ISSN 2320 - 2602
Volume 3, No.2, February 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst03322014.pdf

AN EFFICIENT WEB LEARNING WEB TEXT FEATURE EXTRACTION WITH EXPONENTIAL PARTICLE SWARM OPTIMIZATION

R.Dhanya [1], Mrs.P.S.AnnaKKodi [2]
[1]M.Phil Scholar, Department of Computer Science, Sri Ramalinga Sowdambigai College of Science and Commerce, Vadavalli, Coimbatore-641109, Tamilnadu, India, dhanayarav@gmail.com
[2]Head of the Department , Department of Information Technology , Sri Ramalinga Sowdambigai College of Science and Commerce , Vadavalli, Coimbatore-641109, Tamilnadu, India, annakkodi.ps @gmail.com

ABSTRACT

Due to the growth of World Wide Web the tradition of web patterns also enhanced now a days, extraction of information from web also important. For this reason web mining plays important role to discovery of individual user information and extract information from individual web log files with known text feature. Due to its extensive division, its directness and elevated dynamics, the resources occurrence the web are significantly sprinkled and they comprise no incorporated administration and arrangement. It seriously decreases the effectiveness by means of web information. Because finding text feature most important imperative problem in web mining. To conquer this problem proposed an efficient method to extract the log data to learn user profiles using Back propogation Neural Network (BPNN) and extract web text feature using Exponential Particle Swarm Optimization (EPSO). The proposed representation uses a Back propogation Neural Network (BPNN) structural design with a back propagation knowledge method to determine and investigate helpful information from the obtainable Web log data file then feature extraction is performed. Second the Web transcript feature Extraction procedure is wished-for best feature extraction and finally compares the results. In the present effort develop the best knowledge ability consequence for web log data and decrease the computation strength of an aggressive knowledge BPNN and the EPSO algorithm. Experimental results shows that the enhanced BPNN and EPSO schema extract best text features for web log information for every user.

Keywords: Back propogation Neural Network (BPNN), Exponential Particle Swarm Optimization (EPSO),Web log data ,Text Feature extraction , Web mining.

1. INTRODUCTION

Web mining is second-hand to determine the sample starting the WWW. The web contented mining extensive description on the individual offer explain the repeated exploration and recovery of information and assets accessible beginning millions of sites and on-line records and on the other hand the web usage mining provides the discovery and examination of user admission patterns beginning individual or further on-line services.

Web includes formless information. Consequently semantic comprehensible nonexistence of mechanism, natural language accepting and transcript dispensation are the major technique used in the playing field of Web mining. Web Mining is an extremely significant regulation of data mining and is illustration enormous attention beginning academic world and software business. The WWW serves as enormous, extensively dispersed, worldwide information examine interior for information, announcement, customer information, economic administration, learning, administration and numerous additional information services [1][2].

The WWW is increasing enormously in the numeral of websites and too in the populace of consumer. This self-motivated compilation of information gives rich basis for data mining. Web Mining is a difficult task that exploration intended for Web structures and the reliability and dynamics of Web contents. The web provides an enormous quantity of data existing as dispersed information which are in formless structure. This information that is extorting starting from the web requirements to be cleaned, prearranged, and maintains in organize to allow a well-organized utilize [3].

How to explore used for the information beginning the enormous web information quickly and precisely has grown to be a most important difficulty [4]. There was an urgent necessitate of tools that be capable of rapidly and successfully discover assets and information starting the Web. The major structure of information on the web is web transcript. Consequently how to procedure these Web Texts happen to the explanation difficulty. Data mining can assist determining possible information and in order from mass unprocessed information and successfully resolve the difficulty that data is abundant but information is missing. Though, the majority of traditional Data Mining tools need ordered information. So, Web-Based Text Mining has turn into a novel topic of Data Mining.

In this paper, an examination of Web mining is obtainable exactness the methods of Web mining and information dispensation. A novel Web text feature Extraction technique is anticipated that an EPSO algorithm is use in it. Believe to the text features assortment can be altered into the optimization procedure in the multi-dimensional data space. The proposed representation uses a Back propogation Neural Network (BPNN) structural design with a back propagation knowledge method to determine and investigate helpful information from the obtainable Web log data file. Secondly,

the Web text be supposed to be implied and the text vector is examination as best probability learning result for particles. Since the amount of the text features is unidentified, proposed a new Exponential Particle Swarm Optimization (EPSO) among adjustable measurement. Overturn thoughts particles are also additionally added in the algorithm in order to get better the international search aptitude of the EPSO.

