An Improved Optimized Web Page Classification using Firefly Algorithm with NB Classifier (WPCNB)

Khushboo Bhatt
M. Tech Scholar
Computer Science and
Engineering Department
UIT,BU Bhopal (M.P)

Anju Singh
Assistant Professor
Computer Science and
Application Department
UTD,BU Bhopal(M.P)

Divakar Singh
Associate Professor
Computer Science and
Engineering Department
UIT,BU Bhopal (M.P)

ABSTRACT

The web is a huge repository of information which needs for accurate automated classifiers for Web pages to maintain Web directories and to increase search engines' performance. In web page classification problem each term in each HTML/XML tag of each Web page can be taken as a feature, an efficient methods to select best features to reduce feature space of the Web page classification problem derived here. Classification of Web page content is essential to many tasks in Web information retrieval such as maintaining, web directories and focused crawling. The uncontrolled nature of Web content presents additional challenges to Web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process. As in derived work reviewed in Web page classification, the importance of these Web-specific features and algorithms, describe state-of-the-art practices, and track the underlying assumptions behind the use of information from neighboring pages. This work, our aimed to optimize best features selection for Web page classification problem. Since Firefly Algorithm (FA) is a recent nature inspired optimization algorithm, that simulates the flash pattern 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 FA is used for clustering on benchmark problems which is being found more suitable than Artificial Bee Colony (ABC), Particle Swarm Optimization (PSO), and other nine methods used. The web page optimization using Naïve Bayes classifier (WPCNB) is an improved optimized web page classification using firefly algorithm with NB classifier. this work is tested on research banking data set where firefly algorithm used for web optimization and Naïve Bayes (NB) classifier used for classification of pages in contrast to selected pages with reference to different fireflies. The entitled work is being found better in terms of feature measure(FM),accuracy, precision etc. parameters with respect to existing key concepts.it is also an search optimization approach and can be enhanced by different genetic algorithm(GA)based classifiers use in future.

General Terms

HTML, XML, Web page, Web Mining , websites .

Keywords

Feature, Navie Bayes classifier, j48, f- measure.

1. INTRODUCTION

Web page classification problem defined as the problem of the allocation of a Web page to one or more tags to predefined categories. The rapid growth 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[16].

As the popularity of the band increases, the amount of information on the Web has also increased. This growth of information has led to the need for accurate and rapid classification of Web pages to improve search engine performance. 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 neighbors were developed. Among these methods, decision trees, and support vector machine are suitable for classification problems in which the number of features is low. Classification problem websites, by contrast, is a big problem since each term in every HTML or XML web page of each label can be considered an option [17].

1.1 Web mining

Web Mining is the application of data mining techniques to understand and explore the patterns and interesting use of Web data to meet the needs of Web-based applications. Usage data captures the identity or origin of Internet users and their browsing behavior on a website. Extraction with the band itself can be further classified according to the nature of the intended use of the data as[28]:

Web Server Data: The user logs are collected by the web server. Typical data includes the IP address of time, reference page and access [28].

Application Server Data: Commercial application servers are important to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track the different types of business events and store them in the newspapers of the application server [28].

Application Level Data: New types of events can be defined in an application, and logging can be activated to generate the stories of those specifically defined events. It should be noted, however, that many end applications require a combination of one or more of the techniques used in the above categories[28].

When extracting Web content information using web mining, there are four typical steps [28]:

1. Collect – fetch the content from the Web
2. Parse – extract usable data from formatted data (HTML, PDF, etc)
3. Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
4. Produce – turn the results of analysis into something useful (report, search index, etc)

1.2 Data mining

Data mining refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as 'Knowledge mining from data' or "Knowledge mining [30].

Data: Data are facts, figures or text that can be processed by a computer. Today, organizations accumulate large amounts of data in various formats and various databases more data. This includes [29]:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

Data mining consists of five major elements • Extract, transform, and load transaction data onto the data warehouse system. • Store and manage the data in a multidimensional database system. • Provide data access to business analysts and information technology professionals. • Analyze the data by application software. Present the data in a useful format, such as a graph or table [29].

