Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 3190- 3199

Probability Density Feature and Teacher Learning Based Web Page Recommendation

Rajesh Ku.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— With the drastic increase of the digital content on the servers, it is necessary to develop a algorithm for reducing a latency time of web pages. This work has proposed user behavior based page recommendation by the analysis of web features. Proposed work has utilizes two existing web features named as logs and content. With the help of existing features new probability density web feature was proposed by the work to increase the work efficiency. Due to dynamic nature of web user TLWPP (Teacher Learning Based Web Page prediction) model was proposed in this paper. Experimental work was done on real, live web-portal of a international Journal. Result shows that Use of two phase crossover operation in TLWPP model has reduces the time to get better solution.

Index Terms- Genetic Algorithm, Pattern Extraction, Webpage recommendation, Web mining.

I. INTRODUCTION

These days due to the tremendous increase in the number of web pages and websites the traffic of the network has drastically increased and so is the load that the web servers handle. Although the web clients have been provided with larger bandwidth they still face issues such as high latencies while navigating the websites because of the overloaded network elements such as servers and networks. There is a need for research [1] regarding the reduction of the users who face such latency problems while surfing the websites. Pre-fetching, web caching, and geographical replications are some of the best techniques to overcome such latency problems. Several huge companies use the concept of web replication by implying content delivery network to reduce the access time of their websites but such process is costlier which many small companies cannot afford [2].

Pre-fetching the process by which the users readily receive their desired object before their demand [3]. The process decrease latency of the website at some level. Such pre-process is based on the information of the domain knowledge to calculate the next page a user may desire. Random forest [4], Markov model [5], and the decision tree[5] are some of the primary ways through which the pre-fetching is achieved.

Researchers have made enormous efforts to generate semantic knowledge regarding we pages and such knowledge has been implemented in various formats. The database of the obtained knowledge was implemented in form of spreadsheets, relational databases, and text files. But such heterogeneous databases are difficult to manage and are nearly impossible to generate good quality web page recommendations.

II. Related Work

Chaipornkaew et. al. in [6] Given three machine learning techniques such as Apriri algorithms, TF-IDF, and K means. Out of which the TF-IDF was used to form the vectorization of the word using the webpage heading, K-means for clustering the heading of web pages, and Apoiri is used t find the association of the cluster of the web pages. To obtain an effective number of clusters elbow method was employed.

Bhavithra et. al. in [7] given a case-based reasoning method for webpage recommendation which was the extended part of collaborative filtering. In this, the profile of the users will be generated that contains eight features based on characteristics while two are the content-based features that are revealed by generating search logs of the web access. Case-based reasoning is identified by using the k-NN collaboration of the user profile. To increase the accuracy of the outcome weighted association rule mining is applied that generates the rules of the user profiles and thus predicts the web pages as per the keyword search of the user.

Saradha et. al. in [8] designed an effective way to determine the user profile which was based on usage-based attributes. The attributes of the user such as exit rate, page rank, bounce rate, and conversion rate are stored. Further, the concept of case-based reasoning is applied to the profile of the users to generate the cluster summary based on the search interest of the user. Thus the webpage that fits the requirement of the active reason is presented based on the cluster summary.

Guoguang et. al. in [9], given a unique e-commerce recommendation algorithm that is based on prediction through the BGN link. At first, the data of the user is achieved with the help of the distance formula to get the similarity of the user attributes. Further, the BGN is forwarded using SMN or single-mode network which is essential to achieve the accurate potential links from the BGN and thus the links are predicted by comparing the similarity feature.

Fatma et. al. in [10] given a unique approach for the entertainment industry to get the more valuable suggestion of the user based on their past interactions. It’s simply done to decrease the time duration and frustration of the customer by providing him with valuable content. The author has created an RSMCG or Recommendation System based on Markov Chains and Grouping of Genres to achieve the desired accuracy and to construct an intelligent system that uses Markov chains to predict the current actions of the user based on their past actions. A machine learning algorithm named DBSCAN was also adopted to identify the user's interest and their answers more accurately.

III. Proposed Methodology

Possible set of pages increases the combination set of predicted pages hence genetic algorithm works well in less execution time. This work has resolved same page prediction work problem by another algorithm name as Teacher Learning Based Optimization [12, 13]. This algorithm have two phase first is Teacher phase and other is student phase.
As per school / college teaching pattern a teacher teaches all students in a class is term as teacher phase of learning shown in fig.1. While when students study in their friend circle by one of friend then it term as

student phase learning shown in fig. 2. Working of whole model was shown in fig. 3 and explanation of each block was done in subsection of paper.

