Web Page Recommendation using Random Forest with Fire Fly Algorithm in Web Mining

Pradip Suresh Mane, Ashok Kumar Jetawat, Pravin Jagannath Nikumbh

Abstract*: Nowadays, internet has become the easiest way to obtain more information from the web and millions of users search internet to find out the information. The continuous growth of web pages and users interest to search more information about various topics increases the complexity of recommendation. The user's behavior is extracted by using the web mining techniques, which are used in web server log. The main aim of this research study is to identify the navigation pattern of users from the log files. There are three major steps in the web mining process namely pre-processing the data, classification of pattern and users discovery. In recent periods, the web page articles are classified by the researchers before recommending the requested page to users. However, every category size is too large or manual labors are often needed for classification tasks. A high time complexity issues are faced by some existing clustering methods or according to the initial parameters, these techniques provides the iterative computing that leads to insufficient results. To address the above issues, a recommendation for web page is developed by initializing the margin parameters of classification techniques which considers both effectiveness and efficiency. This research work initializes the Random Forest's (RF) margin parameters by using the FireFly Algorithm (FFA) for reducing the processing time to speed up the process. A large volume of user's interest data is processed by these margin parameters, which provides a better recommendation than existing techniques. The experimental results show that RF-FFA method achieved 41.89% accuracy and recall values, when compared with other heuristic algorithms.

***Keywords* : Clustering, Firefly Algorithm, Margin Parameters, Random Forest, Weblog, Web Usage Mining.**

I. INTRODUCTION

The Internet users access the vast amount of information across the world, where this information are provided by worldwide web. When each user visited a website, he/she leaves the traces in a log files of the pages. The user faces the many problems when they interact with the websites. The consequences of data mining are rapidly developed to solve the problems with the help of developed computing techniques for various applications [1, 2]. The difficult task is to identify the differentiate between interested customer and non-interested customer, where weblog is used to classify the interest of user. When the user accessing the websites, the history of information is available in the weblogs, which is used for the prediction of user behavior [3]. The unstructured format is presented in the weblog, where this format is only used for the analysis of user behavior. Therefore, web mining algorithm is developed to analyze the useful information from the weblog data. The user often surfs the internet due to the large collection of resources, which may leads to information overhead [4,5]. In order to improve the qualities of user experiences and performance of web sites, it is important to recognize the importance of web page personalization techniques and identify the behavior of users on web sites. The various web pages contains different contents includes image, video, audio or text [6,7]. Before, recommending similar web pages to other users, the reading history of users in terms of content are used by the recommendation techniques. The different navigation paths are followed by the same user to browse the similar web pages. Hence, it is impossible to find the next web page of the user only by considering the exact and previous session logs of the user [8,9].

The challenging task is to identify the web page navigation of users and it is necessary to improve the rate of accuracy for searching the information and accessing speed for vast amount of web data sources. The formulation of guidelines is developed by understanding the next page access of user for website personalization [10]. The web page organization is the main step in the personalized recommendation system for web page, which are produced every day in human life [11]. But, the user-item utility prediction method is not simply used in Web Page Recommendation (WPR) system, because it often faces the problem of cold-start issues. Therefore, the web pages should classify before recommending to the user that are carried out by various researchers in recent years. There are four main methods proposed in the web mining process such as association rules [12, 13], clustering, sequential patterns and classification, which is used to discover the patterns in web log. The clustering method are used by some studies [14] for pre-processing the news corpus that lead to better results, however several clustering methods have different properties. For example, according to initial parameters, some clustering methods provided poor results, which is not stable by iterative computing and also it leads to high time complexity. To address this issue, the proposed method initialized the margin parameters of machine learning techniques by using an optimization technique.

These optimized margin parameters are given as input for classification technique to predict the web pages of users from weblogs.

Revised Manuscript Received on February 05, 2020.

* Correspondence Author

Pradip Suresh Mane*, Department of Computer Engineering, Pacific Acadamy of Higher Education and Reseach University, Udaipur, India. Email: Pradipmane510@gmail.com

Dr. Ashok Kumar Jetawat, Department of Computer Engineering, Pacific Acadamy of Higher Education and Research University, Udaipur, India. Email: drashokjetawat@gmail.com

Pravin Jagannath Nikumbh, Department of Computer Engineering, Lokmanya Tilak College Of Engineering, Navi Mumbai, India. Email: pjnilkumbh@gmail.com

The accuracy prediction of the proposed method is validated by conducting the experiments on MSNBC dataset. The validated results proved that the RF-FFA system achieved better results and it is helpful to acquire the social information of users over web navigation, which is based on recommender system.

