Ensemble Web Page Classification by Self & Non-Self with Firefly Optimization

Nikita Sahu[1], Dr. R.K. Kapoor[2]

[1]*Department of Computer Engineering & Application*

[2]*Associate Professor, Department of Computer Engineering & Application NITTTR, Bhopal M.P. INDIA*

Abstract - The problem of assigning pre-defined class labels to incoming, unclassified Web Page is called as Web Page classification. Sample of pre-classified Web Page is the base for the class labels. There are a number of approaches that have been proposed for Web Page classification like probabilistic, information retrieval based and machine learning. Problem of email classification can also be solved by applying these approaches. In web page classification mostly techniques rely on extracting features due to which the importance of extracting a group of related terms that co-occur is ignored and are unable to capture relationships between features. Web Page within a class bind to a set of patterns, and that these patterns closely correspond to, and are derived from the Web Page of the particular class. The features extracted in our method are derived from HTML contents through MATLAB code. We collected 234 web pages of academic class and rest 100 number of malicious web pages. The experimental results show that our method achieves a superior performance: the accuracy was over 95% in detecting malicious web pages.

Keywords— Web Mining, Firefly, Self & Non-Self, Feature Selection, Classifier.

I. INTRODUCTION

As we know there is a rapid up gradation of the WWW that is World Wide Web. With this growth there is a need to offer automation for the web navigators of the web page classification and categorization this approach will become helpful for maintaining the large figure of data in the keyboard depended search engines for the several web documents such as Yahoo directory and MSN directory.

The conventional work process involve the observation And classification of the data of that web page with the help of the digits of the domain expert, but it is invalid for the web page classification as there are several pages are present on the internet. Here we use clustering algorithm associate with the web pages to reduce the difficulties of the classification and speedy. But it is static as the algorithm needed big number of clusters in advance.

The article is organized as follows: Section 2 is all about various optimization techniques. Section 3 describes about web page followed by section 4 with data mining. Web Page Classification is handled in section 5. Section 6 talks about data reduction techniques followed by section 7 of literature review. Section 8 talks about the proposed work. The dataset used for this research is discussed in Section 9 along with experimental setup, results and discussions used for this research is discussed. The conclusion of this research is summarized in Section 10.

II. OPTIMIZATION TECHNIQUES

The development of optimization techniques began during World War II, when they were used to optimize the trajectory of missiles. Subsequently, mathematical programming has been developed to realize optimization through the application of mathematical techniques. Additionally, the optimization technique of meta-heuristics (MH), which imitates physical phenomena and the evolution of living organisms, has been developed since the 1970s. Furthermore, Fuji Electric has begun working to develop an optimization technique that is superior to conventional meta-heuristics. In the past, new optimization techniques had been developed in each era in order to realize customer solutions. But now, those optimization techniques are dated and new solutions are being realized.

a. Particle Swarm Optimization (PSO)

The particle swarm optimization (PSO), originally developed by Kennedy and Eberhart [1], is a technique with the aim of optimizing difficult numerical values on metaphor of social behaviors of that schools of fish and flocks of birds. It is an evolutionary computation technique based on swarm intelligence. A swarm consists of individuals, which known as called particles, which keep changing with respect to time. Each particle shows a potential solution to the error. In that PSO system, particles fly around in multidimensional searching space. At the time of every flight they adjust their position according to the experience and takes the experience of the neighbor to place it on the suitable position.

The effect is that particles move towards the better solution areas, while still having the ability to search a wide area around the better solution areas. The calculation of the performance is done by the fitness function, having a association the problems which is being solved.. The PSO is taken as very efficient and quick in solving nonlinear, non-differentiable and multi-modal problems.

b. Firefly

Consequently each brighter firefly influence by their partners (regardless of their sex), which leads make it more faster . The main rules of the algorithm are as follows:

- All fireflies taken as unisex and they keep going for more better options and brighter ones not concern about the gender.
- If the level of attractiveness is higher then this will leads to brighter fireflies, Also the brightness may decrease as the distance from the other fire flies increases because the air consume the light. If they are unable to detect the brighter fireflies, then they move randomly.
- The measurement of the brightness level of a fireflies is obtained by the objective function of that known problem.

c. Fuzzy Logic

The fuzzy concept networks are then used to personalize the search engine outputs. In [2] Fuzzy Logic was used for offline processing to recommend URLs to users. Fuzzy Logic testing shows slightly lower precision and is harder to program for the fuzzy part. In [3] Bee Colony Optimization was used for IR however this optimization technique is not a widely covered area of research. The process to derive the MRL involves expert advice on plume track direction and plume arrival time using linguistic terms such as very favorable, very fast, which are descriptive in nature.