Generate Population

As per weblog patterns obtained from log feature possible pages were list in the vector for a testing webpage sequence. This vector is a chromosome or student as per teacher learning algorithm. Collection of these students is termed as population matrix [13]. So this work do not generate population by any noise n=generation function like Gaussian.

To represent teacher leaning population and chromosomes work use different notation like 'TLP ' Teacher Learning Population and Sc is student chromosome of TLP. Hence TLP is set of {Sc1, Sc2, Sc3.....Scw}, where w is number of students in the population. Further Sc1 is subset of e number of possible pages obtain from W weblog patterns as per C.

TLP←Teacher_Population(w, e, n, W, C)

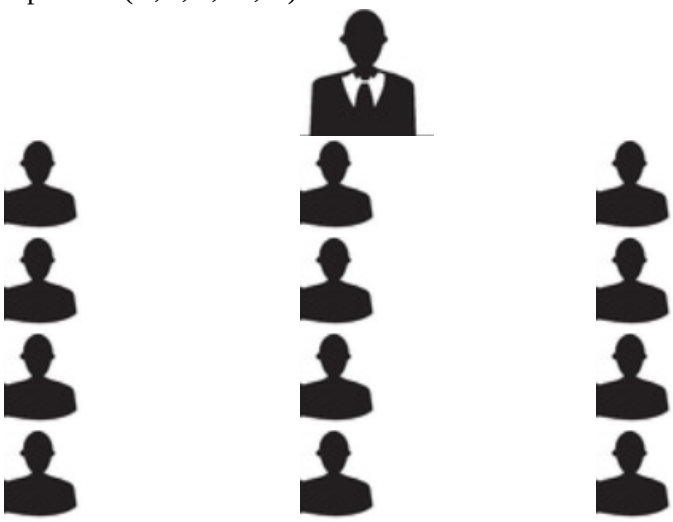

Fig. 1 Teacher phase graphical view.

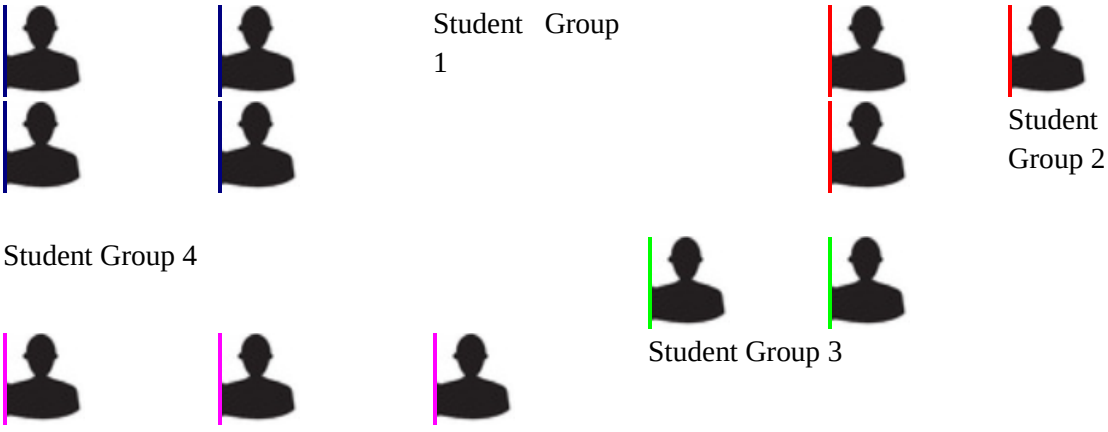

Fig. 2 Student phase graphical view.

Probability Density Feature

This feature combines weblog and web content features to predict user next page with a probability value. Weblog W gives an input of web pattern with S support value obtained from Markov model [14]. Weblog gives an set of next page with S value but each set of page have its own S value, hence chance of getting a particular page need more strength by involving web content feature by evaluating PDF.

