Trivene Khede, 2021, 7:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

Bio-Geography Based Page Prediction Using Web Mining Feature

Trivene Khede, Dr. Avinash Sharma
Dept. of CSE
Millennium Institute of Technology &Science Bhopal
MP, India

Abstract: Website is god place to reach the audience of any field. Many of companies are using these platform for different business. Retaining a web visitor on website depends on available content and intelligence of site. This paper has developed a intelligent model that can predict the web page by understanding the behavior of the user. Biogeography optimization genetic algorithm was used to predict the web page as per past user visits. This work uses web content and web log feature of the website for evaluating the fitness value of genetic algorithm chromosomes. Experiment was done on real dataset with different size. Result shows that proposed model has improved values of different evaluation parameters.

Index Terms- Association Rule, Page Prediction, Genetic Algorithm, Information Extraction, web mining.

I. INTRODUCTION

As the web users are growing day by day, the need of the networking world is becoming moderately high. So as to expand the clearness and promptness within the work great amount of labor depends on this web network [1]. This attracts several researchers for raising the performance of the network and lessens the latency time of the web, so things get less complicated and quick for the daily consumers.

At this point hardware element is the means of optimizing the network however in parallel computer code additionally ought to be updated. This paper concentrates on optimizing the network power by learning the user actions for reducing the latency time of looking out the desired matter of specific interest. As websites are important supply of knowledge for pretty much all necessities, thus these necessities of individuals attract variety of individuals to produce varied services. However targeting the right client is basic demand of the service or business [2]. Analysis during this space has the purpose of serving to e-commerce businesses in their choices, helping within the style of fine websites and helping the user whereas navigating the net. Even despite the fact that these days net users have created higher are broadly put into practice currently these days since they win vital latency savings. Several transnational firms implement net replication by victimisation Content Delivery Networks [4] to lessen their websites access time however this resolution isn't possible because it is pricey and lots of small firms, organizations cannot afford it. Net pre-fetching techniques are reciprocally freelance to caching and replication techniques, so they will be applied along to attain a more robust net performance. Caching and replication techniques are widely enforced in globe; some studies have additionally investigated net pre-fetching in real environments.

Web pre-fetching is utilized to pre-process aim requests before the user makes a specific demand of these objects; so as to lessen the users' observes latency [4, 5]. Net pre-fetching consists of 2 steps mainly; initial, it's a necessity to create correct prediction then next user accesses. These predictions are sometimes created support to previous expertise concerning users' accesses and fondness, and also the relative hints are provided to a pre-fetching device, second, the pre-fetching device makes a conclusion that objects from the anticipated hints are reaching to be pre-fetched.

II.RELATED WORK

Hai Dong, FarookhHussain and Elizabeth Chang in [6] projected net query Classification technique that depends on net distance normalisation. During this design middle categorised queries are send to the target category by normalizing and mapping the net queries. By explaining the frequency, position and position frequency classes are stratified into 3 categories. Within the system Taxonomy-Bridging rule is employed to map target class. The Open Directory Project (ODP) is utilized to create Associate in Nursing ODP-based classifier. This taxonomy is then mapped to the target class's victimization Taxonomy-Bridging rule. Thus, the post-retrieval question is initially classified into the ODP taxonomy, and also the classifications are then mapped into the target classes for net query. Classification of net query to the user intends and question is major duty for any data retrieval system.

Myo MyoThan Naing [7] projected "Query Classification Algorithm". To classify the net query input by the user into the user intended categories, MyoMyoThanNaing utilizes the domain ontology. Ontology is beneficial in matching of retrieve group to focus on group. User questions are extracted in Domain terms are used as input to the question classification rule. Matched terms of every domain term are extracted in more sub division. Reason the likelihood for matched classes. Then all queries are stratified by their likelihood and displays to the user's table.

In [8] web content re-ranking is prepared by the utilisation of net log feature alone. Here multi-damping of the user series is prepared on the premise of linear, page rank, generalized hyperbolic features are used. As this paper has not embrace website feature thus accuracy level is low.