III. WEB PAGES

Search engines offer an easy access to web pages, which makes information extraction an easier task. The preprocessing work like filtering and parsing has to be done before content extraction. The extracted content, which can be used for knowledge classification and knowledge processing, provides information sources for enterprises or other organizations.

IV. DATA MINING

Data mining also term as “Knowledge Discovery and Data Mining” process, it is a computation technique used in the computer science field for the detection of the pattern of the large dataset of that statistics and artificial intelligence etc. The main aim of the data mining is to filter the information and convert it into an specific structure from the large dataset. There is an association of database, data pre-processing model, inference consideration, interestingness metrics, visualization and online updating [4] except the raw observation. The genial data mining work is like automatic or like semiautomatic observations of the big figure of the data in order to filter the unknown useful patterns like as data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). It implies the database methods like spatial indices. then the pattern will be observe as the summarized way of their input data and used as further like the machine learning and predictive analysis. It is also used to detect the multiple groups, which further deals for the exact results by decision support system. Data collection, data preparation and reporting these are the constitute of the data mining.

a. Association rule learning

Detects the connection between variables, As like in a supermarket, data obtained according to the customers shopping, but with the help of the association rule learning the supermarket lets the detection of the higher purchased products, and this will increase their marketing performance and this also know as market basket analysis

b. Classification

Here generalization of familiar data structure are being applied to the fresh data like as a E-mail might be classify the e-mail like “legitimate” or “spam”

c. Clustering

It is the task of discovering groups and structures in the data that are in some way or another "similar" without using known structures in the data.

V. WEB PAGE CLASSIFICATION

As we have seen there is a fast growth of WWW that is World wide web, therefore here also have to facility them the automated assistance to all web users for that web page classifier and categorization, this will assist in collecting the big value of information back fro the keyword-based search engines or generating catalogues, Sometime it becomes very hard to cop up without the automated web page classification method because of the labor-intensive nature , firstly we saw that the web page classification was drawn directly from that machine learning literature for those text classifier. On the close exams the s\results are somewhere on distance as like being straight forward[5].It is having their personal underlying embedded structure of HTML language. They also filled with some loud content like as advertisement banner and navigational bar. For the higher level bias for the classification algorithm is happened by the application of the pure-text classification, by letting it less concentrating on the major and crucial topic, Therefore it is little bit challenging to built up an smart processing

method in order to get the major and important topics of the web page[6]

VI. VARIOUS TECHNIQUES TO REDUCE DATASET SIZE

a. Rough set theory

Rough set theory [7] is taken as a fresh new work process of data mining and data analysis, Its is a mature concept as its almost working since 15 years, Recently the Rough set theory (RST) has become a hot topic and have immeasurable growth and the application is spread out world wildly, It is loaded with several type of concert like as international workshop, conference, seminars etc. Now several type to heavy quality papers are being published on some aspects of the rough sets. RST can be refer as the modification of the previous set theory.

RST is referred as the modified version of the formal set theory which relates with the approximation for decision making (Pawlak, 1982). It is taken as the approximation of the vague concept (set) by set of fixed concept, which is used for the classification of the disjoint categories known as lower and upper approximation, The domain objects are relate with subset of interest known as upper approximation

b. Self and non-self

Sometimes it is hard to match the humeral immune system, It is taken as the "right" cells are match or then this will cause the self-destructive autoimmune reaction. Classical immunology that is (Kuby 2002) guarantee that the response of the immune system is trigger, when the body detects something foreign particle. Even presently the concept of the self-non-self determination is not clear. but several immunologist are capable of finding the difference between them. The maturation method plays a key role to obtain the self-bearing power by removing the T and B-cells by itself. then additionally a "surety" signal is needed, his is might be for B-cell or T(killer) cell and the T (helper) cell also suppose to be active, This 2 way activation, secure in future time the possibility of accident.

c. Genetic

The relationship of the genetic search with connectionist computation is very reknown. It has been well recognized the genetic coding at the sub symbolic level. The reason behind this approach of "genetic connectionism" is the synchronic emergence of connectionist systems, It has been analyzed that the lower level computation is suitable for the higher level behavior. The genetic data is mostly deals with the connection power of neural network (e.g Belew, Mcinerney & Schraudoph 1990), or to convert the whole topology of networks (Miller, Todd & Hegde 1989, Whitley, Stark weather & Bogart 1989). These all techniques are providing excellent outcomes for the optimization of the network design.

