IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

A New Technique of Web Page Classification and Optimization

R Khan[1*], R K Gupta[2], V. Namdeo[3]

[1,2,3]Department of Computer Science, SRK University, Bhopal, India

[*]*Corresponding Author: rkresearch2019@gmail.com, Tel.:9191537831*

Available online at: www.ijcseonline.org

Accepted: 24/Jan/2019, Published: 31/Jan/2019

***Abstract*—** The rapid development of the internet and web publishing techniques create numerous information sources published as HTML pages on World Wide Web. WWW is now a popular medium by which people all around the world can spread and gather the information of all kinds. The importance of these Web-specific features and algorithms, describe the state-of-the-art practices, and the following hypothesis. This work is for a better description of Web page classification problem. Since Firefly Algorithm (FA) is a recent nature inspired optimization algorithm, which simulates the flash patterns and characteristics of fireflies. Clustering is a popular data analysis technique to identify homogeneous groups of objects based on the values of their attributes. Here is used for clustering on benchmarks which is more suitable than Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and other nine methods used. The webpage optimization using Naïve Bayes classifier is an improved optimized web page classification using firefly algorithm with NB classifier. The inclusion of Naïve Bayes is an expert in the field of firefighting. Current classification techniques use word consistency and grouping techniques for classifying web pages. These Techniques use an ad hoc approach to review and reconcile whole keywords on a website for classification. These methods are effective, but not without problems like slow Processing, word meaning differences, poor identification of sentences also disregard the homonymy of the words. Hence this work is better, in the accuracy, precision, etc. parameters with respect to existing concepts.

***Keywords*—** Accuracy, Artificial Bee, Classification, Clustering, Colony, Firefly, Features, Homogeneous, HTML, Information, Optimization, Precision, Web, etc.

I. INTRODUCTION

The rapid growths of Internet usage and advances in communication technology have led to a rapid increase in the amount of text information online. Following this, it has become difficult to manage the huge amount of information online. To resolve this problem, many new techniques have been developed and used by search engines. Several tests are used to provide more accurate and faster results for users. One of the most important studies in this field is the text classification. Text categorization or classification, which is widely used by search engines, is one of the main techniques for handling and organizing text data. This growth of information has led to the need for accurate and rapid classification of Web pages to improve search engine performance [1]. Automatic classification of the website is a supervised learning problem in which a set of web documents tagged for training a classifier, then the classifier is used to assign one or more predefined category labels web pages for future use. Automatic classification of the website is not only used to improve the performance of search engines, it is also essential for the development of web directories, discussion of specific topics Web, contextual advertising links on the analysis of the structure current site and to improve the quality of web search. Several methods of classification such as decision trees, Bayesian classifier, support vector machines, k-nearest neighbours were developed [2-4].

WEB PAGE CLASSIFICATION TECHNIQUES

Ranking web pages is the technology developed in the form of text classification. The main text classification and classification of web pages difference is that websites have a lot of other information such as text links sound image, etc., which are very important in classification [3-6]. So is important to combine with information from the websites of web analysis. It plays a vital role in reaching characteristics of Web pages using technology [7]. This common text processing is the fundamental question of the text classification. Web pages based on maximum entropy model categorization is an effective method of experience. Web pages are divided by a pretreatment before the network structure [8-11].

II. RELATED WORK

[1] In this paper Author present a novel sensitive information classification algorithm and topic tracking

algorithm for Web pages contents. First, a text sensitive information classification method is proposed based on a vector space model and cosine theorem. Experiments show that the classification of the text sensitive information is very effective and result of topic tracking is ideal.

[2] In this paper Author presented a way to identify the category in which the article falls, which enrich the reader's knowledge with direct access the content or locate to the relevant titles. The paper describes a model to perform categorization on article related to child development and parenting's contexts, which starts with catalogue generation through identification & analysis carried with the consideration of category keyword along with relevant information been extracted from various sources over web to achieve the classification with expected accuracy.

[3] In this paper Author discussed about the classification of duplicate web pages. In this paper, we are proposing a five stage algorithm for the detection of near duplicate web pages, which include pre-processing, minimum weighting, filtering and verification and classification of the web page using apirori algorithm.

[4] In this paper Author discussed about A classification approach for Tibetan web pages is introduced in this paper. It takes advantage of the class feature dictionary and Rocchio classification algorithm to classify the Tibetan web pages into the predefined classes rapidly and accurately. The experimental results present that the approach has better classification accuracy for Tibetan web pages classification. It is useful and helpful for the construction of the statistical and rule-based classification of Tibetan texts as well as construction of high-quality Tibetan corpus.

[5] In this paper Author explores the use of formal source code structure for classifying a large collection of the web content. Is focused on use of schemas collection Schema.org to classify web pages and categorize them unambiguously.