In [9] this analysis work, web content health recommender systems are introduced via the exercise of bound agents so as to produce very acceptable web content for patients. The essential feature of Particle Agent Swarm Optimization (PASO) is that the creation of the rule is denoted by a group of Particle agents World Health Organization join forces achieve the target of the task into account. Within the analysis technique, 2 forms of agents are presented: net user particle agent and linguistics particle agent. PASO primarily based web content Recommendation (PASO-WPR) system is Associate in Nursing intermediate program (or a particle agent) containing a programme, that sagely produces a set of information that suits Associate in Nursing individual's necessities. PASO-WPR has carried out dependent upon incorporating linguistics data with data processing techniques on the net usage information still as clump of pages dependent upon similarity in their linguistics. Because the web content with transmission files are viewed as ontology people, the pattern of patients' routing are like instances of ontology instead of the uniform supply locators, and with the assistance of linguistics similarity, page clump is carried out.

Frikhaetal.[10] concentrates on enhancing typical recommender systems through means of group action data in social media; consist of the first choice of a user still as influence from his peers who are on-line. So, ontology is based on the interest of the user is created to form recommendations that are customized. Symentic social recommender system is projected for applying user interest ontology still as Tunisian Medical Tourism (TMT) ontology. Finally, social recommendation rule is developed with the purpose to be used during a Tunisia tourism Website to aid users to attract in visiting Tunisia for medical reasons.

Gulzar et al. [11] projected a recommender system, which might advise as well as the user in selecting the courses based on his want. The Hybrid technique was used within the company of ontology to require out valuable data and create accurate recommendations. These strategies are helpful to learners to boost their performance still as progress their satisfaction level. The given recommender systems can do higher by assuaging the constraints of basic individual recommender systems.

III.PROPOSED METHOD

Explanation of proposed Bio-Geography Optimization Based Web page Prediction (BOWPP) model was done in this section of paper. To increase understanding of the work fig. 1 shows steps of different BOWPP sections. Detail explanation of each block of BOWPP is given below.