The rest of this paper are organized as: The review of recent techniques used in predicting the user navigation from the weblogs are described in Section II. The problem statement of this research work and their solutions with explanations are presented in Section III and IV. The various experiments are conducted to validate the effectiveness of RF-FFA algorithm is explained on Section V. Finally, the conclusion of this research work with future development are given in Section VI.

II. LITERATURE REVIEW

Various studies have been developed to improve the efficiency of recommender systems and predict the user's navigations from the weblog data. In this section, a survey of recent techniques is presented with its advantages and limitations.

D. Anandhi, and MS Irfan Ahmed, [15] identified the user and their navigation pattern from the web log files, where the process includes major steps such as pre-processing the raw data and classify the pattern discovery and users. This approach identified the user sessions in a web log server and then found out the important sessions by applying the rough set clustering technique, which was based on the maximum pages visited in the websites. There were three types of user namely the users of potential, frequent and synthetic users, but this paper focused only on classifying the potential user from a web log data.

X. Xie, and B. Wang, [16] developed the TClus algorithm-based architecture by combining the two approaches namely k-means and density-based clustering techniques, which was used to tackle the personalized WPR. The problem of news-evolving in the WPR were solved by TClus algorithm by considering the clusters of irregular web page content. In order to improve the effectiveness of selection strategy, the utility between web page and given user was predicted by surplus function in TClus algorithm. In addition, the problem of cold-start was also solved by using this effective selection strategy. However, there is no relationship between web page objects and the users, which is considered as a drawback of the TClus algorithm.

E. A. Neeba, S. Koteeswaran, and N. Malarvizhi [17] completed the task for selecting the features by using swarm based cluster algorithm (SBCA), which was used to provide the optimized feature clusters for weblogs and data classification. The results stated that the method achieved high accuracy with less output errors and also consumed minimum time for web page classification. When compared with other attribute selection algorithms, SBCA provided higher optimization ability. However, a high-dimensional data was not handled by this method, where these data were related to real-time applications.

E. Sen, I. H. Toroslu, and P. Karagoz, [18] identified the next web page by using semantic user session generation module and content-based semantic similarity method with extended standard clustering approach. A set of concepts are described as semantics of web-pages, and sequence of sets modelled as user session. The amount of searching time was reduced by this method and also the accuracy rate was dropped at an acceptable level. The cost of clustering and keeping clusters up-to-date that didn't affect the prediction process, since these calculations were done offline while the prediction module was active. But, it is not possible to keep all previously recorded sessions, which lead poor performance. The execution time for every configurations is very close and short to each other, due to small size of the dataset.

R. Katarya, and O. P. Verma, [19] recommend the web-pages according to the sequential information of user's navigation by employing Fuzzy C-Mean (FCM) clustering for receiving the top-N clusters. The method found out a user's next Web page visit by identifying the similar users to target user. The top page was predicted and recommended to the target user by determining the weights for every page. The validation of FCM approach was carried out by several experiments on real world dataset called MSNBC. There is no privacy of user's data and trust of social websites, which is considered as the major problem of FCM method.

H. Jindal, N. Sardana, and Raghav Mehta, [20] described the various aspects of visualization for web navigation mining. The navigation details of users were identified by three process such as cleaning, pre-processing and recognition of patterns. After the pattern identification, the visualization of web data was illustrated by two components. The investigation of user behavior, structure and evolution of website, rare and frequent patterns or anomaly detection were carried out by web data visualization. In addition, the hidden information from vast data were analyzed and discovered by visualization. However, the method provides poor performance when the data contains outliers.

III. PROBLEM DEFINITION

- The processing time is reduced and data size will increase due to continuous growth of internet users', and in the current situation, internet access among vast amount of users are also increased.
- According to the user's interest, the classification of user pattern is very difficult for personalized recommendation in large-scale dataset.
- The another difficult task in web mining is to predict the behavior of user. The history of information about user's websites are presented in weblog.

Solution to this problem: The margin parameters used in classification are tuned by initializing the optimization techniques to speed up the process and reduce the processing time. This initialization of margin parameters is used to process the large volume of user's interest data and also provides better recommendation than existing techniques.

IV. PROPOSED METHODOLOGY

The user navigation patterns are discovered by developing an integrated framework using data mining techniques.

A better browsing experience is created in an efficient manner by extracting and finding the desired resources and information, where these techniques helps both web users and web administrators. The estimation of users is used to minimize the time for analyzing the irrelevant entries and also used for identification of robot entries. The classification techniques are used to classify the web pages of a user from the log files. Figure 1 shows the working procedure of the research work.