[6] In this paper Author studies the process and methods of text classification. Based on Naive Bayesian algorithm and the semi-structured feature in Web page information, this paper proposes an improved Algorithm for Web page text Information classification which utilizes Html tag Information in classification. Experiments show that this algorithm is feasible and effective and can apply to information extraction in topic search engine, which can enhance the theme fitness of the search results and further improve the searching efficiency.

[7] In this paper Author shows that the algorithm of web page classification provides a good approach to facilitate Internet-related information analysis, and the E-government is useful for the government department standardizing their job procedure.

[8] In this paper Author discussed detailed design and implementation of a Chinese Web-page classification system is described in this paper and some methods on Chinese Web-page pre-processing and feature preparation are proposed. Experimental results on a Chinese Web-page dataset show that methods we designed can improve the performance from 75.82% to 81.88%

[9] In this paper Author discussed about multiple classifiers are built, one for each training domain, and the block classification proceeds through combining them.

[10] In this paper Authors adopt a novel method of Web page expression, and make use of summarization algorithm to reduce the noise of Web pages. A preliminary experimental comparison is made showing encouraging results.

III. METHODOLOGY

This new proposed work is being conceptualized as given in figure 1 where objective data set of web content applied with feature extraction algorithm like FA where appropriate extraction of feature takes place and applied to existing classification with J48 classifier and also its result applied and compared with NB classifier [12-15]. On comparative study, as expected, NB classifier overall produce better result in terms of two parameters that is accuracy and F – measure.

There are so many optimization methods are available for feature extraction by using different operations like GA, ACO, PSO algorithms [10].

The various optimized algorithms for feature selection are:

a. Firefly Algorithm(FA)
b. Ant Colony Optimization(ACO)
c. Particle Swarm Optimization(PSO)
d. Genetic Algorithm(GA)

Where FA is a based on wrapper technique which finds the best features for Web pages. ACO is based on probabilistic method for finding optimal path using graphs. PSO is a computational intelligence-oriented, population-based, stochastic, global optimization technique[6 & 15-17].GA The GA-based feature selector determines the best weight for each feature to find the most similar feature vector to the positive Web pages in the training dataset [14 & 16].

PROPOSED ARCHITECTURE

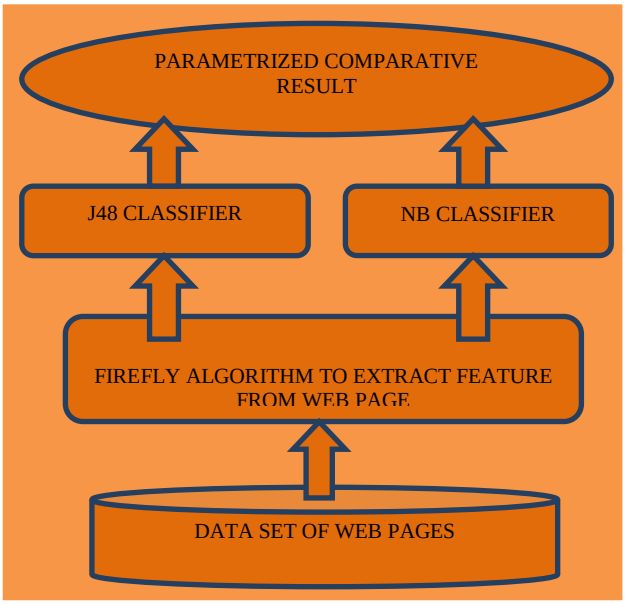


Figure 1: Proposed Architecture

PROPOSED ALGORITHM

To accomplish the desired task and its goals here an efficient algorithm produced below for which certainly some assumptions exists too. By assuming the following limitations and parameters this proposed work has been carried out.

(1) f = Features (2) fs = sub features,
(3) m = no. of web pages (4) n = no. of features,
(5) Load Bank Search m webpage dataset.

Proposed-Algorithm(){
Feature-Extraction(WebPages){
a) for loop i = 1:m
i)if found in web-page
Write f else
leave empty space
if condition end
ii)if fs found in web-page
Write fs along f
else
leave empty space
if condition end
iii)fi = f(i,1)
iv)fi = fs(i,2)
for loop end}
Feature-Index(fi){
A) for loop j= 1:m
i) If f or fs ==0
fi = 0
if condition end
ii) if f ==0 and fs==1
fi =lv, lv>0
if condition end
iii) if fs==0 and f==1
fi = iv, iv> lv
iv) if f ==1 and fs==1
fi = lvi, lv<lvi>l
if condition end
Nd = fi(j,1) (Nd =Numerical data)
for loop end}
Firefly(){ i. Pass Nd to Firefly Algorithm
ii. Firefly generate initial population of fireflies
xi = LB + rand* (UB-LB)
(Where LB= lower bounds, UB = upper bounds of ith firefly)
i) Obtain attractiveness, which varies with distance r
ii) Find new solutions and update light intensity for each generation (iteration) the firefly with maximum light intensity is chosen as potential optimum solutions
iii) Got firefly optimized output}