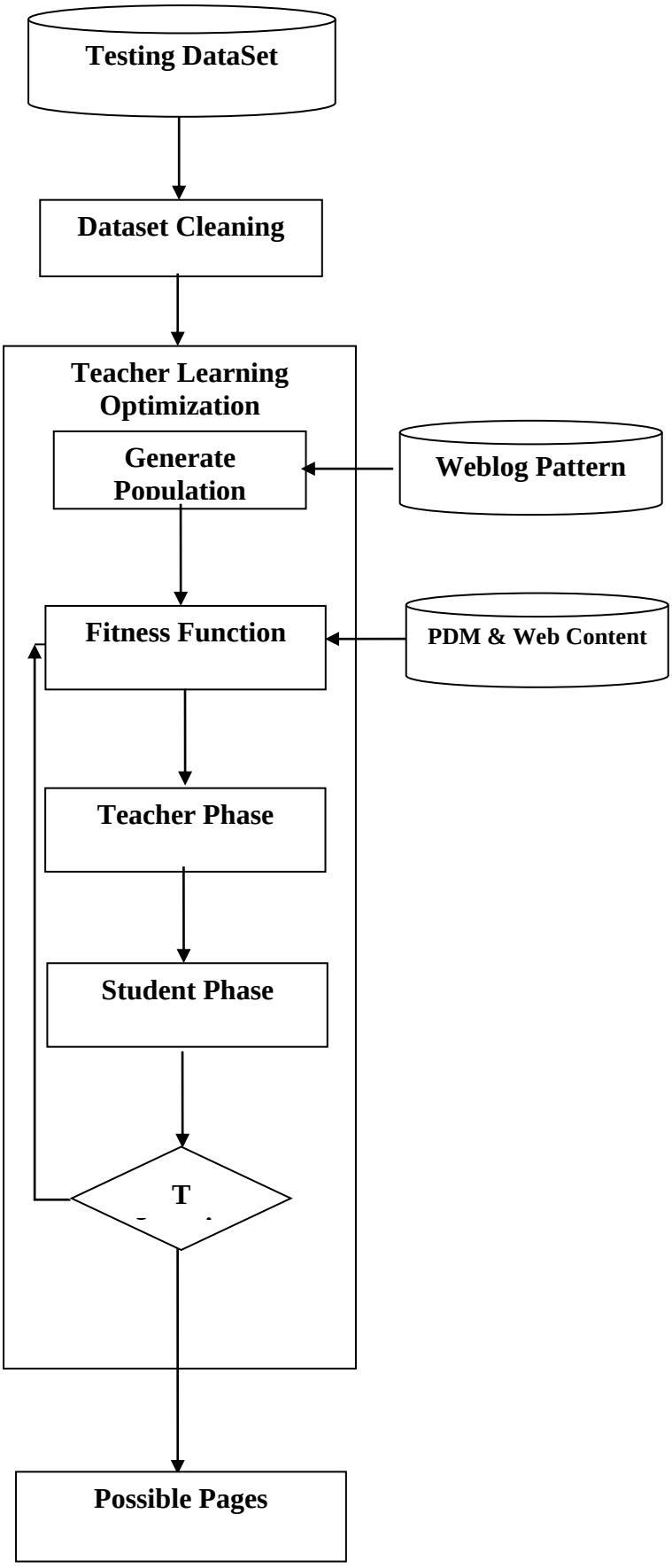
Testing DataSet
Dataset Cleaning
Teacher Learning Optimization
Generate Population
Weblog Pattern
Fitness Function
PDM & Web Content
Teacher Phase
Student Phase
T
Possible Pages

Fig. 3 Block diagram of proposed TLOWPP model.

Unique page U obtained from weblog feature of input dataset was search in each weblog pattern W. So a U_i page present in W_m $\{U_1, U_2, U_i\}$ having support value S_m use as a weblog input in PDF. Set of other pages present in W_m keywords were used to find the distance from U_i this act as the web content feature input in PDF. Distance between two pages i, j were evaluate by eq. 1 where matrix Pi column value was subtract from Pj column where j^{th} page found in pattern Wm.

$$D_{i,j} \leftarrow \sum_{r=1}^{n} |P_{i,r}, P_{j,r}| \text{----Eq. 1}$$

So this D is a vector of j number of elements. Each j^{th} element was added by the S_m support value. Equation 2 gives weblog and web content combined input vector to PDF.

$$F_{pdf} = D_{i,j} + S_m \quad \text{i.e. } m \in W \cap U_i \text{----Eq.2}$$

Further F_{pdf} vector was use to get mean μ and variance σ for getting PDV of ith unique webpage by Eq. 3 .

$$PDV_i = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{P_{PDF}-\mu}{\sigma}\right)^2} \text{----Eq. 3}$$

This PDV value store in Probability Density Matrix (PDM) have nxn dimension.