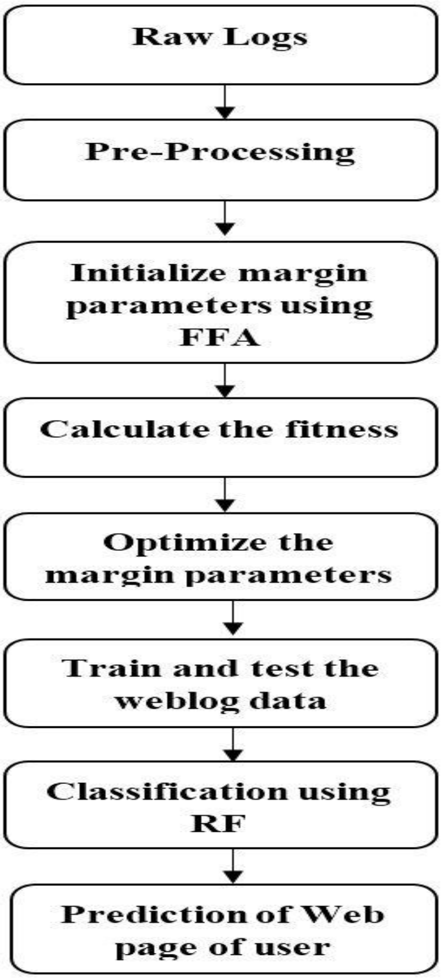
Raw Logs
Pre-Processing
Initialize margin parameters using FFA
Calculate the fitness
Optimize the margin parameters
Train and test the weblog data
Classification using RF
Prediction of Web page of user

Fig. 1. Working Procedure of Firefly Algorithm with Random Forest

A. Data Pre-processing

There are three main steps included in the web usage mining process i.e. pattern discovery, data pre-processing and pattern analysis. Pre-processing is the most time consuming and complex process, when compared with other process due to changes in log data. While comparing with other phases, it takes much time to pre-process the log files. Label encoder is used as pre-processing techniques in this research work, which is used to convert the categorical data or text data into numbers. According to the operations of labels, the machine learning algorithm works in a better way to predict the user's web blog data. The label encoder is considered as one of the important pre-processing steps for structured dataset in supervised learning. The output of the data pre-processing step is to split the database into training and testing data, which will be given as an input for further processing. To overcome the issues of clustering techniques, this research work initializes the margin parameters by using FFA. The steps are further explained as follows:

B. Parameter Initialization

The existing techniques use the clustering of data to divide the data into groups or classes, where the same items present in the same group of clusters is called as similar objects and the various items are presented in the other group of clusters is defined as dissimilar objects. The various items are placed into classes by using several actions, which is based on the determination of clustering techniques and nature of data. The similarity measures are used to control the clusters formation. The browsing history of users is considered as the major criteria to evaluate the similar users. There are two steps involved in pattern recognition based recommendation system, which includes clustering as the first step and followed by classification tasks. The system classification accuracy is significant, because the system is provided with sufficient learning in the initial stages. Once the system learned, a set of recommendations with good rankings are generated. Sometimes, a multiple cluster are formed by the same web users, because he/she may have multiple interests. Therefore, number of clusters are formed and also the identification of web pages from the log data that leads to difficult tasks. To overcome this issues, this research work uses the Machine Learning techniques for predicting web pages.

But, before applying to the classification technique, margin parameters of RF should be initialized. Firefly algorithm is used to optimize these margin parameters by calculating their fitness function. The firefly algorithm is described as follows:

C. Firefly Optimization Algorithm for initialization

A new FFA is designed by the fireflies' behavior, which are used to fly in the tropical summer sky. A prey searching, communication and identifying the mate using flashing patterns with bioluminescence are the basic characteristics of FFA. Therefore, a various meta heuristic algorithms are implemented by using FFA's natural properties. In this research work, a FFA-inspired algorithm is developed by some important characteristics of FFA. The following pseudo code explains the basic steps in FFA:

```
Firefly Algorithm
Consider $f(x), x=(x_1,......x_d)^T$ as objective function
Consider $x_i\ (i=1,2,....,n)$ as a fireflies population
Define light absorption coefficient $\gamma$
While $(t < MaxGeneration)$
for $i=1:n$   all n fireflies
   for  $j=1:n$  all n fireflies
        $f(x_i)$ is used to determine the light intensity $I_i$ at $x_i$
        if $(I_j > I_i)$

Towards $j$, move firefly $i$ in all $d$ dimensions
        end if
            Attractiveness
varies with distance $r$ via
```

```
$\exp\left[-\gamma r^2\right]$
            Update the intensity of light and calculate the new solutions.
        end for $j$
    end for $i$
    Find the current best value and rank the fireflies
    end while
    Calculate the results for process and post the visualization.
```