Classification-method(){
i) Make features groups
ii) Naive Bayes classifier Classify feature group and optimized data
iii) Make predict class from Naive Bayes model
iv) Create confusion matrix by (Feature group == Predict Class)
v) Got measuring parameters}}

Proposed Flow Chart

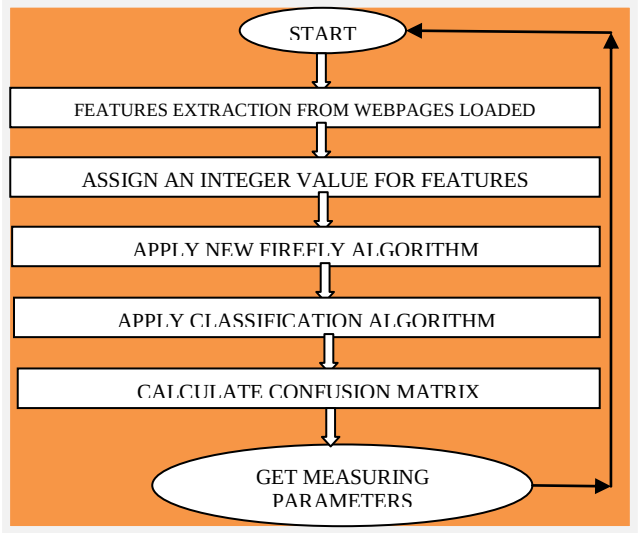


Figure 2: Proposed Flow Architecture

IV. RESULTS AND DISCUSSION

There is already lots of work already being done and still more other may going on the web page classification and optimization algorithms development. Out of them three are quite efficient algorithms. They are such as FA, PSO, ACO and optimized feature based methods. Out of them a lot of work done more on enhanced version of FA algorithms the experimental result and analysis is truly based on contents of pages. Here all the data set to test are stored in database they are called to test.

The tested result of previous work and proposed one is analyzed and observed by following parameters:

1. Accuracy
2. F-Measure

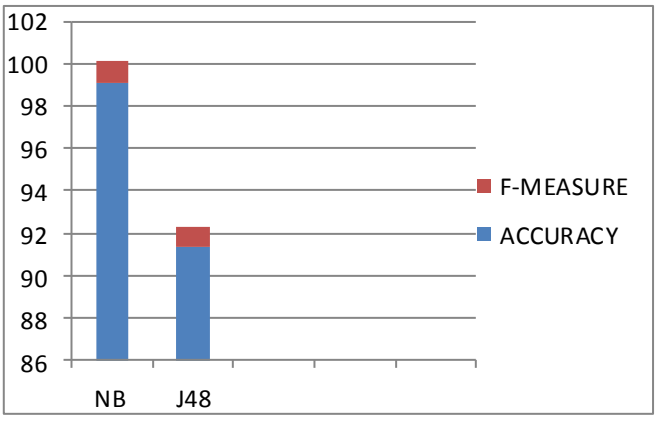
The test result collected in Table: 1 and corresponding Graph: 1 can easily distinguish betterment of proposed approach. It would analyze that the dataset has been implemented as test set up to 100%. And the basis of analyzing it would produce up to 99% correctly classified on behalf of different features.

Table 1 represents the comparison between numerical value of base and web page classification using NB classifier.

Table 1 Test Result Analysis

CLASSIFIER	F-MEASURE	ACCURACY
NB	0.9907	99.074
J48	0.9141	91.412

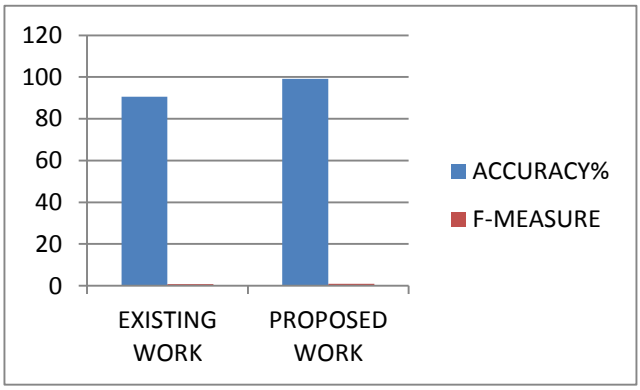
The graphical representation of table 1 placed below in Graph 1: Classifier Comparison



Finally a comparative analysis of existing work and this newly proposed work has been prepared here on the basis of concentrated parameters like accuracy and F-measure in

Table 2: Comparison of Base and Existing Work

CLASS	F-MEASURE	ACCURACY
EXISTING WORK	0.711	90.56
PROPOSED WORK	0.975	99.06



Graph 2: Comparison Graph of Existing and Proposed Work

V. CONCLUSION AND FUTURE SCOPE

As it can be seen that the proposed technique was found very useful from the existing techniques. The main drawbacks of existing methods are searching efficiency and content matching according to the feature measure. Secondly the need for frequent updates web content request success result and this is possible on testing the dataset over internetwork data content. Hence these aspects cause the main motive of the research to enable enhanced web optimization which is basically concentrated on the two facts.