1.3 Web page classification techniques

The ranking of web pages are developed on the basis of the text classification technique. Text representation and feature extraction are critical issues. Both text mining and information retrieval in the text classification indicating text information with the words characteristics quantified from texts. Better mutual information calculated to describe the structure of web pages accurately formula is proposed. This is accomplished by introducing the factor hyperlink to improve the accuracy of the classification of web pages. One type of Web pages more link hypertext factor studied in this paper. Hypertext link sections are given a weight value based on their contribution to the classification of texts first. Factor hyperlink is carried out with the inclusion of weight according to the text below. Classification accuracy improved after hyperlink factor in the formula set mutual information of the experience. How to find a more convincing thanks to a proven factor and practice the theory hypertext link is the problem should be solved the next due to the factor that you hyperlink in this experiment is done by balancing weight of eleven types tag. Web page categorization becomes a key technology in the transformation and organization of a mass of documents and data. The function is selected to improve the processing technology text hyperlink factor that believes in the maximum entropy model. Experiment found that the process is more efficient. Not only can you get the most consistent distribution but to ensure the accuracy and universality in the classification of the classification of the website as well. So the method of classifying the Web page based on maximum entropy model is a method of relatively perfect website ranking.[31]

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. So is important to combined with information from the websites of web analysis. It plays a vital role in reaching characteristics of Web pages using technology. 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. A more detailed classification can be achieved by considering the characteristic bands hyperlink optimized MEM parameter and characteristic function. Smoothing technique is used to solve the problem of scarce when the characteristic feature is constructed. Of course, the study of categorization of web pages based on MEM is far from being limited, for example, how to find an effective value should be studied further and the absolute value of Discounting directly impacts the accuracy of the classification. Accuracy in the process of pretreatment website threshold. How to find a more specific set of labels, while the construction of the characteristic function of pages considered for Ongoing identification of new words in this work. [33] Site accuracy classifiers can be improved by extracting the main channels with a method of clustering.

1.4 Classifiers

After selecting features of training web pages, several machine learning methods and classification algorithms can be applied for categorization of web pages. When a new web page is to be classified, the classifiers use the learned function to allocate the web page to particular categories.

There are different types of classifiers [34], such as:

A. Profile Based Classifiers

For profile based classifiers, a profile (or a representation) for each category is extracted from a set of training web pages that has been predefined as examples of the category. After training all categories, the classifiers are used to classify new web pages. When a new web page is to be classified, it is first represented in the form of a feature vector. The feature vector is compared and scored with profiles of all the categories. In general, the new web page may be assigned to more than one category by thresholding on those webpage-category scores and the thresholding methods used can influence the classification results significantly. In the case where a web page has one and only one category, the new web page is assigned to the category that has the highest resulting score. Examples of classifiers using this approach are Rocchio classifier, Support Vector Machine, Neural Network classifier, and Linear Least Square Fit classifier [34].

B. Rule Learning Based Classifiers

The one of the most expressive and human readable representations for learned hypotheses is sets of if-then rules. The most important property of rule induction algorithms is that they allow the interrelationships of features to influence the outcome of the classification.

In general, for rule learning based classifiers, the training web pages for a category are used to induce a set of rules for describing the category. A web page to be classified is used to match the conditions of the rules. The matched rules predict the class for the web page based on the consequents of the rules. The three representatives of the rule learning based classifiers are: Disjunctive Normal Form (DNF) rule, Association rule, and Decision tree.

C. Direct Example Based Classifiers

For such classifier, a web page that is to be classified is used as a query directly against a set of examples that identify categories. The web page is assigned to the category whose set of examples has the highest similarity with the web page. These classifiers are called lazy learning systems. For example, K-nearest-neighbors classifier is a representative.

D. Parameter Based Classifiers

Training examples are used to estimate parameters of a probability distribution in parameter based classifiers. For example, Probabilistic Naïve Bayes classifier [34].

1.5 Optimization techniques

An optimization algorithm is a procedure which is executed iteratively by comparing various solutions till an optimum or a satisfactory solution is found. There are so many optimization methods are available for feature extraction by using different operations like GA, ACO, PSO.a further description of algorithms.

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.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.