The Eq. (1) is used to identify the firefly i movement, which is attracted to another more brighter firefly as j. The mathematical expression for Eq. (1) is as:

$$x_i^{t+1} = x_i^t + \beta_0 e^{-\gamma r_{ij}^2}\left(x_i^t - x_j^t\right) + \alpha\varepsilon_i^t$$

(1)

Where, at $r=0$, the attractiveness is illustrated as β_0, according to the attraction, the identification of second term is carried out. With the random variables vector as ε_i, the third term is randomization, where the Gaussian distribution is used to obtain those random variables. The Cartesian distance $r_{ij} = \left\|x_i - x_j\right\|_2$ or the $I_2 - norm$ is used to find the distance between any two fireflies i and j at x_i and x_j.

The fitness value of objective function and the position of fireflies are the major components used in FFA techniques. The best objective values are provided by FF, which is shown by the optimistic position and every FF has own range of view because more fireflies are attracted towards their direction. When the firefly range is high and there are less number of fireflies, the neighbor fireflies affected their range and more fireflies are attracted. The optimal solution position is found out, only when the firefly range is small.

Let assume that N is firefly swarm id, the objective function is illustrated as $f\left(x_i, y_j\right)$, which is used to match the position of firefly as i and T_i describes the fluoresce of firefly. The Eq. (2) presents the updating formula for every firefly view range as.

$$f_k^i\left(u+1\right) = \min\left\{f_t, \max\left\{0, f_k^i\left(u\right) + \beta\left(t_u - \left|t_u\left(u\right)\right|\right)\right\}\right\} \qquad (2)$$

Whereas, in $u+1$ range, the i^{th} firefly is represented as $f_k^i\left(u+1\right)$, the neighbor firefly's threshold value is described as t_u. The control constant is presented as β and range of firefly with high fluoresce is depicted in $t_u\left(u\right)$. Therefore, the formula of $t_u\left(u\right)$ is mathematically represented in Eq. (3),

$$t_u\left(u\right) = \left\{j : \left\|y_j\left(u\right) - y_i\left(u\right)\right\| \mathrm{p}\ f_k^i l_j\left(u\right)\right\} \qquad (3)$$

Whereas, in t generation, the position of j firefly is described as $y_i\left(u\right)$, the fluoresce value of this firefly is represented as $l_j\left(u\right)$ and the range of neighbor firefly is depicted as f_k^i. Eq. (4) shows the probability selection of firefly neighbors.

$$f_{ij}\left(u\right) = \frac{l_i\left(u\right) - l_j\left(u\right)}{\Sigma_{k\in n_i(u)} l_k\left(u\right) - l_i\left(u\right)}$$

(4)

The position updating formula of firefly is shown in Eq. (5).

$$f_i\left(u\right) = f_i\left(u-1\right) + s\frac{f_j\left(u-1\right) - f_i\left(u-1\right)}{\left\|f_j\left(u-1\right) - f_i\left(u-1\right)\right\|}$$

(5)

The Eq. (6) is used to describe the Fluoresce in value's formula.

$$f_i\left(u+1\right) = \left(1-t\right)l_i\left(u\right) + \gamma k\left(c_i\left(u+1\right)\right) \qquad (6)$$

In the Eq. (6), the function value is measured by using the parameter called γ, where the fitness function value is illustrated as $k\left(c_i\left(u+1\right)\right)$. By using the fitness function, the FFA optimize the margin parameters of RF to predict the weblogs pages of end users. Then, these best model are given as an input to classification algorithm for predicting the weblogs. The classification of RF are given as below:

D. Classification of Random Forest

Among the research community, ensemble learning algorithms namely boosting, bagging and RF gained more interest due to its robustness against outliers and noises than single classifiers. In general, it uses hundreds of diverse classification trees as a composite classifier. While applying the majority rule of a sample over the votes of single classifiers, the decisions are carried out for selecting the final classification for that given sample. Each tree is grown by using a reduce samples to avoid the correlated and similar predictions in a training set. The best splits are identified by introducing the randomness in an algorithm for increasing the diversity between the training samples. There are many advantages presents in the application of RF, which are developed by Breiman and their benefits are stated as below:

- It works on large scale dataset effectively and efficiently.
- It can process more than thousands of input variables without deleting any variables in the tree.
- It predicts which variables plays an important role in a classification process.
- It generates an internal unbiased estimate of the generalization error.
- The pairs of cases are used to locate the outliers and their proximities are identified by RF algorithm.
- It can able to handle the presence of noise and outliers.

