers the application, and the user is navigated to other web pages. If there is some way that the expected user's interaction is prefetched in the browser beforehand, then it can make the next page accessible to the user within a fraction of a second. Prefetching of the web pages is to store several multimedia from the external domains to reduce the waiting time of the users who need such resources. However, it is not an easy process to predict the interaction of the user on web pages. Collecting relevant and detailed information on the user's expected response is difficult due to the target application's limitation and the platform. Many previous researchers have attempted [3]-[5] to collect the user's information from the server-side, but that was limited to only some web-based applications. Such applications are found helpful to collect the prediction of the users who surf several web-based applications. Some other works were related to the client-side and focused on collecting application executions, phone calls, clicks, and user interaction data. But such client-side data can only be implemented with modification of the source code of the web browser, or else they only are confined to be used on mobile platforms. The measurement overhead creates a problem in the scalability of the data collection, leading to limited information regarding user's interaction in real-time applications. The remaining paper was arranged in a few sections, while the second section gave several web mining methods to extract the user's information. The third section describes the complete proposed models with feature extraction—the fourth section brief experimental work with comparing tables. In the last section whole work was concluded with future work.

II. Related Work

M. A. Awad et al. in [7] Used the kth Markov and the Markov model to predict the user's interaction in web-based applications. They designed a new, improved version of the Markov model in which the problem of scalability and the number of paths problem was addressed. Their experiment showed that the new, improved model was better as it reduced the number of paths.

D. Xu et al. in [8], Clustering of data in an unknown area under unsupervised learning that deals with data partition to filter the required data for further learning was discussed. Clustering is the concept of separating the sessions of the users based on the similarity of the user's actions. *T* is usually done to divide the web session into tiny groups in which the as the difference of the cluster is more the distance between them is more and vice-versa.

A. Verma et al. in [9] proposed a k-means algorithm to find a similar user session pattern. The work was based on using the k-nearest neighbour or KNN algorithm.

R. Manikandan et al. in [10] provide useful web pages to patients recommender system was introduced that uses certain agents. The prominent feature of Particle Agent Swarm Optimization (PASO) was that it creates an algorithm that is denoted by a set of particle agents whose work was to corporate in attaining any task. Web user particle agent and semantic particle agent were the two agents shown by this research. PASO Based Web Page Recommendation (PASO-WPR) is an intermediate particle agent or program that contains a user interface that has a collection of information as per the user's requirements. PASO-WPR was carried along with data mining techniques on web data to study similarity in web pages. As the web pages with multimedia files were viewed, patient navigation patterns were like ontology instances and were not from uniform resource locators. At the same time, the semantic similarity was used for page clustering.

Arta Iftikhar in [11] gave an improved method for collaborative filtering, which was based on triangle similarity. However, there was some drawback in triangular similarity as it gives only common ratings of the users. The proposed similarity theory proposed both common and non-common ratings of the users. Finally, the outcome was complemented with URP(user rating preference) to study the behaviour of rating performance. Ying Jin et al. in [12] proposed UIQPCA as a different and unique approach for the hybrid user and covering algorithm and item-based quality. UIQPCA collects information from both users and web applications based on the quality of the web services. As the information is collected, the target user and the target web application is matched and selected. UIQPCA worked well in dealing with target users and its match with the target website.
E.J. Thomson Fredrik et al. in [13] authors presented work that focuses on cosine the based similarity feature of a k-mean algorithm rather than the Euclidean distance k-mean algorithm, which was generally used traditionally. K-mean algorithm works on the grouping of similar data to find out the behavioural patterns and guess the next path of the browsing to make a group, and the k-means examine fixed k clusters available in a data set. The k-mean method is then made to learn to analyze the next step of the user.

III. Proposed Methodology

This section briefs a web prediction model into figure 1 shows a block diagram of the proposed model where the cleaning technique pre-processes the raw input dataset. After cleaning, patterns were extracted from the weblog using the Markov model with hierarchical; structure.

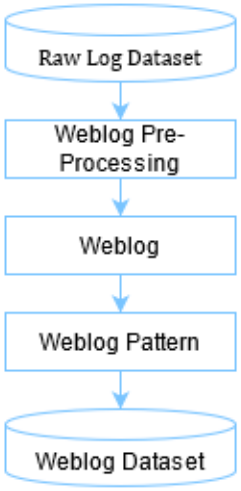

Figure 1 Block diagram of weblog Feature extraction.

Weblog Pre-processing

The work used cleaning techniques for pre-processing of the input raw log dataset. Each raw log file was treated separately to remove a few pieces of information from the log as the log has information of client web browser, operating system, visiting website page. The above log has the number of information related to client and website instance. Out of different information, this work utilizes a few of information for the work. Some data are selected, while other information is removed from the log and considers this step as cleaning in the weblog pre-processing. So, log dataset after cleaning looks change and processed information further to extract patterns.