2. LITERATURE SURVEY

Optimization problem is one of the most difficult problems in the field of operational research. The goal of the optimization problem is to find all the variables resulting from the optimum value of the objective function, including all values that satisfy the constraints. Many new types of optimization algorithms have been explored. One is a type inspired by nature. Algorithms of this type are such as ant colony optimization (ACO) proposed by Marco Dorigo in 1992 has been successfully applied to the scheduling algorithm. ACO is inspired by the social behavior of ants to find food sources and the shortest paths to their colony, marked by its pheromone released. Another example of this type of algorithm is an optimization algorithm particle swarm (PSO) developed by Kennedy and Eberhart in 1995. PSO is based on the behavior of swarms of schools of fish and birds in nature. PSO has been successfully applied to a problem of forecasting of wind energy where it is estimated that wind power on the basis of two meta-heuristic intelligence attributes swarm. Firefly algorithm is another example. It is an algorithm based on population, inspired by the social behavior of fireflies [3]. Fireflies communicate by blinking. Dimmers fireflies are attracted to the brightest and move them to mate. FA is widely used to solve problems of reliability and redundancy [14]. **Firefly Algorithm**: There are three idealized rules incorporated into the original Firefly algorithm (FA) [14]:

i) All fireflies are unisex so that a firefly is attracted to all other fireflies.

ii) a firefly's attractiveness is proportional to its brightness seen by other fireflies, and so, for any two fireflies, the dimmer firefly is attracted by the brighter one and moves towards it, but if there are no brighter fireflies nearby, a firefly moves randomly and

iii) The brightness of a firefly is proportional to the value of its objective function.

2.1 Related Work

The rapid growth of Internet usage and advances in communication technology has led to a rapid increase in the amount of text information online. Consequently, it has become difficult to handle the large amount of information online. To solve this problem, many new techniques have been developed and used by search engines. Several tests are used to provide faster, more accurate results for users. One of the most important studies in this field is the text classification. Categorization or classification of text, which is widely used by search engines, is one of the fundamental techniques for the management and organization of text data. The purpose of text categorization is to classify documents in a number of presets using the characteristics of the categories of documents. Text classification plays a crucial role in many tasks of recovery and information management. These tasks are; information retrieval, information extraction, filtering materials and the hierarchical construction of Web directories [11, 12]. When the text classification focuses on web pages is termed as Web Ranking or rank webpage. However, the sites are different from text, and contain a lot of information such as urls, links; HTML tags are not supported by text documents. Because of this distinction, ranking Web is different from traditional text classification. Nature inspired techniques including genetic algorithm (GA) optimization ant colony (ACO) and particle swarm optimization have also been proposed (PSO) algorithms for text and web classification problems. Gordon (1988) used the gas to find the best illustrator document for each user specified document as used and the relevance judgments requests made during the recovery process. This is one of the first studies of gas is applied to the search domain of information. Chen and Kim (1995) proposed a hybrid system based on neural network genetic algorithm called MASCATO [13]. They used a GA to choose the best keywords that describe the documents selected by the user, and a neural network, weights keywords are determined. In addition, Boughanem et al. (1999) applied a technique based on GA to optimize document descriptions and improved formulations of consultation. Ribeiro proposed a web page classifier based on the extraction rule. Navie are used both Bayes and a genetic algorithm for classification. In their study, the fuzzy membership function works best with a classifier naive Bayes classifier with a GA based. Liu and Huang proposed a fuzzy semi-supervised classification algorithm based on a genetic algorithm. Both labels and documents labels are taken together to obtain a classifier. Each document is represented as a vector of weighted word frequency of TF-IDF, and stripping and removal of stop words are not used. HTML tags not taken into account and compared with the naive Bayesian classifier, and the improve result shown with the classification accuracy [12].

3. PROPOSED WORK

The entitle 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. On comparative study, as expected, NB classifier overall produce better result in terms of two parameters that is accuracy and F – measure.