When compared with other ensemble methods, RF is computationally
light-weight method.

A set of decision trees with controlled variations are constructed by combining the randomly selected features with “bagging” ideas of Breiman's.

Algorithm: Random forest classifier

Input: Consider the dataset N, which is a collection of observed and their associated class values in training process.

Output: Prediction of weblog pages for providing the recommendations to the end users

According to the following steps, the construction of every tree are as follows:

- Consider N as a number of training cases and M as a number of variables in the classifier.
- At a node of the tree, the decisions are identified by using the number of input variables as m and the variables in the classifier M should be higher than input variables m.
- From all N available training cases, n times with replacement are chosen to select the a training set for this tree. The errors of the trees are identified by using the remaining cases and also by calculating their classes.
- Randomly choose m variables for each node of the tree, and according to these m variables, the best splits are calculated in the training set.
- To construct a normal tree classifier, each tree should not be pruned and new samples are pushed down the tree for prediction. It ended up when the training samples labels are assigned in the terminal node and the iterations are conducted over all trees in the ensemble. The RF predictions are reported by finding the average vote of all trees.

While entering in the model, the relevance features are identified by RF in a way, where each features are shuffled at a time, then prediction error on this shuffled dataset are estimated by Out-of-Bag (OOB). The shuffled features' relevance is closely related to the relative loss in performance between the shuffled and original dataset. While OOB subsets are used by RF algorithm, the features importance is estimated and their computational efforts are not increased. By using these RF classifier, this research work can predict the web pages of users from weblogs.

V. RESULTS AND DISCUSSION

The validation of proposed RF-FFA method against existing techniques are presented through various experimental evaluation using MSNBC dataset, which are briefly explained in this section. In these experiments, classifier RF are considered as an ideal conditions and then, various heuristic algorithms namely Particle Swarm Optimization (PSO) and Grey Wolf Optimizer (GWO) are also implemented with RF-FFA. The efficiency of the proposed RF-FFA are tested against these heuristic algorithms in terms of accuracy, precision, recall and F-measure.

A. Experimental Setup and Parameter Evaluation

The experiments were implemented using Python 3.7.3 on a computer with Intel Core i5 CPU 2.2 GHz with 8.00 GB RAM. There are 5000 users with six page visited data is presented in MSNBC database, which is a large dataset among other database in WPR. This database is used for predicting the next visit of user i.e. seventh page. The parameters such as accuracy, precision, recall and f-measure are used to validate the effectiveness of RF-FFA against several heuristic methods. The formulas for measuring those parameters are given in Eq. (7) – Eq. (10).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (7)$$

$$\Pr ecision = \frac{TN}{TN + FP} \qquad (8)$$

$$\operatorname{Re} call = \frac{TP}{TP + FP} \qquad (9)$$

$$F - Measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \qquad (10)$$

Where, TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. In next section, the validation of RF-FFA are explained in detail with a graphical representation.

B. Performance of RF-FFA against various heuristic methods

In this section, the performance of an ideal solution of RF are validated against heuristic methods such as GWO, PSO and proposed FFA by using accuracy parameter. The experimental results are tabulated in Table 1 and their graphical representation of accuracy are presented in Figure 2. The best values are represented by bold.

Table- I: Comparative Analysis of RF in terms of Accuracy

Methods	**Accuracy (%)**
RF	37.12
RF - PSO	38.86
RF - GWO	39.29
Proposed RF - FFA	**41.89**

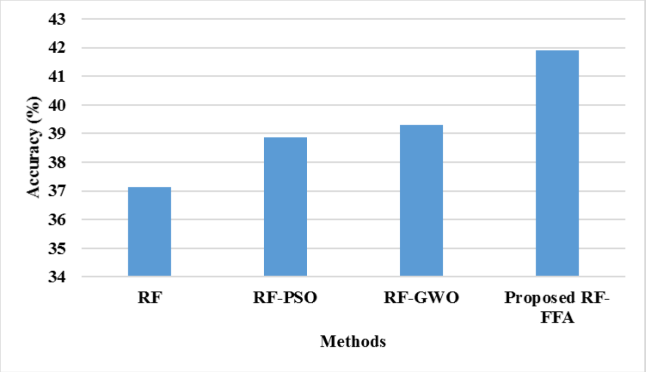

Fig. 2. Performance of proposed RF-FFA

From the Table 1 and Figure 2, the experimental results clearly stated that the performance of RF is increased by using the proposed FFA algorithm in terms of accuracy.