2. RELATED WORK

Improvement of effectual method for Web personalization is a significant part of investigation through numerous instant appliances for Web information arrangement. Appreciative the users' navigational inclination and performance is a necessary step in learning the usefulness of a Web site. For instance, the detection of the majority probable admission patterns authorize e-business contributor to assess and develop the superiority of a site formation by adaptation the hyperlinks converting into a Web page for conducting the consumer to motivating information.

Web usage mining [5] is an extensively second-hand move toward to confine and representation of Web user behavioral sample beginning the log information created by Web and appliance server. A variety of Web usage mining method has been second-hand to grow well-organized and successful recommendation scheme to afford individualized contented to user support on their partiality and precedent behavior. For instance, association rule mining has lately been premeditated as a move toward to determine models for suggestion scheme [6-8]. It is a non-sequential mining method to do not conserve the ordering information amongst page views in consumer gathering.

Sequential pattern mining and the discovery of frequent navigational path obtain into explanation the ordering constraint inherent in navigational patterns. The employ of navigational and sequential patterns for analytical user representation have been expansively considered [9]. Conversely, the most important focal point of these studies has been on wonderful of Web pages to progress server presentation. In a unified prescribed structure is accessible to confine a variety of navigational substructure in the tradition information, together with frequent itemsets and a widespread move toward to personalization is anticipated by means of navigational passageway fragments [2-10].

The general performance of the PSO whilst Gaussian distributed arbitrary noise was additional to the fitness function and rotation of investigation search space with randomly performed [11]. The experimental results shows that the performance of PSO remained successful with presence of noise, and, in a number of belongings, noise still help out the PSO keep away from individual attentive in neighboring optima.

In the investigation the PSO was evaluate to a noise-resistant variant wherever the major PSO round was adapted so that numerous assessment of the similar applicant explanation are comprehensive to improve evaluate the definite fitness of this exacting clarification. The comparison well thought-out a number of arithmetical problems by means of added noise, as well as unsupervised knowledge of obstruction prevention by means of single or additional robots. The noise Resistant EPSO showed significantly improved performance than the unique [12].

3. BACKPROPOGATION NEURAL NETWORK (BPNN) LEARNING METHOD AND EXPONENTIAL PARTICLE SWARM OPTIMIZATION (EPSO) BASED TEXT FEATURE EXTRACTION

Web personalization extract the characteristic consumer profiles beginning the huge quantity of past data stored in admission logs. The generally procedure of Web personalization normally includes of three phases:

1. Data training and conversion
2. Pattern discovery
3. Recommendation

In conventional collaborative filtering methods, the pattern detection stage as well as the suggestion part is achieved in real instance. In compare, personalization scheme supported on Web usage mining [13], achieve the pattern detection stage offline. Data training phase transforms unrefined web record files addicted to connect stream data with the purpose of is able to be procedure by data mining responsibilities. The personalized contented be able to acquire the structure of suggested associations to the user's supposed preferences as strong-minded by the corresponding procedure patterns. In this paper, our focal point is particularly on association rule mining and the appropriateness of the resultant patterns designed for personalization.

Web mining is the nontrivial procedure to determine applicable, potentially functional information beginning web information by means of the data mining procedures. It might provide information with the purpose is functional for civilizing the services presented by web threshold and information admission and recovery apparatus. With the rapid expansion of further researchers includes the clustering procedure to dissimilar ground in modern days. When clustering move toward is functional to the web practice data it mechanically confines the concealed browsing patterns beginning it in the structure clusters.