VII. LITERATURE REVIEW

Recently much work has been done on Web-page summarization [8][9][4]. Ocelot [8] is a system for summarizing Web pages using probabilistic models to generate the "gist" of a Web page. The models used are automatically obtained from a collection of human-summarized Web pages. In[9], Buyukkokten et al. introduces five methods for summarizing parts of Web pages on handheld devices where the core algorithm is to compute the words' importance using TF/IDF measures and to select important sentences using Luhn's classical method [10]. In [4], Delort exploits the effect of context in Web page summarization, which consists of the information taken from the matter of all the documents linking to a page. It is shown that summaries that take into account of the context information are usually more relevant than those made only from the target document.

The increase figure of the information lets to the demand of the exact automated classifiers for the web pages for the maintenance of the Web directories and to enhance the performance, All the tags and terms are taken as the features but still it is essential to detect a method for the selection of the best features for the reduction of the web page space error. Here we are trying to work on the optimizations method like as "Firefly algorithm" to choose the correct feature. The firefly algorithm is taken as the meta heuristic algorithm , which is being inspired by the life style of the fire flies. With the help of the FA we choose the correct feature then evaluate it j48 classifier by the action of the data mining tool. For the evaluation purpose we take the Web KB and the conference dataset as the Web KB and conference data set are being classified in absence of loss of accuracy, but consume little bit more time. Hence the Web page lessen as the amount of features got reduced.

In this paper [8] authors present firefly Algorithm brought into existence by Xin-She Yang in 2007-2008 at Cambridge University, this method has coined its idea from the nature and behavior of fireflies. The assumptions introduced in this method are fireflies are unisexual, attractive feature of fireflies is directly dependable on their brightness , evaluate the brightness by the measure of the objective function etc.

VIII. ARCHITECTURE OF ENSEMBLE WEB PAGE CLASSIFICATION BY SELF & NON-SELF WITH FIREFLY OPTIMIZATION

Here we are discussing a new method of web page classification named Ensemble Web Page Classification by Self & Non-Self with Firefly optimization (EWPCSNFO).

The basic idea behind EWPCSNFO is to mine web pages from a HTML pages dataset. This dataset of classification is filter by Self & Non-self concept of danger theory which is optimized by Firefly technique.

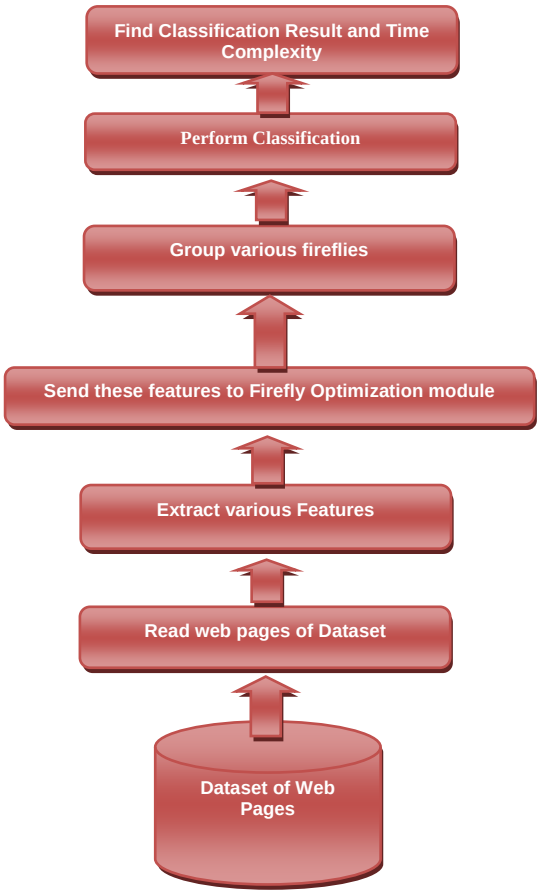

Figure 2: Flowchart of Proposed EWPCSNFO Work

While dealing with web-page classification there is a situation which decreases the performance of the classification process. And this issue is handled by the concept of Self and Non-self.

