Computer Science and Engineering 2016, 6(1): 7-13
DOI: 10.5923/j.computer.20160601.02

Optimal Web Page Classification Technique Based on Informative Content Extraction and FA-NBC

A. M. James Raj*, F. Sagayraj Francis, P. Julian Benadit

Department of Computer Science Engineering, Pondicherry Engineering College, Pondicherry, India

Abstract 'Web Mining' refers to a group of techniques that derive interesting patterns of information from the World Wide Web. 'Web content mining', which is one of the Web mining techniques aims at retrieving interesting patterns of information from the raw data that exist in the Web pages. The source data primarily contain textual data in Web pages such as the words and their tags. General applications on these data are content-based categorization and content-based ranking. This paper proposes a method that is made up of three phases for classifying the Web pages, viz., feature extraction, information learning and classification. It first extracts object based features and utilizes these features to retrieve informative contents. Next it takes both terms and HTML tags at the same time on a Web page as features to extract the informative contents from the Web pages. The decision tree learning method is used to extract the rules from the features calculated. Based on the rules extracted, the Web pages are classified using optimal Firefly Algorithm (FA) based Naive Bayes Classifier (FA-NBC) in the final phase. Here the FA are used to optimize the rules extracted. The method was implemented in Java and its performance is compared with the existing classifiers. It is shown that this new method provides better performance than the existing classifier KNN.

Keywords Web Mining, Decision Tree, Firefly Algorithm(FA), Naïve Bayesian Classification (NBC), K-Nearest Neighbor (KNN)

1. Introduction

The World Wide Web is a huge repository of data that enables people all over the world to share and exchange their information. Although the World Wide Web provides numerous advantages to users, mostly it contains harmful, illegal or irrelevant contents [1]. In web browsing most of the search results, are obtained based on ranked list that gives irrelevant and inadequate information to the people when the search items are related to different titles. It is possible to retrieve more useful information from the Web very quickly if similar search results are grouped together [2]. Many studies have been conducted to classify related information and to support the manipulation of texts available on the Internet [3]. With the growing popularity of the Internet, and the availability of search tools and techniques like Web Servers, Browsers, Visual tools for Web page makers, dynamic HTML, Web-based databases and so on are making the web content into a heterogeneous in nature [4]. Many businesses and day today activities highly rely upon Internet applications and these applications have much vulnerability and typically are targeted by a large number of cyber-attacks due to their high exposure, access by browsers, and integration with databases [5].

The web, contains of volumes of web pages that increase every second and also beneficial to academicians, researchers, and general public. In such pages data are mixed with various contents like code, online advertisement, business links or Web page navigational links that make a Web page as a human friendly, but not a machine friendly [6]. Some of the pages are automatically created using general templates related to its contents for effective searching [7]. Google that services approximately 64.6% of traditional web search queries over 1 billion of Web documents has become the yardstick for web-search in terms of user interaction, response times, and prioritization of results.

Nowadays web sites are combined with more complicated information such as n-grams, HTML tags, JavaScript, URL information and Part-of-Speech (POS) tags [8] and categorizing these heterogeneous web documents into a set of meaningful classes which are predefined earlier is becoming increasingly challenging [9]. Moreover, Web page development strategies are also changing continuously from static to dynamic nature in order to increase the flexibility, user friendliness, and scalability.

With the drastic growth of Web based information, Web page classification process becomes one of the major challenges in organizing and maintaining such enormous collection of pages. Now-a-days many researches are in

* Corresponding author:
jamdiva05@yahoo.co.in (A. M. James Raj)

Published online at http://journal.sapub.org/computer

progress to enable the web categorization process to be a simple and effective one. This paper too is an attempt in that direction and proposes as novel method with verified efficiency.

This paper has been organized as follows: Section 2 outlines the related work in Web page classification. The proposed approach and the methods used in the approach are explained in section 3. The experimental results and its discussion are summarized with graphs are given in section 4. Finally, section 5 concludes the work and suggested ideas to further enhancement.

2. Related Work

Many Web page classifiers have been presented in literature over the years in which different perspectives have been taken to improve the performance of web classifiers. We found various kinds of classifiers developed on different perspectives in literature. They are Feature based classification (dimensionality reduction approach), Content based classification (word count in a Web page), Link based classification (by the association between different pages), Query log based (by the relationship among queries and Web pages) and Structure based classification (using the structure of the Web page, the images, links contained in the Web page) [10].

In recent years, the advent of the social media in which huge amounts of data posted by different online users, which have been the potential for researchers to classify the web data appropriately within a short span of time. So that, Web page classification plays a vital role in order to organize and maintain online documents automatically.

The K-Nearest neighbor, KNN in brief, is a sophisticated classifier which has been applied in many areas, especially in pattern recognition for over many years. A text document is given as an input to the KNN algorithm and it forms k-nearest possible groups. We have utilized the KNN classifier, for the comparative study of our proposed work.

Support Vector Machine (SVM) is one among the most popular kernel-oriented categorization method and provides good accuracy on many research applications. It is an efficient machine learning method developed by Vapnik et. al. and widely applied in text classification [11].

Naive Bayes classifiers are widely applied in machine learning and utilized it in our experiment. For huge amounts of training data, SVM, KNN shows higher performance, but SVM is very slow while training and at testing time, which again places an importance to the naive Bayesian classifier [12]. We have carefully implemented some of the most relevant features, extraction methods, according to explanations found in the corresponding papers and we have tested their performance with the existing classifiers.