Back-propagation (BP) representation is the majority well-liked in the supervised knowledge structural design since of the weight error approved regulations. It is measured a simplification of the delta regulation for nonlinear establishment functions and multilayer schema. In a back-propagation neural system, the knowledge steps have two types of phases. Primary, a preparation key pattern is accessible to the system input level. The network proliferate the input example beginning layer to layer in anticipation of the yield pattern is produced by the output layer. If this pattern is dissimilar beginning the preferred output, errors are calculated and then proliferate backward all the way through the network beginning the output layer to the contribution layer. The weights are adapted as the error is

circulated. The back-propagation preparation algorithm is an iterative ascent calculated to reduce the mean square error among the concrete yield of multi-layer feed forward observation and the preferred output. It necessitate constant dissimilar non-linearity

Step1: Initialize the weights of user log file for web user and offsets position every one weights and node offsets to little random principles.

Step2: current input and preferred outputs there are continuous respected input vector $X_0, X_1, \dots\dots, X_{N-1}$ and indicate the preferred output vector data for log user file $d_0, d_1, \dots\dots, d_{M-1}$.

If the network is second-hand as a classifier those every one of preferred outputs are normally set to zero excepting for with the aim of equivalent to the group the input is beginning. That preferred output is 1. The input might be original on every we log file beginning a training set might be obtainable regularly pending steady.

Step 3: estimate definite Output exercise the sigmoid non linearity beginning higher than and formulas as to estimate output $Y_0, Y_1, \dots\dots, Y_{M-1}$.

Step 4: Familiarize yourself weights employ a recursive algorithm preliminary at the output nodes and effective back to the primary hidden coating. Regulate weights by

$$W_{ij}(t+1) = w_{ij}(t) + \eta\delta_j x_i^{'}$$

Within this equation $w_{ij}(t)$ is the weight beginning hidden node i to node j is considered as one log file to another log file during the time t , $w_j^{'}$ is moreover the output of node i as input , η is gain period and δ_j is an error time for node log file j ,if node j is an output node of the log file then ,

$$\delta_j = y_j(1-y_j)(d_j - y_j)$$

Where $d_j$the preferred output log is file of node j and y_j is the definite output . If node j is one log file to another log file data is an interior concealed node then

$$\delta_j = x_j^{'}(1-x_j^{'})\sum_k \delta_j^m w_{jk}$$

Where k is in excess of every one node in the layers beyond the node j .

Internal node threshold are modified in a comparable approach by high and mighty they are association weights on links beginning supplementary steady–valued inputs .Convergence is occasionally earlier if a force expression is supplementary and weight modify are rounded by

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j x_i^{'} + \alpha\left(w_{ij}(t) - w_{ij}(t-1)\right), where\ 0 < \alpha < 1$$

Step 5:Repeat and go to step 2 until all the web log data files are learned from Back propogation neural network

Particle swarm optimization optimizes a difficulty by pleasing into description of the position and velocity. It is guided in the direction of the most excellent be familiar with location in the exploration space. This is predictable to give the greatest resolution. It can be used to investigate large investigate space. EPSO (Extended Particle Swarm Optimization) can be mutual among the further intelligent optimization technique to propose numerous of composite optimization systems. EPSO can be moreover lead into distribution scheme to expand EPSO's purpose assortment. EPSO is the appropriate method for systematize the problems in the web pages collection intended for supporting the user navigation all the way through the web pages. It can also be presented with the large amount of data. Back propagation progress the global explores capability of the EPSO. Particle position and the velocity give the greatest fitness attained.

PSO was stimulated by the communal behavior of a bird assemble. The representation of these bird assemble is named as particles. These particles can be measured as straightforward agents all the way through a difficulty freedom. A particle's position in the solution search space defines the solution for each and every problem .If the particles moves from one solution search space to another solution search space the new solutions are found for each and every particles .The solution of each particle is estimated based on fitness function of the every particle with the intention of offer a qualitative result assessment of the solution usefulness [14], [15]. Each particle is considered as text feature $\mathrm{TF_d}$ preserves the subsequent information: $\mathrm{TF_d}$ the present location of the text feature particle , $v_i$the present velocity of the text feature particle should be definite by parameters v_{min} andv_{max} . The individual text feature particle most excellent location of the text feature particle is characterized by y_i.