Fitness Function

Each candidate or chromosome in the population has its own existence but fitness function extimate value of likeliness to be user visit. This paper has evaluate probability density matrix value for different page as per web log and content feature. Eq. 4 was used in the model for getting fitness value of the work. To represent fitness value in this model TF_m=notation used.

$$TF_m = 1/\sum_{i=1}^{e}\sum_{j=1}^{c} PDM(C_j, TLP_{i,m}) \text{--------4}$$

Teacher Phase

As per fitness value obtain from the fitness function, best fitted chromosome or lowest fitness value chromosome was consider as teacher [15, 16]. Chance of opening the page set available in best chromosome is high as compared to other, so crossover operation perform by this step by replacing random page of teacher (Local Best Chromosome) to the other set of student chromosome.

This phase is used for the crossover of the chromosomes by the single best solution from the population. Best solution $Sc_{teacher}$ act as a teacher and its selection is based on the fitness value. Now selected teacher will teach other possible solution $Sc_{student}$ by replacing possible page unique id present in teacher solution. By this all possible solution which acts as student will learn from best solution which act as teacher. In order to do crossover operation random position range from 1 to 0 generate by Gaussian function, copy unique page id from the teacher chromosome and put same at same position in student chromosome. This improves the population quality.

$$Sc_{new,i} \leftarrow Crossover(Sc_{student,i}, Sc_{teacher,i}) \; i \in \{1,2, \ldots \ldots m\} \text{---------} 5$$

Where Scnew,i is the updated value of $Sc_{student,i}$. j is random position generate by Gaussian function.

In this phase best chromosome which has good fitness value relative power values as compared to existing act as teacher chromosome.

Student Phase

In this phase some random group of chromosome are made automatically. Each group was used for the crossover of the chromosomes by the single best solution in that group. Best chromosome in a group act as a trainer among other chromosomes and its selection is based on the fitness value. In order to do crossover operation random position range from 1 to e generate by Gaussian function, copy unique page id from the teacher chromosome. Each new chromosome was cross verified that either its fitness value improved then previous, if fitness improves then new chromosome is included in the population and older one get removed. Vice versa if fitness value not improves. In this phase student groups were developed initially which undergo into crossover operation.

Possible Page prediction

Once iteration reach to max number t then phases of learning get stop and final population was further evaluate for prediction. Fitness value of final set of possible page were estimate where current set of visited pages C play asn important factor.

IV. Experiment and Results

In order to evaluate proposed work website of two different domain were used. Each website log set was used for the prediction of user next page. Detailed description of each site was mention below:

Journal dataset: Its an research article publication journal "International Journal of Science, Engineering and Technology", have valid ISSN: 2348-4098 and domain www.ijset.in. Table 1 gives description of various features and corresponding values.

Tables 1: Journal dataset feature description.

Features	Value
Pages	158
Logs	10000
User	2264
Date	Feb-Mar-2021

Result

Experiment was done on real dataset mention in 5.2 and comparison of proposed models were done on different dataset size with existing model **PASOWPR** proposed in [38]. Comparing models are categorize as per feature vector as well. So variation of genetic algorithm with feature set combination were also shown in this section of the thesis.

Table 2 Coverage based comparison of web page prediction models.

Dataset Percentage	PASOWPR	TLWPP
30	0.1263	0.3010
40	0.1465	0.2230
50	0.1575	0.1964
60	0.1692	0.1687
70	0.1589	0.1677

Table 5.6 shows coverage values of page prediction models, it was obtained that PASOWPR has lower coverage value as compared to proposed models TLWPP. Use of weblog and web content features in the work has increases the work efficiency of proposed genetic based models. It was obtained that proposed models TLWPP has highest coverage value in all dataset percentage. So coverage average percentage enhancement done by TLWPP is % as compared to PASOWPR.

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

Dataset Percentage	PASOWPR	TLWPP
30	0.2016	0.4007
40	0.2340	0.2970
50	0.2516	0.2616
60	0.2704	0.2248
70	0.2544	0.2835

Table 3 shows that proposed model TLWPP has increases the M-metric value of the work by % as compared to PASOWPR model. It was also shown from table 3 that TLWPP m-metric value was always

high in all set of dataset percentage as compared to PASOWPR. Fig. 3 shows that TLWPP proposed genetic based model has high M-metric value always above 0.22, this efficiency of correct page prediction was achieved by two phase learning.

Table 4 MAE based comparison of web page prediction models.

Dataset Percentage	PASOWPR	TLWPP
30	**0.0695**	**0.0440**
40	**0.0522**	**0.0517**
50	**0.0502**	**0.0404**
60	**0.0470**	**0.0343**
70	**0.0704**	**0.0299**

Table 4 shows MAE values of page prediction models, it was obtained that PASOWPR has higher MAE value as compared to proposed models TLWPP. Use of weblog and web content features in the work has increases the work efficiency of proposed genetic based models. TLWPP has reduced the MAE value by 30.76% as compared to PASOWPR.