While using PSO, the accuracy of RF is 38.86% and GWO achieved 39.29% only. The ideal condition of RF achieved less than other algorithms i.e. 37.12% accuracy. But, when combining with FFA, the RF achieved higher accuracy i.e. 41.89% due to lighting intensity of fireflies, which will obtain the optimal solutions easily. The Table 2 shows the validated results of proposed RF-FFA in terms of precision. Figure 3 describes the graphical representation of precision values for the RF.

Table- II: Precision Performance of Proposed RF-FFA

Methods	Precision (%)
RF	37.56
RF - PSO	39.32
RF - GWO	38.92
Proposed RF - FFA	**39.62**

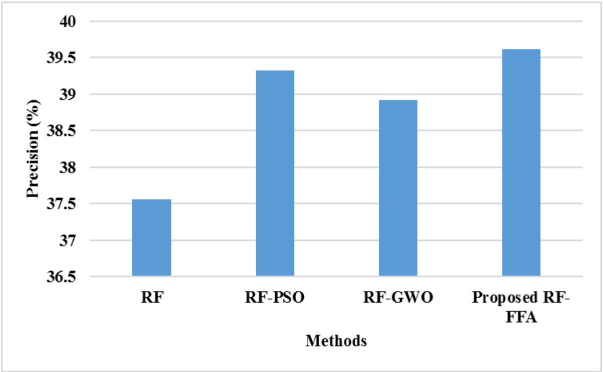

Fig. 3. Comparative Analysis of proposed in terms of Precision

When compared with ideal RF, the proposed RF-FFA algorithm achieved higher precision values. However, these methods achieved less precision value when compared with PSO algorithm, but higher than GWO. For instance, RF-PSO obtained 39.32% precision, RF-GWO obtained 38.92% precision, but the proposed RF-FFA achieved only 39.62%. According to random movements of both PSO and GWO finds the optimal solution, but FFA achieved the optimal solution according to the attracting behavior nature of fireflies. Table 3 presents the performance of ideal RF with several heuristic algorithms in terms of recall values, whereas Figure 4 shows their graphical representation.

Table- III: Performance of Recall for Proposed RF-FFA

Methods	Recall (%)
RF	37.12
RF - PSO	38.86
RF - GWO	39.29
Proposed RF - FFA	**41.89**

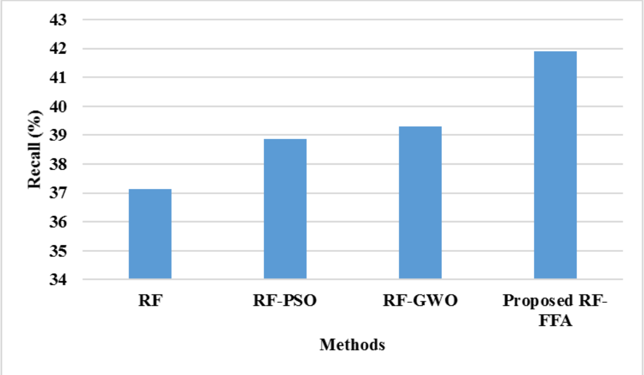

Fig. 4. Recall values of proposed RF-FFA against PSO and GWO

The performance of the proposed RF-FFA are validated by several experiments and their results are described in Table 3. From this, it is clearly explained that the RF produced higher recall values, while combining with proposed FFA. i.e. RF-FFA achieved 41.89% recall. However, the other heuristic algorithms obtain higher recall values than the ideal RF algorithm. The PSO and GWO achieved nearly 39% recall, but ideal RF achieved only 37.12% recall. The optimal solutions are identified by only random movements of PSO and GWO, which will automatically reduce the recall values of RF. But, FFA directly searches the optimal solutions based on the lighting intensity of fireflies, that will increase the recall values of RF. Finally, the performance of proposed RF-FFA are compared with several heuristics algorithms namely PSO and GWO are given in Table 4, where their results are represented in Figure 5.