This research work enable requester to fast extraction of web pages with corresponding web contents over store web server's data by appropriate metafiles. The entitled methodology will guarantee to generate best suited matched content pages.The processing overhead is tried to reduce or maintained on lower rates.

The entitled research work has justified its working efficiency for different datasets of long length. It can be further optimized by using some better approaches based on Artificial intelligence and training algorithms.

REFERENCES

[1] Guixian Xu ; Ziheng Yu ; Qi Qi, Efficient Sensitive Information Classification and Topic Tracking Based on Tibetan WebPages,IEEE Access, 2018

[2] Ankit Dilip Patel ; Vimal N. Pandya, Web page classification based on context to the content extraction of articles 2nd International Conference for Convergence in Technology (I2CT), 2017
[3] Eldhose P Sim, Classification & detection of near duplicate web pages using five stage algorithm, IEEE, 2015
[4] Guixian Xu ; Chuncheng Xiang ; Xu Gao ; Xiaobing Zhao ; Guosheng Yang, Automatic Classification of Tibetan Web Pages, International Conference on Computer Science and Electronics Engineering, 2012
[5] Jonáš Krutil ; Miloš Kudělka ; Václav Snášel, Web page classification based on Schema.org collection,2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012
[6] He Youquan ; Xie Jianfang ; Xu Cheng, An improved Naive Bayesian algorithm for Web page text classification, 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011
[7] Boyi Xu ; Jing Wang ; Hongming Cai, A Web page classification algorithm and its application in E-government system, 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 2010
[8] Weitong Huang ; Luxiong Xu ; Yanmin Liu, Preprocessing and Feature Preparation in Chinese Web Page Classification, 2009 International Conference on Computer Engineering and Technology, 2009
[9] Jinbeom Kang ; Joongmin Choi, Block Classification of a Web Page by Using a Combination of Multiple Classifiers, 2008 Fourth International Conference on Networked Computing and Advanced Information Management,2008
[10] Yong Zhang ; Bin Fan ; Long-bin Xiao, Web Page Classification Based on a Least Square Support Vector Machine with Latent Semantic Analysis, 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008
[11] Moonis Javed ; Aly Akhtar ; Akif Khan Yusufzai, Classification of Web Pages as Evergreen Or Ephemeral Based on Content, 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015
[12] Feiyue Ye ; Zhian Yu, Finding the Semantic Relation between Web Pages through Topic Knowledge Repository, 2009 Ninth IEEE International Conference on Computer and Information Technology, 2009
[13] Chinese Web-page Classification Study, Weitong Huang ; LuXiongXu ; Junfeng Duan ; Yuchang Lu, Chinese Web-page Classification Study, 2007 IEEE International Conference on Control and Automation, 2007
[14] Sumaia Mohammed Al-Ghuribi ; Saleh Alshomrani, A Simple Study of Webpage Text Classification Algorithms for Arabic and English Languages, 2013 International Conference on IT Convergence and Security (ICITCS), 2013
[15] Daya Gupta ; Harsh Tripathi ; Mayukh Maitra, Classifying web hierarchically using multi label tree classifier, 2015 Annual IEEE India Conference (INDICON), 2015
[16] Prabhu, Yashoteja, Manik Varma, FastXML: a fast accurate and stable tree-classifier for extreme multilabel learning, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014
[17] E. Lee, J. Kang, J. Choi, and J. Yang., Topic-specific web content adaptation to mobile devices,e 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pages 845-848. IEEE Computer Society, 2006

Authors Profile

Mr.R.Khan is pursuing Master of Technology in Computer Science & Engineering Department of Sarvepalli Radhakrishnan University, Bhopal (MP). He focuses on the field of computer security to strengthen computer security learning and importance.

Mr.R.K.Gupta is a well known Professor of Computer Science & Engineering Department of Sarvepalli Rdhakrishnan University, Bhopal (MP). He is pursuing Ph.D. in Computer Science from Barkatullah University Bhopal. His research interest and specialization falls upon intrusion Detection System and various Computer Security Strategies.

Dr.V.Namdeo Head of Computer Science & Engineering Department of Sarvepalli Radhakrishnan of Sarvepalli Rdhakrishnan University, Bhopal (MP). She is equiped with huge knowledge and experience of various fields of Computer Science fields.