So the particle's location is familiar according to

$$v_{tf,k}(t+1) = wv_{tf,k}(t) + c_1 r_{1,k}(t)\left(y_{tf,k}(t) - x_{tf,k}(t)\right) + c_1 r_{1,k}(t)\left(\hat{y}_k(t) - x_{tf,k}(t)\right) \rightarrow (1)$$

$$x_{tf}(t+1) = x_{tf}(t) + v_{tf}(t+1) \rightarrow (2)$$

Where w is the inertia weight whose series is [0-1], c_1& c_2 are the knowledge factors called, correspondingly, cognitive factor and social factor, $r_{1,tf}(t), r_{2,tf}(t) \sim \mathrm{U}(0,1)$ and $k = 1, \dots, N_d$.

$$y_{tf}(t+1) = \begin{cases} y_{tf}(t)\ if\ f(x_{tf}(t+1)) \geq f(y_{tf}(t)) \\ x_{tf}(t+1)\ if\ f\left(x_{tf}(t+1)\right) < f\left(y_{tf}(t)\right) \end{cases} \rightarrow (3)$$

$$c_2 r_{2,k}(t)\left(\hat{y}_{tf,k}(t) - x_{tf,k}(t)\right) \rightarrow (4)$$

Where $\hat{y}_{tf,k}(t)$ is the most excellent text feature particle in the locality of the $\mathrm{TF_d}^{\mathrm{th}}$ particle. The PSO is implemented with frequent submission of the equation (1), (2) for each and every text feature until a précised amount of steps has been exceeded and equal to the velocity of the each and every text features are close to zero over a number of iterations. This communal collaboration helps them to determine moderately good solutions quickly. Though, it is accurately these immediate communal relationships that formulate text feature particles idle on neighboring optima and not succeed to come together at overall best possible. Once a novel g_{best} is established, it spreads in excess of text feature particles instantaneously and so all text feature particles are concerned to this location in the following steps

until one more enhanced resolution is found. Consequently, the stagnation of PSO is basis by the generally speed distribution of recently found g_{best}.

An development to unique PSO is represented by the information with the intention of *w* is not reserved steady during implementation; rather, preliminary from a maximal significance, it is linearly decremented as the numeral of steps raise downward to a minimum significance

$$w = (w - 0.4)\left(\frac{MAXITER - ITERTATION}{MAXITER}\right) + 0.4 \rightarrow (5)$$

$MAXITER$ is defined as the maximum number of iterations to complete the text feature extraction result for learning web data log file result from BPNN and $ITERATION$ characterizes the number of iterations to extract best text features . EPSO has a enormous result on global and local examination it is hypothetical to get out the investigate performance rapidly and wisely as it keep away from the text feature particles beginning stagnation of local text feature extraction result by varying this internal weight exponentially from equation (6) consequently with the intention of the association of the text feature particles determination be further earlier and far-away beginning every other

$$w = (w - 0.4)e^{\left(\frac{MAXITER-ITERTATION}{MAXITER}\right)^{-1}} + 0.4 \rightarrow (6)$$

Evaluation task is an significant principle for evaluator the qualities of text feature particle position. The location of every text feature particle is collected of the feature weight of the equivalent documents are obtained. So the best text feature results must be capable to reproduce the web document fine and it moreover must contain additional text features distribution. Consequently, the individual's fitness necessity be calculated by every one text features which current the similar document. The further comparable they are, the superior their condition must exist .The fitness function be supposed to be characterized as subsequent:

$$fitness\left(P_{tf}^{d}\right) = \sum_{j=1}^{swarm\_size} \frac{similar\ (P_{tf}^{d}, P_{tf1}^{d})}{swarm_{size}} \rightarrow (7)$$

The similarity of BPNN learning result can be calculated by the direction among two text feature learning results particles using the following formula

$$similar(P, Q) = \frac{\sum_{tf=1}^{d} p_{tf} \times q_{tf}}{\sqrt{\sum_{tf=1}^{d} p_{tf}^{2}} \times \sqrt{\sum_{i=1}^{d} q_{tf}^{2}}}$$

p, q is the elements of learning results P ,Q. In adding up, the learning result of web log data is also an important condition. If several particles have the same fitness, smaller the learning web log data file result then select the most important text features are extracted then page view results are found in the experimental results