1. Pre-Processing- Web mining two features were extract for the page prediction, first weblog and other is web content.

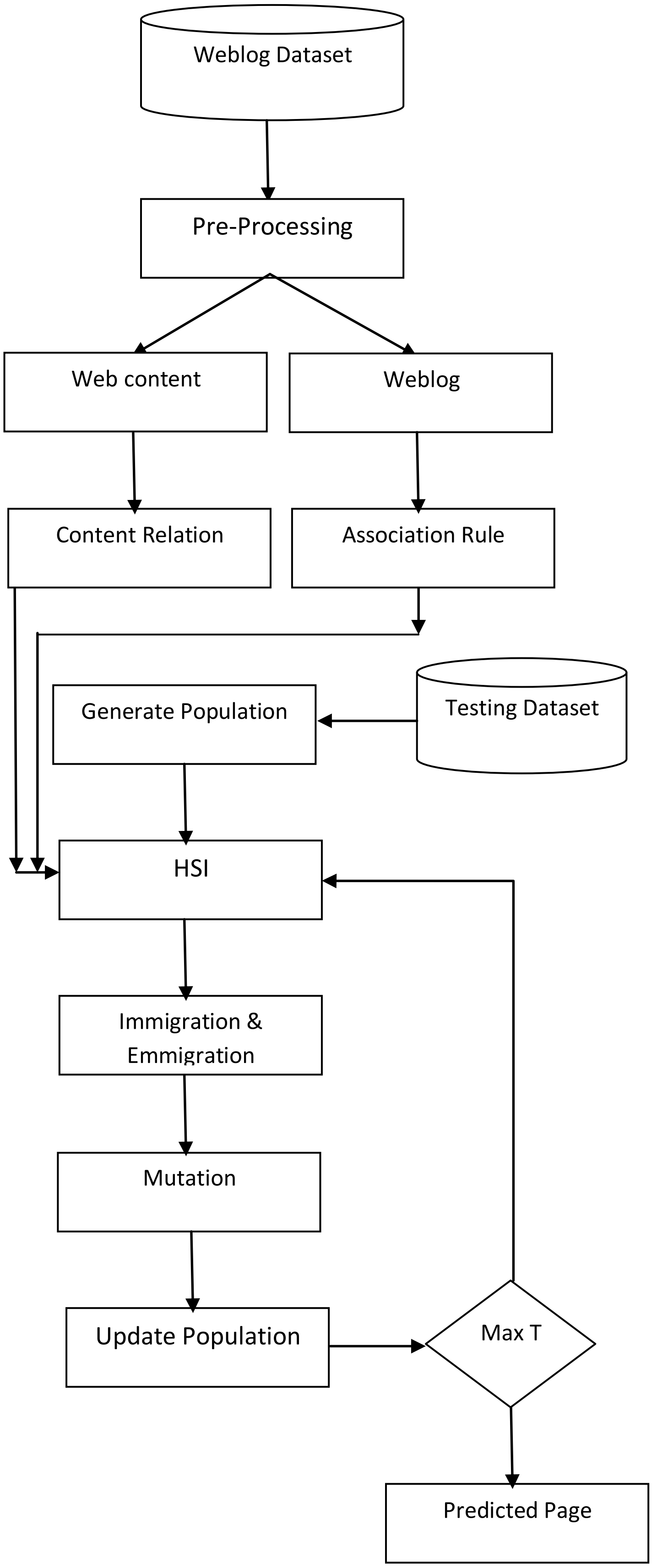

Fig. 1 Bloch diagram of BOWPP.

2. Web content- Keywords are generate from the URL of weblog. As most of url in these days are keyword oriented as per search engine optimization rules. Although some pre-processing is required for the URL that is to remove the unwanted words or the stop words from the URL. So a stopword dictionary was used to filter content and update in W a matrix. Weblog: As per W matrix that is collection of page keywords each page get a unique ID. So a set of pages visit by a user in single span of weblog dataset is consider its weblog pattern. Similarly other weblog patterns generate from the weblog dataset by identifying page and its corresponding user IP address.

Content Relation: Web content was used for finding the relation between the pages in form of bonds. Page having good set of common keywords are strong bond mostly 2 or 3 common words. But page have 0 or 1 common word are weak connected pages. This relation help genetic algorithm to dicide which page has more chance to open after a sequence of previous page views.

Weblog Association rule: In order to extract information from the dataset association rules were generate form the weblog dataset. Three element association urles were generate where rule looks like A, B →C. Some set of rules were removed the pattern that have very low support value below 1%, as these rules are noise in the information of rule mining.

Testing Phase: Here the dataset is again preprocess for the web log portion in order to get the logs that are use for testing the built model. Pre-Processing steps are similar as done in previous steps of model. The only difference here is that pre-processed logs are break such that each sequence first few pages are in the testing part and the next page after that sequence is store for the evaluation of result.

Generate Population: Here assume some possible solution set that are the combination of the possible pages as per generated rules having visited web pages elements. This was developed by the random function shown in eq. 1. This can be understand as let the number of possible set be n and number of initial chromosome is IP, then one of the possible chromosome solution is H ={P_1, P_5,….P_m} this can be assume as the solution set [14].

$$H \leftarrow (m, n, ARS)\text{----Eq. 1}$$

Habitat Suitability Index (HSI): This is term as fitness value of the habitat, means higher value shows that poor place to live while low value means good place to live in terms of resources, life, etc.

Immigration and Emigration Rate Some of basic terms of immigration λ and Emigration α was done by Eq. 1, 2 [15]:

$$\lambda_R = (1 - \alpha_R)\text{-------------Eq. 1}$$

$$\alpha_R = \frac{R}{h}\text{-------------Eq. 2}$$

Where R is rank of habitat in terms of HSI value, while h is total number of habitats.