Table- IV: F-Measure of proposed RF-FFA

Methods	F-Measure (%)
RF	37.18
RF - PSO	38.90
RF - GWO	38.73
Proposed RF - FFA	**38.66**

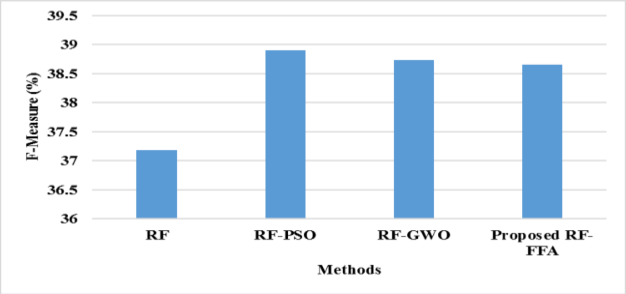

Fig. 5. Comparative Analysis of proposed RF-FFA in terms of F-measure

Compared with other parameter evaluations, the F-measure of proposed RF-FFA algorithm achieved low values than other existing heuristic algorithms such as PSO and GWO. The existing GWO and PSO achieved nearly 38.74% and 38.90% f-measure respectively, but ideal RF and proposed RF-FFA achieved only less f-measure (i.e. 37.18%, and 38.66%). This variation of proposed RF-FFA is due to the number of occurrence presents in the light intensity of fireflies.Because, the light intensity is directly propositional to TP, where the variations of occurrence will affect the intensity of light, which will automatically reduce the mean value of precision and recall. However, when compared with existing techniques, the proposed RF-FFA predicts the weblog data effectively. The overall performance of the proposed RF-FFA against heuristic algorithms in terms of all parameters is given in Figure 6.

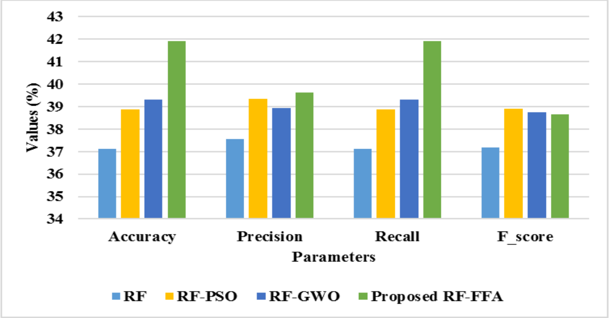

Fig. 6. Overall Performance of proposed RF-FFA

From the various experimental results, it is concluded that the proposed RF-FFA algorithms provides better performance than ideal RF, PSO and GWO in terms of accuracy, precision, recall and F-measure, which also predicts the weblog data effectively from the MSNBC dataset. The execution time of RF-FFA for every metrics is very short and close to each other, since the size of the data set is small. From the experiments, it is concluded that when selecting the huge number of web pages, the distribution of those list is much closer to the long-term reading interest of user in various centers. Hence, the selection of web page items is highly similar in RF-FFA model.

The proposed algorithm outperforms the other methods very significantly, because it considers the web page organization according to the light intensity value of fireflies. Therefore, when more web pages are recommended to the users, the computation overheads are highly reduced by using the proposed RF-FFA algorithm.

VI. CONCLUSION

Nowadays, number of users are started to browse the internet due to development of digital information. Hence, the developer community considers it as one of the important factors for focusing on user experience. In this paper, RF-FFA algorithm is used to identify the future visit of a user in a website. An issues in clustering method is addressed by initializing the margin parameters of RF are optimized by using FFA. The maximum number of users visited pages in websites are identified by applying the RF in important sessions of web server log. The experiments are conducted on MSNBC dataset to identify the effectiveness of RF-FFA algorithm, which provides higher prediction accuracy than other existing techniques. The validation of RF-FFA algorithm provides 41.89% accuracy than ideal RF (i.e. 37.12% accuracy), and existing PSO and GWO algorithms (i.e. 38.86% and 39.29% accuracy). The user's likes and perception are identified by the frequent visit of the user in weblogs, which is demonstrated by the results of RF-FFA. In the future work, the relation between the web pages and user should be identified by extracting the objects in the web page content. Among the users and web page objects, the relationship should be constructed and a hybrid intelligent systems should be used in the research work for privacy and trust of the user.