1. Initialize every particle to contain TF_d randomly selected feature.
2. for t = 1 to t max do
3. for each particle TF_d randomly do
4. for each data vector F_p
5. The similarity of two position vectors in each and every text feature can be measured by the angle among two vectors by the distance among two particles.
6. Assign feature vector F_p to text data such that

$$similar(P, Q) = \frac{\sum_{tf=1}^{d} p_{tf} \times q_{tf}}{\sqrt{\sum_{tf=1}^{d} p_{tf}^{2}} \times \sqrt{\sum_{i=1}^{d} q_{tf}^{2}}} \rightarrow (8)$$

7. Calculate the fitness using equation (7)
8. Update the global best and local best positions
9. Update the text feature data using (1) and (2)

In EPSO text feature extraction is different from normal text feature mining algorithm update the text features extracted results equations (1) and (2) make use of exponential inaction weight as specified in equation (6) as an alternative of linear inaction weight which specified in equation (5).

4. EXPERIMENTAL RESULTS

In this work second-hand the Web log data file beginning the http://startcompanyinindia.com as our test data. The data contains upto the data size of 23MB. The following Figure 4.1, Figure 4.2 and Figure 4.3 , shows the example of the Visitor's Page views, Page view and page custom details. The numeral of web pages in a web site and the normal numeral of out degrees of pages. The proposed EPSO analyzes the result of each web page those are extracted from web pages and what type of feature is extracted from page in specific time. In this thesis aspect explanation of page view are known. Pages in the website determination moreover are visited by the user not including thorough the site. The information of the pages stayed along with the normal page break time is showed. Page view concludes the traffic with the intention of website. Pages are consisting of files and each image in a page is an individual file. When a user looks at a page view they might see frequent images and produce numerous hits.

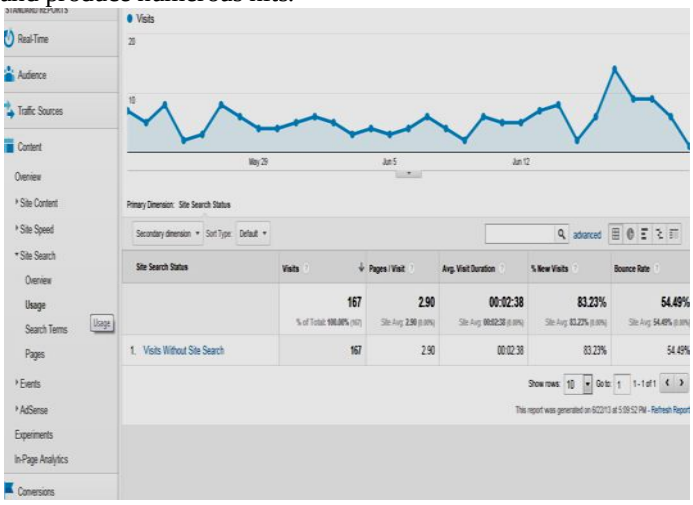

Figure 4.1 Page View

Website Visitors

The numeral of visitors with the purpose of a website obtains is a considerable assessment for numerous web extraction results. Google Analytics, broadly used platform for long-term situation, with make use of this numeral of visitors. At this time page view details are specified momentarily with numeral of calculation in Page views, distinctive page views, moment in time spend in page throughout their visit and the proportion of outlet page

Figure 4.2 Site Visitor

Due to the growing quantity of information available web log file, it has suitable individual of the mainly important resources. Web mining technology is the accurate end result for information finding on the Web. In the current work, proposition a novel method to improve the knowledge competence and decrease the computation concentration of an aggressive learning BPNN. The conception following it to describe center positioned in dissimilar values of text features and associates the similar text feature data from learning BPNN. It learns and examines valuable information from the obtainable Web log data in efficient manner.

Figure 4.3 Page Flow

5. CONCLUSION AND FUTURE WORK

In this paper a new schema for regular finding of user session and web log data in page view with different sessions and a new Back Propagation Neural Network (BPNN) to extract each and every user navigational patterns from web logs. The session of each user individual a temporally compacted series of web access through an each and every user. In this thesis can finish that to categorize frequent patterns in Web by using Extended Particle Swarm Optimization (EPSO) improved than additional methods. EPSO has an improved result of text feature extraction information for each and every web user log files are learned from web log file with well suited learning result. In this thesis Extended Particle Swarm Optimization method is established to the study of mining text feature from learning Web log data result from BPNN. Normal BP representation has completed full association of every node in the layers from input to output layers. Consequently, it take a assortment of computing point in time and iteration computing for high-quality results and less established error rate when we are responsibility a number of pattern generation. The investigational outcomes based on the data beginning the Web log file of the server shows that our EPSO system is very useful for a particular area. The results of the clusters generated beginning the Extended Particle Swarm Optimization demonstrates that it can successfully determine usage patterns. Our results are able to moreover be second-hand to forecast the user's browsing behavior based on the history knowledge.