We have excluded various web-pages from the dataset which would not belong to 'Course' or non-Course category from the dataset as they are treated ad non-self element for our purpose.

IX. SIMULATION RESULT

In this section, we will evaluate the performance of artificial immune system's self- non-self based firefly feature selection based classifier with simulation test.

Table 1 : Course Category- with Sub-categories

S. No.	Course-Related Sub categories	Course-Related 2nd level Sub categories
1	Theory and Practical	Theory and Practical
2		Operating System
3		Computer
4	Theory	Theory
5		Numerical Methods
6		Algorithms
7	Practical	Practical
8		Artificial Intelligence
9		Multimedia
10	Contents	Contents
11		Syllabus
12		Course
13	General Information	General Information
14		Announcements
15		Cornell University

a. Collecting Sample Information

The dataset which is required to perform the experiments is a web pages dataset. This dataset is taken from the organization Department of Computer Science, Cornell University. It is very know university which includes various departments with various activities and provide their dataset for research purpose.

Number of Categories: mainly two.

Name of Categories:

1. Course
2. Non-Course

Total Number of Instances: 334.

Number of Course Instances: 234.

Total Number of Non-Course Instances: 100.

Table 2: Non-Course Category- with Sub-categories

S. No.	Non- Course-Related Sub categories	Non- Course-Related 2nd level Sub categories
1	Student	Student
2		Graduate
3		Engineering
4	Department	Department
5		Institute
6		Committee
7	Professor	Professor
8		Investigator
9		Guide
10	Research Associate	Research Associate
11		Ph.D
12		Thesis
13	Academics	Academics
14		Programs
15		Publications

b. Simulation Test

The system used for execution of the Self and Non-Self based Firefly feature selection method with existing [8] methods is as follows:

All the experiments were conducted on a PC, Dual-Core (2.20 GHz), with 4GB of RAM, running a Windows7 operating system and 32-bits operating system.

Time complexity

The effectiveness is measured in terms of the time requirement in web page classification techniques of paper [8] and proposed EWPCSNFO work. It compares between execution timing of the web page classification of paper [8] and EWPCSNFO work. Experiment is performed 10 times (as required time always gets change due to execution of many daemons processes at the same time).

Table 3: Execution time comparison between proposed work with existing work .

S. No.	WPC-FF	EWPCSNFO
1	10.579	7.541112
2	10.405	7.478047
3	10.326	7.36577
4	10.339	7.396642
5	10.393	7.387252
6	9.6687	6.942458
7	9.7143	6.922075
8	9.7632	6.955101
9	9.7141	6.957932
10	10.579	7.541112

Average of ten times of execution is calculated and shown in figure 3.

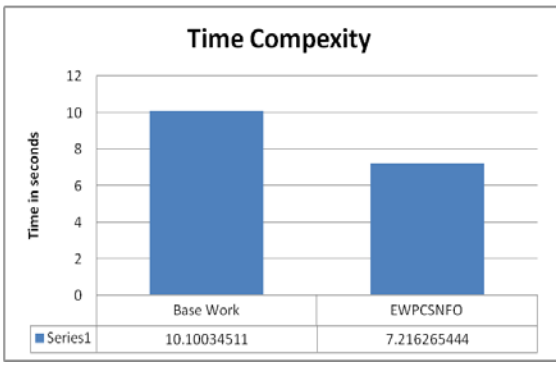

Fig3: One Instance of execution time of proposed EWPCSNFO work.

1. Web page classification

Classification could be find as follows:

$$classifcaton = \frac{numberofrecordsclassifiedproperly}{sizeoforiginaltotalnumberofrecordsindatasetldata} * 100$$

In the similar fashion web page classification could be like this:

$$Classificationofwebpage = \frac{Numberofwebpagesclassifiedpropoerly * 100}{SizeoforiginaTotalnumberofwebpagesindatasetl}$$

Table 4: Classification Result comparison between proposed work with existing work.