Fitness Function (HSI): Habitat suitability Index of any habitat depends on the distance. So its an summation of support value of rules obtained from the weblog feature and relation value between possible pages. Chromosome having high summation value have good chance to predict desired page.

$$F_h = \sum_{x=1}^{n} Support(H, V) + \sum_{y=1}^{n} Relation(H, H) \text{ ------Eq.2}$$

Hence rank i of the habitat depend on the $F_h$value.

R = Rank(F_h, H)----Eq. 3

Crossover

Emigration of page in form of species from one habitat to other is depend on emigration rate. While permitting species to enter in a habitat is depend on immigration rate. Hence for crossover from one habitat to other both type of rate need to find. So crossover depends on following condition.

```
Loop x=1:h
    If Cross_Over_Limit > λ_R
        Loop y=1:h
            If Cross_Over_Limit > α_R
                M←Rand()
                H[x, m]← H[y, m]
            EndIf
        EndLoop
    EndIf
EndLoop
```

Where Cross_Over_Limit is random number range between 0-1, x and y is habitat position specify immigration, emigration operation.

Mutation

In this work after crossover mutation was also perform so chance of new solution get increases. For this paper has involved mutation probability where as per HSI value mutation was performs in selected habitats.

$$M_h = \frac{R_h}{sum\,(h)}\text{-------------Eq. 6}$$

$$M_p = \frac{M_h}{Max\,(M_h)}\text{ -------------Eq. 7}$$

Hence habitat which cross a constant Mutation_Cross_Limit range in 0-1, M_h gives an mutation rank for the habitat as per HSI value. So higher value have higher mutation rank. Hence those habitats which have higher mutation rank have higher mutation probability. So habitat which has lower Mutation Probability as compared to Mutation_Cross_Threshold undergoes to mutation.

Final Solution

In this work after sufficient number of iteration best possible chromosome obtained and set of those pages are recommended pages for the proposed model of genetic algorithm.

IV.EXPERIMENTS AND RESULTS

All calculations and utility measures were executed utilizing the MATLAB platform. The tests were performed on a 2.27 GHz Intel Core i5 machine, furnished with 8 GB of RAM, and running under Windows 10 operating system.

Results:

Table 1. Web page prediction comparison b Precision value.

Dataset Percentage	BOWPP	Previous Model [16]
30	0.8462	0.4196
40	0.7592	0.4031
50	0.7143	0.3824
60	0.6643	0.3706
70	0.6847	0.3694

Table 1 shows that proposed BOWPP model ha increases the web page precision evaluation parameter value. T was found that proposed model has improved page accuracy by use of crossover and mutation operation both in one iteration.

Table 2. Web page prediction comparison by coverage value.

Dataset Percentage	BOWPP	Previous Model [16]
30	0.4261	0.0853
40	0.3816	0.0777
50	0.3586	0.0715
60	0.3333	0.0679
70	0.3434	0.0667

Coverage values present in table 2 shows that proposed BOWPP model has increases the value as compared to model proposed in [16]. Use of Biogeography fitness function where content and web log feature combination increases the user page prediction coverage parameter.

Table 3. Web page prediction comparison by M-Metric value.

Dataset Percentage	BOWPP	Previous Model [16]
30	0.5667	0.1418
40	0.5079	0.1303
50	0.4775	0.1205
60	0.4439	0.1148
70	0.4574	0.1131

M-Metric values present in table 2 shows that proposed BOWPP model has increases the value as compared to model proposed in [16]. Use of Biogeography fitness function where content and web log feature combination increases the user page prediction M-Metric parameter.

Table 4. Web page prediction comparison by time (Seconds) value.

Dataset Percentage	BOWPP	Previous Model [16]
30	4.2096	18.4969
40	4.6184	25.8503
50	5.1629	31.3389
60	6.0562	35.1036
70	7.4281	41.875

Table 4 shows that proposed BOWPP model has reduces the web page prediction time evaluation parameter value. It was found that proposed model has improved page precision value in less time by use of content and web log features for estimating the fitness value of the biogeography chromosomes.