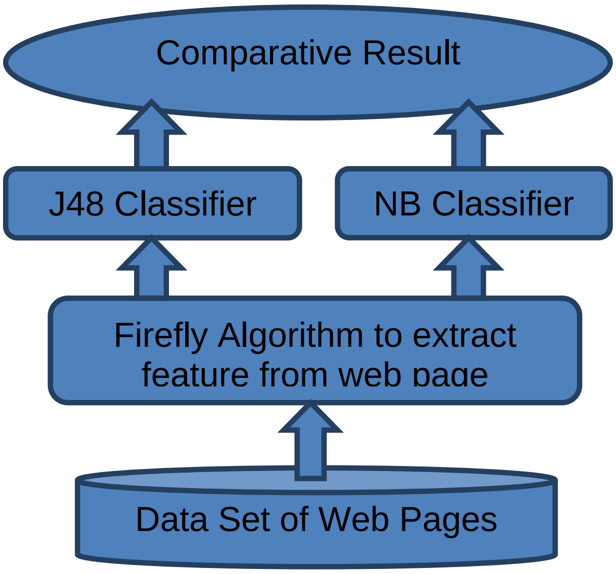

Figure.1 Block representation of Proposed Architecture

3.1 Proposed Algorithm

Algorithm for Web Page Classification by Firefly Optimization and NB classifier are as follows -

The proposed algorithm consists of four steps.

Step1 shows search and extraction of n features from web pages loaded from the data set in the form of html files. If a feature (f) is found in a web page write f or else leave empty space and following if condition ends. If sub features (fs) is found in web page write fs along f or else leave empty space and following if conditions ends. Now store the value of f and fs in fi as shown in (iii) and (iv) of step 1.

Step 2: Assigns an integer value 0-1 to each feature combination extracted from data set. If value of f or fs is 0 then value of function (fi) is also 0 following if condition ends here.if value of f is 0 and fs is 1 then the value of fi is Iv where Iv>o condition end here. Again if the value of fs is 0 and f is 1 then the value of fi is iv where iv>Iv. Lastly if value of f is 1 and fs is also 1 then the value of fi will be Ivi where Ivi is greater then I and Iv condition ends. Now the values obtained from above execution is store in Nd (Numerical data) and the Loop ends here.

Step 3 : Applies firefly algorithms on the numerical data extract from step 2. Firefly algorithm generate initial population of Firefilys as seen in (ii) of step 3.now the attractiveness is obtained and new solution are found and the light intensity for each generation (iteration) is updated. Finally with maximum light intensity is chosen as potential optimum solution and in this way Firefly optimized output is obtained.

Step 4 : NB classifiers Classifies the feature groups to obtain optimized data and a confusion matrix is created to get measuring parameters. Therefore the optimized output is executed. Execution of Proposed algorithm is as follows:

The required Assumption are :

f = Features, fs = sub features, m = no. of web pages, n = no. of features, Load Bank Search m webpage dataset.

```
STEP 1: Search and extract n features from 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
STEP 2 : Assign an integer value0-l to each feature combination
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
STEP 3 : Apply Firefly Algorithm
    i.     Pass Nd to Firefly Algorithm
    ii.    Firefly generate initial population of fireflies
            xi = LB + rand* (UB-LB)
(Where LB denotes lower bounds and UB denotes upper bounds of ith firefly)
    iii.   Obtain attractiveness, which varies with distance r
```

iv. Find new solutions and update light intensity for each generation (iteration) the firefly with maximum light intensity is chosen as potential optimum solutions

v. Got firefly optimized output

STEP 4 : Classification

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

3.2 Proposed flow chart

The proposed algorithm starts followed by loading of web page in the data set .then searching and extracting feature from each web page is done .if a feature or sub feature is found it is written in the Mat file and stored in an array then integer value to each feature combination is allotted and this numerical data is passed to firefly algorithm and this algorithm is applied on the above stated data. Then this firefly lot a number to each feature on priority bases.After applying the numbering to features the algorithm will provide number of iteration to each feature or sub feature. The firefly algorithm will optimize the numerical data (Nd) according to their brightness (availability). Now a firefly optimized output will be provided which will in turn create feature group. These feature groups will be subjected to Naïve Bayes classified model. A confusion matrix is made which compares feature group and NB predicts class and finally gives measuring parameters.