Table 5 RMSE based comparison of web page prediction models.

Dataset Percentage	PASOWPR	TLWPP
30	**0.2637**	**0.2097**
40	**0.2284**	**0.2273**
50	**0.2241**	**0.2010**
60	**0.2169**	**0.1851**
70	**0.2654**	**0.1729**

Table 6 Execution time based web page prediction models comparison.

Dataset Percentage	PASOWPR	TLWPP
30	**11.4074**	**2.6321**

40	**15.4247**	**3.2406**
50	**20.2031**	**2.9260**
60	**22.2819**	**3.0729**
70	**25.9928**	**3.8502**

Table 5 and 6 shows that proposed model TLWPP has reduces the RMSE value of the work by % as compared to PASOWPR model. Table 5 shows that TLWPP proposed genetic based model has low RMSE value this was achieved by merged PDM feature set.

V. Conclusions

Website rank on search engine depends on users retention, so suggesting relevant page for the user increase its weight for others. Paper has detailed teach leaning based optimization algorithm for web page prediction. Use of hybrid probability density matrix feature obtained from web log and web content in the fitness function evaluation increases the chance of predicting a more desired page. In order to increase the efficiency of work in less time with high precision model has include genetic algorithms. Experiment was done on real live site weblog and content dataset. Result shows that Proposed model has improved the TLWPP coverage value by %, while reduce MAE by % and RMSE value by % as compared to PASOWPR. In future scholars can perform same work on other language sites with different features set.

References

[1] N. Ahmad, O. Malik, M. ul Hassan, M. S. Qureshi and A. Munir, "Reducing user latency in web prefetching using integrated techniques," *International Conference on Computer Networks and Information Technology*, 2011.

[2] Soundharya V , Ram kumar R, Prakash B, Sowndarya B, Prathiksha B. "A Survey on Pattern Discovery of Web Usage Mining". International Journal of Research in Engineering, Science and Management Volume-1, Issue-8, August 2018.

[3] Pabarskaite, Zidrina. Decision trees for web log mining. Intell. Data Anal.. 7. 1 -2003-7205.

[4] Kolan A., Moukthika D., Sreevani K.S.S., Jayasree H. (2020) Click-Through Rate Prediction Using Decision Tree. Proceedings of the Third International Conference on Computational Intelligence and Informatics. Advances in Intelligent Systems and Computing, vol 1090. Springer, Singapore

[5] Dr R.S. Vetrivel, Manju.J, P Jeyanthi Rani. "Visual Intensive Web Data Extraction Using Markov Chain Classifier For Web Document Categorization". International Journal of Advanced Research in Computer Science & Technology (IJARCST 2017) 64 Vol. 5, Issue 1, Jan. - Mar. 2017.

[6] Chaipornkaew P., Banditwattanawong T. A Recommendation Model Based on User Behaviors on Commercial Websites Using TF-IDF, KMeans, and Apriori Algorithms. In: Meesad P., Sodsee D.S., Jitsakul W., Tangwannawit S. (eds) Recent Advances in Information and Communication Technology 2021.

[7] Bhavithra, J., Saradha, A. Personalized web page recommendation using case-based clustering and weighted association rule mining. Cluster Comput 22, 2019.

[8] Saradha, A., Aiswarya, J.: An improved mechanism for user profiling and recommendation using case-based reasoning. IIOAB J. 8(2), 2017.

[9] Guoguang Liu" An ecommerce recommendation algorithm based on link prediction", Alexandria Engineering Journal, 2021.

[10] Fatma Mlika, Wafa Karoui. "Proposed Model to Intelligent Recommendation System based on Markov Chains and Grouping of Genres", Procedia Computer Science, Volume 176, 2020.

[11] 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 10 January 2019.

[12] Ashwani Mathur "Hybrid Combination of Error Back Propagation and Genetic Algorithm for Text Document Clustering" International Journal of Computer Trends and Technology 68.11 (2020):64-68.

[13] M. M. Puralachetty, V. K. Pamula, L. M. Gondela and V. N. B. Akula, "Teaching-learning-based optimization with two-stage initialization," *2016 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2016

[14] R. Venkata Rao, Vivek Patel, An improved teaching-learning-based optimization algorithm for solving unconstrained optimization problems, Scientia Iranica, Volume 20, Issue 3, 2013,

[15] Nomaan Jaweed Mohammed. Neural Network Training by Selected Fish Schooling Genetic Algorithm Feature for Intrusion Detection. International Journal of Computer Applications 175(30):7-11, November 2020.