Weblog Example
157.49.79.189 - - [27/Feb/2021:22:30:14 +0530] "GET /good-publishing-international-journals/ HTTP/1.1" 200 751 "Mozilla/5.0 (Windows NT 6.3; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.69 Safari/537.36

Weblog Example
157.49.79.189 [27/Feb/2021:22:30:14 +0530] "GET /good-publishing-international-journals/

Weblog

To generate weblogs of a user cleaned dataset was further processed and assign each requesting page to get a unique number U. This U help to prepare the user weblog sequence. To identify the user's IP address was used, and the timestamp log feature was further used to generate the log sequence. It can be understood by relating the User client IP and timestamp with page unique id. Now for getting the weblog, if IP 157.49.79.189 is considered as a user, then its weblog as per timestamp has page sequence /good-publishing-international-journals/

Weblog Pattern

A weblog is a sequence of web pages visited by a user in fix range of timestamp., Markov second and third-order model was used by the work to get the pattern from this dataset. Markov model generates support value from the weblog for various set of second and thirst order patterns. In order to reduce the database scan, the proposed model read the dataset file and increased the counter of the second, third-order pattern. The hierarchical structure was prepared for each node in the structure to act as a counter when a node has a matching pattern and then counter increases. Let us understand this pattern generation by figure 3. where nodes act as a counter of pages.
Nodes were arranged in three levels of hierarchy first level for first order, the next level for a second order, and the final level for third order. This structure works for each weblog were as per page id counter of matching order nodes were increases. So, for weblog {1, 3, 4}if first-order nodes [1, 3, 4] counter get increase by one, similarly counter second-order nodes of next level [(1,3), (1,4) (3,4)]. Finally, for third-order nodes [{1,3,4}] counter gets increased by one. The set of this pattern and the counter

was arranged in table 3, with the final support value having n number of weblogs.

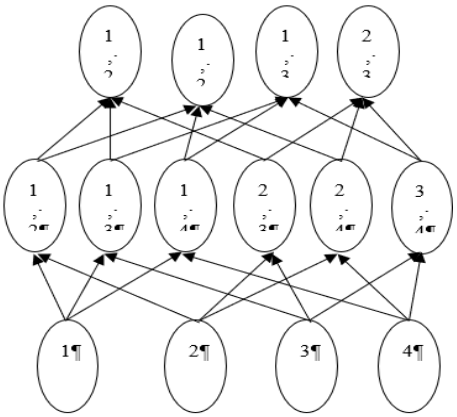

Figure 2 Hierarchical structure for Markov pattern counter of the second and third-order.

Prediction Model

Genetic algorithm steps were used for the prediction of the next page as per the user previous visits.

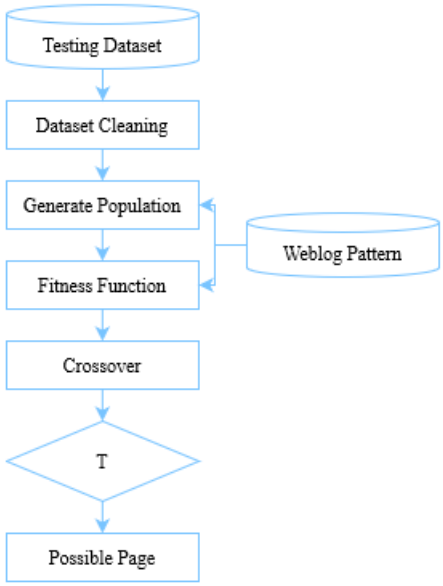

Figure 3 Block diagram of GAWPP.

Generate Population

The population collects chromosomes, and each chromosome is a subset of possible pages as per the currently visited web page. So, the generation of the population in this algorithm was the same as done by the weblog pattern. Random Gaussian function was used for the page prediction from possible pages of web pattern.

Fitness Function

In order to identify feasibility for a set of chromosomes present in the population. The fitness value of each need to be calculated. As work has utilized the weblog feature from the raw dataset to generate the PDM feature So this model uses Markov support value sum as a fitness value for a chromosome.

$$F_m = \sum_{i=1}^{p} \sum_{j=1}^{c} Web\_Pattern(C_j, P_{i,m})$$

F is the fitness of the m^{th} chromosome, where C is the current visits of the user, and P is the population of chromosome having n number of possible pages.

Crossover

As per fitness value, the best chromosomes was select from the population. This chromosome has a subset of possible pages that most likely to visit by a random user. So other sets of chromosomes in P get some characteristics from the best chromosome. It increases the learning of the whole class/population.

Possible Page prediction

After the t number of iteration steps (fitness function, crossover) final P population passes through the fitness function, the best-fitted chromosome is considered a possible page prediction subset. These chromosome element pages are predicted possible set of pages for the C visited pages of the user.

IV. Experiment and Results

The precision of a transaction is provided as the ratio of the number of web pages appropriately predicted, and the overall amount of web pages predicted.

Precision = Approximate_Correct_pages / All_predictions

Coverage is calculated as the ratio of the number of web pages appropriately predicted and the overall number of web pages visited by the user.