The above process is depicted below in the form of a flow chart fig.2

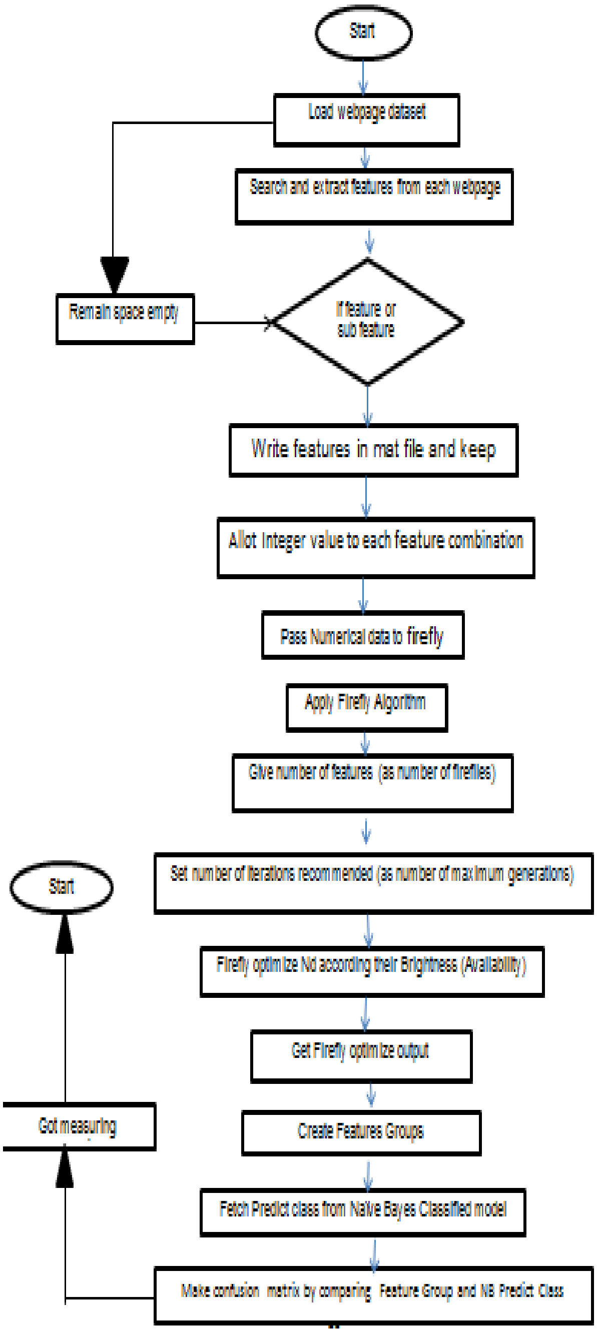
Start
Load webpage dataset
Search and extract features from each webpage
Remain space empty
If feature or sub feature
Write features in mat file and keep
Allot Integer value to each feature combination
Pass Numerical data to firefly
Apply Firefly Algorithm
Give number of features (as number of fireflies)
Start
Set number of iterations recommended (as number of maximum generations)
Firefly optimize Nd according their Brightness (Availability)
Get Firefly optimize output
Got measuring
Create Features Groups
Fetch Predict class from Naive Bayes Classified model
Make confusion matrix by comparing Feature Group and NB Predict Class

Figure 2. Proposed Flow Chart

4. RESULTS ANALYSIS

In the result stress should be given down that there are already lots of work is 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 .

Firstly they are applied with an optimized firefly algorithm with respect to correctness of corresponding features then they are classified by NB classifier. The tested result of previous work and proposed one is analyzed and observed by following parameters:

1. F-Measure
2. Accuracy

The test result shows that the F- measure value for base classifier is 0.709 and the accuracy percentage is 90.60% as compared to the WPCNB for which the F- measure value is 0.962 and accuracy percentage is 98.90%. 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 98% 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

F-MEASURE	ACCURACY Percentage	CLASS
0.709	**90.60**	**Base**
0.962	**98.90**	**WPCNB**

The Graphical Representation of the Result is shown in the Fig.3. Graphical representation shows the comparison between J48 and NB classifier on the basis of accuracy and F-measure parameters. The accuracy of NB classifier is 98.90% as compared to J48 classifier which is 90.60 %. F-measure parameter for NB classifier is 0.987 as compared to J48 classifier where it is 0.906.