S. No.	EWPCSNFO work	Base
1	79.91453	61.676647
2	79.91453	61.676647
3	79.91453	61.676647
4	79.91453	61.676647
5	79.91453	61.676647
6	79.91453	61.676647
7	79.91453	61.676647
8	79.91453	61.676647
9	79.91453	61.676647
10	79.91453	61.676647

Average of ten times of execution is calculated.

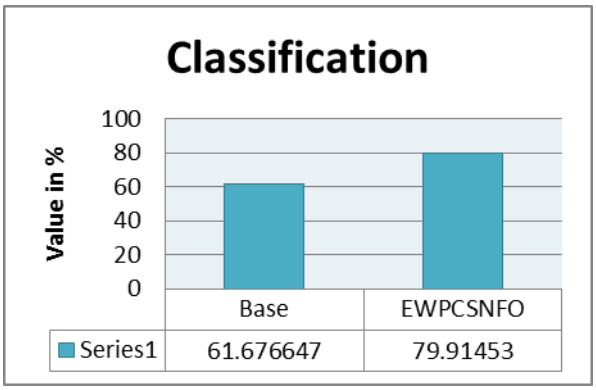

Fig 4: Classification result of proposed EWPCSNFO work

X. CONCLUSION

Ensemble Web Page Classification by Self & Non-Self with Firefly optimization (EWPCSNFO) demonstrates that the performance of web page classification is close to published results for the collection of data provided by Department of Computer Science, Cornell University, which can be used for speeding up the slow process. Thus, by considering at least that the user can prepare a representative collection of data, so that the resulting classifier should be exact enough to be used. To prepare such a required data set and to guide the user through describing candidate web pages, a graphical interface is

provided. This interface works in such a way that its backside searches through a pool of candidate web pages on behalf of a user and tries to find out the document whose label matches with the features will definitely improve accuracy.

After analysis of Table 3 and 4, which is output of the proposed method for Classification is further better.

After performing various operations on web page dataset, which contents html files, we land up with the result of improvement over base paper method. These improvements are in following direction:

1. Reduce Time Complexity
2. Improve Classification Result.

we can summarize that the following areas could be improved.

1. Classification of web pages can be included according to various features of Images based on input image's parameters.
2. Classification of web pages can be Included according to various features of Videos based on input video's parameters.

Classification of web pages can be Included according to various features of longitudinal parameters of query's physical location.

REFERENCES

[1] C. Leacock and M. Chodorow, "Combining local context with WordNetsimilarity for word sense identification," in WordNet: A LexicalReference System and itsApplication, MIT Press, 1998, pp. 265-283.

[2] Liu,T., Yang, Y., Wan,H., Zeng,H., Chen,Z.,& Ma,W. (2004). Support Vector Machines Classification with A Very Large-scale Taxonomy. SIGKDD Explorations , 1-36.

[3] Liu, T., Yang, Y., Wan, H., Zhou, Q., Gao, B., Zeng, H., Chen, Z. & Ma, W. (2005). An Experimental Study on Large-Scale Web Categorization. WWW 2005., Chiba, Japan, pp. 1106-1107.

[4] Melvin Cohn, "A biological context for the Self-Nonself discrimination and the regulation of effector class by the immune system," The Salk Institute for Biological Studies, Conceptual Immunology Group, 10010 N. Torrey Pines Rd. La Jolla, California 92037.

[5] Thomas Boehm, "Quality Control in Self/Nonself Discrimination," Leading Edge Review, Cell, 125, June 2, 2006 ©2006 Elsevier Inc.

[6] Aanshi Bhardwaj and Veenu Mangat,"A Novel Approach for Content Extraction from Web Pages," Proceedings of 2014 RAECS UIET Panjab University Chandigarh, 06 – 08 March, 2014.

[7] Xin-She Yang, "Firefly Algorithms for Multimodal Optimization," arXiv-2010.

[8] Esra Saraç and Selma Ayşe Özel,"Web Page Classification Using Firefly Optimization,"2013 IEEE.

[9] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for Web browsing on handheld devices. Proc. of WWW10, Hong Kong, China, May 2001.

[10] H.P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development, Vol. 2, No. 2, April 1958, 159-165.