Coverage = Approximate_Correct_pages / All_Visited_Pages

M-metric is utilized to obtain a single evaluation measure, and it is defined in this manner.

M-metric = (2xPrecisonxCoverage) / (Precision + Coverage)

Shopping dataset: A shopping website provides code files for scholars "Project Tunnel" [14] and domain www.projecttunnel.com. Dataset has 238 pages with 12000 sessions.

Results

Table 1 Precision based comparison of web page

Dataset %	Previous work	GAWPP
35	0.375	0.708333
45	0.421622	0.718919

55	0.368889	0.657778
65	0.390977	0.661654
75	0.368078	0.664495

Prediction models

The precision value in table 1 shows that the GAWPP model has increased the page prediction precision value compared to the existing model in [11]. This increase in value was done by the use of Markov based weblog feature using a hierarchical structure. The use of genetic algorithm has also increased the possible page matching.

Table 2 Coverage based comparison of web page prediction models.

Dataset %	Previous work	GAWPP
35	0.073469	0.356643
45	0.079511	0.361413
55	0.067977	0.330357
65	0.070893	0.332075
75	0.065966	0.333333

The precision value in table 1 shows that the GAWPP model has increased the page prediction precision value compared to the existing model in [11]. This increase in value was done by the use of Markov based weblog feature using a hierarchical structure. The use of genetic algorithm has also increased the possible page matching.

Table 3 M-metric based comparison of web page prediction models.

Dataset %	Previous work	GAWPP
35	0.122867	0.474419
45	0.133791	0.481013
55	0.114799	0.439822
65	0.120023	0.442211
75	0.111881	0.443961

The M-metric value in table 3 shows that the GAWPP model has increased the page prediction m-metric value compared to the existing model in [11]. The genetic algorithm fitness function uses Markov support values for comparing chromosome sets in a population. This use of Markov support has increased work efficiency.

V. Conclusions

The Internet has become a huge source to obtain any information. As there is a transfer of information from physical media to digital online, collecting such information is crucial. This paper has proposed a weblog feature-based page prediction model. In order to extract the webpage feature log was pass through a hierarchical pattern counting structure. The genetic algorithm fitness function uses Markov support values for comparing chromosome sets in a population. This use of Markov support has increased work efficiency. The experiment was done on real data obtained from the "ProjectTunnel" site. The result shows that the proposed genetic algorithm of web page prediction has increased the precision value by 43.58%. In the future, the scholar can involve more feature in work to increase the prediction accuracy.

References

[1] Google Chrome. Google Chrome Privacy Whitepaper. Accessed: Sep. 18, 2019. [Online]. Available: https://www.google.com/chrome/privacy/whitepaper.html

[2] I. Grigorik. Resource hints. W3C Working Draft. Accessed: Jul. 2, 2019.

[3] B. Mobasher, J. Srivastava, and R. Cooley, "Automatic personalization based on Web usage mining," Commun. ACM, vol. 43, no. 8, pp. 142–151, Aug. 2000.

[4] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream Analysis for Sybil detection," in Proc. SEC, Washington, DC, USA, 2013, pp. 241–255.

[5] C. Li, Y. Lu, Q. Mei, D. Wang, and S. Pandey, "Click-through prediction for advertising in the Twitter timeline," in Proc. KDD, Sydney, NSW, Australia, 2015, pp. 1959–1968.

[6] H. Obendorf, H. Weinreich, E. Herder, and M. Mayer, "Web page revisitation revisited: Implications of a long-term click-stream study of browser usage," in Proc. CHI, San Jose, CA, USA, 2007, pp. 597–606.

[7] M. A. Awad and I. Khalil, "Prediction of user's web-browsing behaviour: Application of Markov model," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1131-1142, 2012.

[8] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," Annals of Data Science, vol. 2, no. 2, pp. 165-193, 2015.

[9] A. Verma and B. Prajapat, "User Next Web Page Recommendation using Weight based Prediction," International Journal of Computer Applications, vol. 142, no. 11, 2016.

[10] R. Manikandan. "A novel approach on Particle Agent Swarm Optimization (PASO) in semantic mining for web page recommender system of multimedia data: a health care perspective". Springer Science Business Media, LLC, part of Springer Nature Jan. 10 2019.

[11] Arta Iftikhar, Mustansar Ali Ghazanfar, Mubbashir Ayub, Zahid Mehmood, And Muazzam Maqsood. "An Improved Product Recommendation Method for Collaborative Filtering". IEEE Access, volume 8, July 20, 2020.

[12] Ying Jin, Guangming Cui, Yiwen Zhang, "Quality Prediction of Web Services Based on a Covering Algorithm", Complexity, vol. 2020, Article ID 8572161, 17 pages, 2020.

[13] E.J. Thomson Fredrik and Jothish Chembath. "Enhancing Web Page Prediction by Using Modified K-Means Clustering Method". Jardcs volume 11 issue 4, 2020.
[14] Dataset: https:projecttunnel.com