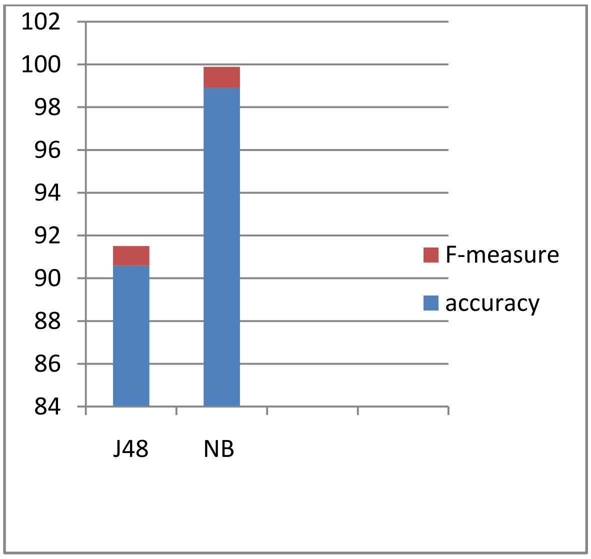

Figure 3. Comparison Graph with J48 and NB Classifiers

5. CONCLUSION

The conclusion of work is that although, there are already a lot of good as well as efficient methods are available in the web world but they are based on the different aspects and parameters refinement. That means they are able to detect and search those web contents whose features or their identification properties are already stored in the database analytical system of module. 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 servers 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. This work is tested over for different collections of data sets eg. 500 html files,1000 html files,1500 html files etc. here the result for different length data set comparatively better than the existing base method and it is performing till 98%(apprx.) in contrast of f-measure parameter. here the main concern of research was to enhance the classification and optimization using the different classifier in place of existing one and NB (Naïve Bayes) classified tested as better.

6. REFERENCES

[1] Esra Saraç, Selma Ayşe Özel ,”Web Page Classification Using Firefly Optimization”, IEEE, vol. 6, pp1-5, 2013.

[2] Amarita Ritthipakdee, Arit Thammano, Nol Premasathian, and Bunyarit Uyyanonvara, ”An Improved Firefly Algorithm for Optimization Problems”, ADCONP,HIROSHIMA, vol.4, pp 2-6,2014.

[3] Xin-She Yang, Xingshi He, “Firefly Algorithm:Recent Advances and Applications”, School of Science, Xi‟an Polytechnic University,Vol. 1, issue 1, 2013.

[4] Ben Choi and Zhongmei Yao,”Web Page Classification”, Louisiana Tech University, Ruston, LA 71272, USA,Vol.180,pp 221-224, 2008.

[5] Daniele Riboni,”Feature Selection for Web Page Classification” ,D.S.I., Universita‟ degli Studi di Milano, Italy,2009.

[6] Comparative Study of Firefly Algorithm and Particle Swarm Optimization for Noisy Non-Linear Optimization Problems I.J. Intelligent Systems and Applications, 2012.

[7] Adil Hashmi, Nishant Goel, Shruti Goel, Divya Gupta,”Firefly Algorithm for Unconstrained Optimization”, IOSR-JCE,2013.

[8] XIAOGUANG QI and BRIAN D. DAVISON, ”Web Page Classification: Features and Algorithms”,ACM, 2009.

[9] Xin-She Yang,M. Bramer et al. (eds.), “Firefly Algorithm”, L´evy Flights and Global Optimization Research and Development in Intelligent Systems Springer,2010.

[10] Selma Ayse Ozel, Esra Sarac, “Feature selection for web page classification using the intelligent water drops algorithm”,01330 Turke, 2011.

[11] Xin-She Yang , “Firefly Algorithms for Multimodal Optimization”,2010.

[12] Maybin K. Muyeba, Liangxiu Han, “Fuzzy Classification in Web Usage Mining using Fuzzy Quantifiers”, IEEE/ACM,2013.

[13] Sankalap Arora,Satvir Singh, “The Firefly Optimization Algorithm: Convergence Analysis and Parameter Selection”,IJCA,2013.

[14] Nikita Sahu, Dr. R. K. Kapoor, “A Review on Optimization in Web Page Classification”, IJAFRC 2014.

[15] Prabhjot Kaur , “Web Content Classification: A Survey”,IJCTT,vol.10, no.2,pp: 97-101, April 2014.

[16] Jie Chen,Jian Li,Hao Liao,Qingsheng Yuan, Xiuguo Bao, “Study on Meaningful String Extraction Algorithm for Improving Webpage Classification” IEEE,2011..

[17] Jiao Lijuan, Feng Liping,”Improvement of Feature Extraction in Web Page Classification” IEEE, 2010.

[18] JIAO Lijuan, Feng Liping, “Web Page Categorization based on Maximum Entropy Model” ,2010.