REFERENCES

1. P. O. Prakash, and A. Jaya. (2016). Analyzing and predicting user behavior pattern from weblogs. International Journal of Applied Engineering Research, 11(6). pp. 6278-6283.
2. M. Timilsina, W. Khawaja, B. Davis, M. Taylor, and C. Hayes. (2017). Social impact assessment of scientist from mainstream news and weblogs. Social Network Analysis and Mining, 7(1), pp. 48.
3. T. R. Ramesh, and C. Kavitha. (2013). Web user interest prediction framework based on user behavior for dynamic websites. Life Science Journal. 10(2). pp. 1736-1739.
4. R. Geetharamani, P. Revathy, and G. J. Shomona. (2015). Prediction of users webpage access behaviour using association rule mining. Sadhana, 40(8). pp. 2353-2365.
5. V. Pushpa, and V. Vidyapriya. (2016). An Efficient Preprocessing Method to Detect User Access Patterns from Weblogs. International Journal of Computer Science and Mobile Computing 5(9). pp. 16-22.
6. B. Li, Y. Lin, and S. Zhang. (2017). Multi-Task Learning for Intrusion Detection on web logs. Journal of Systems Architecture, 81. pp. 92-100.
7. L. Chen. (2017). An effective LDA-based time topic model to improve blog search performance. Information Processing & Management. 53(6). pp. 1299-1319.
8. X. Wei, Y. Wang, Z. Li, T. Zou, and G. Yang. (2015). Mining users interest navigation patterns using improved ant colony optimization. Intelligent Automation & Soft Computing. 21(3), pp. 445-454.
9. M. Jafari, F. SoleymaniSabzchi, and S. Jamali. (2013). Extracting users' navigational behavior from web log data: a survey. Journal of Computer Sciences and Applications. 1(3). pp. 39-45.
10. F. Khalil, J. Li, and H. Wang. (2009). An integrated model for next page access prediction. IJ Knowledge and Web Intelligence. 1(2). pp. 48-80.
11. A. Garg, and B. Kaur. (2014). Enhancing performance of web page by removing noises using LRU. International Journal of Computer Applications. 103(6).
12. N.T. Da, N. Thon, T. Hanh, and P. H. Duy. (2018). A Survey of Webpage Access Prediction. International Conference on Advanced Technologies for Communications (ATC).
13. S. P. Malarvizhi, and B. Sathiyabhama. Frequent pagesets from web log by enhanced weighted association rule mining. Cluster Computing. 19(1). pp. 269-277.
14. L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan. (2011). SCENE: a scalable two-stage personalized news recommendation system. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM. pp. 125-134.
15. D. Anandhi, and M.S. Irfan Ahmed. (2017). Prediction of user's type and navigation pattern using clustering and classification algorithms. Cluster Computing. pp. 1-10.
16. X. Xie, and B. Wang. Web page recommendation via twofold clustering: considering user behavior and topic relation. Neural Computing and Applications. 29(1). pp. 235-243.
17. E. A. Neeba, S. Koteeswaran, and N. Malarvizhi. (2017). Swarm-based clustering algorithm for efficient web blog and data classification. The Journal of Supercomputing, 1-14.
18. E. Sen, I. H. Toroslu, and P. Karagoz. (2016). Improving the prediction of page access by using semantically enhanced clustering. Journal of Intelligent Information Systems, 47(1). pp. 165-192.
19. R. Katarya, and O. P. Verma. (2017). An effective web page recommender system with fuzzy c-mean clustering. Multimedia Tools and Applications. 76(20). pp. 21481-21496.
20. H. Jindal, N. Sardana, and Raghav Mehta. Analysis and Visualization of User Navigations on Web. Data Visualization and Knowledge Engineering. Springer, Cham, pp. 195-221.

AUTHORS PROFILE

Pradip Suresh Mane is currently pursuing PhD (Computer Engineering) and working as an Assistant Professor in Computer engineering department at Shah and Anchor Kutchhi Engg. College, Chembur, Mumbai. He has teaching, research experience of 7 years. Big data, Data Optimization, Network security are just some of his areas of interests.

Dr Ashok Kumar Jetawat is a Professor in Pacific University ,Udaipur and chairman of Motivational Society of Achievers. After completing graduation in engineering, he did MTech, MBA and Masters in three other disciplines. He did his Doctorate from MLSU in E-Governance. He received "Rashtriya Ratana Award" in 2002 for individual outstanding performance. Dr Jetawat is Corporate Trainers having specialization in Study Techniques, Mind Management, Stress Management, Concentration, Soft Skills and Personality Development. More than 1,50,000 people from various organizations have been benefitted by the motivational seminars given by Dr. Jetawat.

Dr. Pravin Jagannath Nikumbh Ph.D. in Engineering at Jadavpur University. His research work is published in the book form titled as: Product Mix Analysis using Soft Computing: An Eclectic Solution by Lambert Academic Publishing Gremany (978-3-659-22388-4). He is working as Professor and Department Head at Lokmanya Tilak College of Engineering, Navi Mumbai. His area of interest are Advanced Manufacturing Technology, deals with high flexibility, high speed, least waste and computer based soft-computing techniques and application of hybridized artificial intelligence.