The investigation and execution has obtainable in this proposal is in a hopeful step and is merely Website precise. Though these individual explanations of the algorithm are extremely inspired and recommend an assortment of range to be extensive on to additional difficulty domains. Additionally, somebody concerned in this field can take a comparable move toward and adapt this method to develop them to a universal situation to a dissimilar region. Usage of data collection to distributed manner is well efficient both non scalable and impractical. Therefore, their desires to be an approach where information mined beginning various logs can be incorporated mutually into an additional complete representation.

REFERENCES

1. Shafiq Alam, Gillian Dobbie, Patricia Riddle and Asif Naeem *M. International Conference on Web Intelligence and Intelligent Agent Tech*nology, 2010.

2. Van Der Merwe D.W. and Engelbrecht A.P *Congress on Evolutionary Computation*, Canberra, Australia,2003 .

3. Omran, M. "**Particle Swarm optimization methods for pattern Recognition and Image**

Processing", Ph.D. Thesis, University of Pretoria, 2005.

4. Wang Shi, etc.. Web mining〔J〕. *Computer Science*, 27 (4) ,pp 28-31 , 2000.

5. J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. ,"**Web usage mining: Discovery and applications of usage patterns from web data**", *SIGKDD Explorations*, 1(2):12–23, 2000.

6. W. Lin, S. A. Alvarez, and C. Ruiz**," Efficient adaptive-support association rule mining for recommender systems**", *Data Mining and Knowledge Discovery*, 6:83–105, 2002.

7. X. Fu, J. Budzik, and K. J. Hammond**," Mining navigation history for recommendation",** *In Proceedings of the 2000 International Conference on Intelligent User Interfaces, New Orleans,* LA,January 2000. ACM Press.

8. B. M. Sarwar, G. Karypis, J. Konstan, andJ. Riedl. **"Analysis of recommender algorithms for e-commerce",** *In Proceedings of the 2nd ACME-Commerce Conference (EC'00),* Minneapolis, MN, October 2000.

9. M. Deshpande and G. Karypis**," Selective markov models for predicting web-page accesses",** *In Proceedings of the First International SIAM Conference on Data Mining*, Chicago, April 2001.

10. L. Schmidt-Thieme W. Gaul**.," Recommender systems based on navigation path features",** *In Proceedings of Web KDD Workshop at the 7th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining (KDD01),* San Francisco, August 2001.

11. Parsopoulos, K. E. & Vrahatis, M. N,**"Particle swarm optimizer in noisy and continuously changing environments ",** *InM. H. Hamza (Ed.), Artificial intelligence and soft computing.* Anaheim: IASTED/ACTA. pp. 289–294

12. Pugh, J., Martinoli, A. & Zhang, Y**," Particle swarm optimization for unsupervised robotic learning**", *In Proceedings of IEEE Swarm Intelligence Symposium (SIS)*, pp. 92–99), Piscataway: IEEE,2005.

13. W. Lin, S. A. Alvarez, and C. Ruiz**," Efficient adaptive-support association rule mining for recommender systems",** *Data Mining and Knowledge Discovery*, 6:83–105, 2002.

14. Cui, X., Potok, T., Palathingal, P., **"Document Clustering using Particle Swarm Optimization",** *Swarm Intelligence Symposium, Proceedings* IEEE, pp. 185- 191,2005.

15. Li-ping, Z., Huan-jun, Y., Shang-xu, H.," **Optimal Choice of Parameters for Particle Swarm Optimization**", *Journal of Zhejiang University Science*, Vol. 6(A)6, pp.528-534, 2004.

16. Finkel, Jenny Rose and Christopher D. Manning, **"Joint parsing and named entity recognition",** *In: Proceedings of NAACL 2009.*