I. CONCLUSIONS

Web content increases day by day so attracting users and retaining on site is getting tough. In order to increase some intelligence in the web portal researcher work on web page prediction. This paper has proposed a BOWPP model that efficiently utilize different features of the website logs and content. Based on association rule and webpage relation page prediction done by BOWPP model. Experiment was done on real website dataset with different dataset size. Result shows that proposed model has increases the precision value by % as compared to model proposed in [16[. In future scholars can train a mathematical model to get more accurate value in less time.

REFERENCE

[1]. An Ontology-based Webpage Classification Approach for the Knowledge Grid Environment by Hai Dong, Farookh Hussain and Elizabeth Chang, 2009 Fifth International Conference on Semantics, Knowledge and Grid (IEEE-2009).

[2]. Fan Jiang Carson K. Leung, Adam G. M. Pazdor." Web Page Recommendation Based on Bitwise Frequent Pattern Mining". 2016 IEEE/WIC/ACM International Conference on Web Intelligence.

[3]. Mohammad Amir Sharif, Vijay V. Raghavan. "A Clustering Based Scalable Hybrid Approach for Web Page Recommendation". 2014 IEEE International Conference on Big Data.

[4]. Dr. V. Sujatha1 , Dr. M. Punithavalli and Dr. Renjit Jeba Thangaiah. "Classifier and Clustering for Web Page Prediction". International Journal of Scientific & Engineering Research Volume 8, Issue 9, September, 2017.

[5]. Neha V. Patil , Dr. Hitendra D.Patil. "Prediction of Web User's Browsing Behavior using All Kth Markov model and CSB-mine". International Journal of Computer Trends and Technology (IJCTT) – Volume 43 Number 1 – January 2017 ISSN: 2231-2803.

[6]. Ms. Dhaslima Nasrin.S , Ms. Mubina. A, Mrs.K.Shanmuga priya. New Event Detection For Web Page Reccomendation Using Web Mining. International Journal of Advanced Research in

International Journal of Science, Engineering and Technology

An Open Access Journal

Computer Science Engineering and Information Technology Volume: 6 Issue: 3 Mar,2017.

[7]. Zhao, Y., Karypis, G. 2002. Evaluation Of Hierarchical Clustering Algorithms For Query Datasets, Acm Press, 16:515-524.

[8]. San San Tint1 And May Yi Aung. "Web Graph Clustering Using Hyperlink Structure ".Advanced Computational Intelligence: An International Journal (Acii), Vol.1, No.2, October 2014

[9]. Khan, M. S., & Khor, S. W. (2004). Web Query Clustering Using A Hybrid Neural Network. Applied Soft Computing, 4(4), 423-432. 17

[10]. Kleinberg, J. 1997. " Web Usage Mining For Enhancing Search Result Delivery And Helping Users To Find Interesting Web Content",‖ Acm Sigir Conf. Research And Development In Information Retrival (Sigir '13), Pp. 765-769,2013.

[11]. Mamoun A. Awad And Issa Khalil "Prediction Of User's Web-Browsing Behavior: Application Of Markov Model". Ieee Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.

[12]. Thi Thanh Sang Nguyen, Hai Yan Lu, Jie Lu " Web-Page Recommendation Based On Web Usage And Domain Knowledge" 1041-4347/13/$31.00 © 2013 IEEE.

[13]. Zhen Liao, Yang Song, Yalou Huang, Li-Wei He, And Qi He. "Task Trail: An Effective Segmentation Of User Search Behavior" . Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 12, December 2014.

[14]. Simon D. Biogeography-based optimization. IEEE Transactions on Evolutionary Computation. 2008;12(6):702–713.

[15]. Mohammed Alweshah. "Construction biogeography-based optimization algorithm for solving classification problems". Neural Computing and Applications, Springer volume 28 February 2018

[16]. 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.