The existing Web page classification system has some issues such as

i. The number of features that are considered in the classifier is very less

ii. Either HTML tags or terms are considered as the feature.

In order to overcome these drawbacks and to provide an innovative and enhanced classification model, we have proposed an optimal Web page classification technique based on a hybrid of Firefly- Naive Bayes algorithm.

3. Proposed Work

The proposed method comprises of three phases: feature extraction, information learning and classification. This approach first extracts object based features and utilizes these features to extract informative contents. Here our proposed method takes both terms and HTML tags at the same time on a Web page as features to extract the informative contents from the Web pages. The major flow of the proposed work is illustrated in Figure 1.

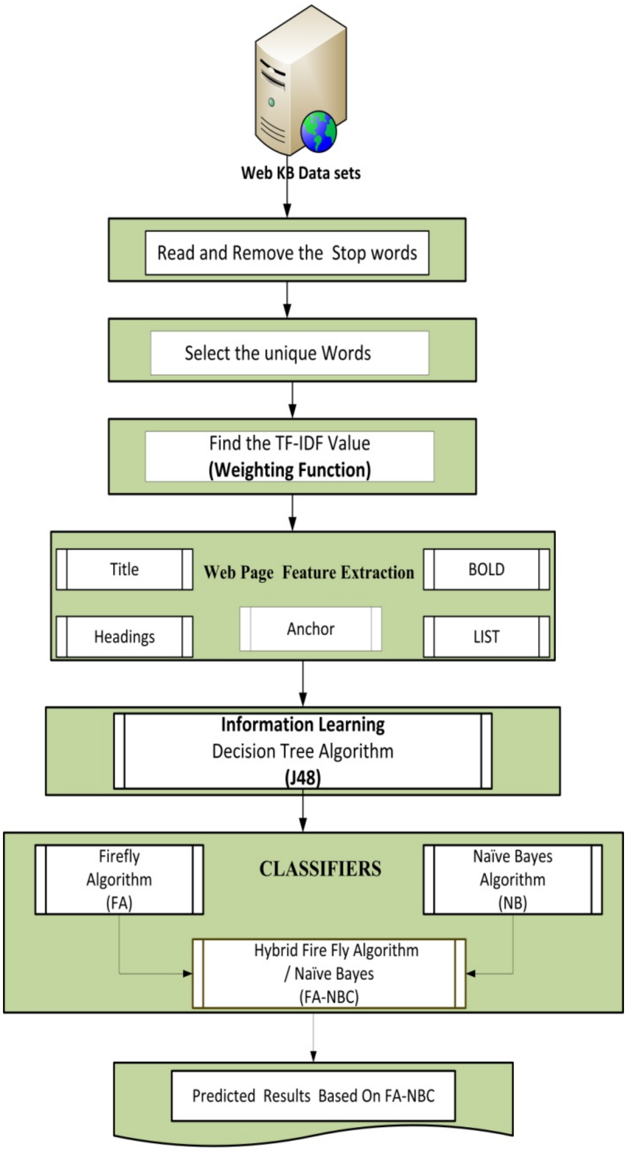

Figure 1. Proposed Working Flow of Hybrid Firefly Naïve Bayes Web Page Classification

The decision tree learning method is applied to extract the rules from the features calculated. Based on the rules extracted the Web pages are classified using optimal FA based Naive Bayes Classifier (FA-NBC). Here the FA is used to optimize the rules extracted. Initially, Web page data set is given as an input then read the file, remove problems, then unique words are identified from the Web pages, and then we will calculate TF-IDF value.

In feature extraction phase the features of Title, Heading, Bold, List, and Anchor are extracted and then converted the string value into a numerical value with the aid of 10×4 matrix. Based on the threshold value, classification is performed. The major three phases of our proposed method is explained in the following section.

3.1. Feature Extraction

A Web page is given as input, to the proposed model that removes the stop words and some unique words from the Web page and the TF-IDF (Term Frequency-Inverse Document Frequency) values are calculated to determine feature extraction process. The following five tags are considered as frequently used features and hence, that are extracted from the web page. They are Title, Headings, List, Bold, and Anchor.

Once the feature extraction process is completed, the decision tree method will be used to extract the rules for Information Learning. In the decision tree, learning method ID3 algorithm is applied. With the aid of information learning, decision tree is built for well-formed document. Finally, a hybrid Firefly - Naive Bayes classification is applied to obtain well- categorized documents. ID3 algorithm is called as Iterative Dichotomiser 3, developed by Ross Quinlan and it is widely used in the process of machine learning's. ID3 models are fast, accurate and easy to maintain in constructing training models in text classification. It is necessary to pick up an attribute that is noteworthy in the classification process. A statistical property is known as information gain that measures how significant a given attribute, splits the training data of their target classification method. This information gain is used as a measure in selecting from the candidate at each phase while constructing the tree.

Once a decision tree is built, it is used by each tuple in the dataset and the classification results are obtained. Most of the decision tree algorithms are following these steps

i. Selecting splitting attributes
ii. Ordering of the splitting attributes
iii. Total number of splits
iv. Balancing of tree structure and tree pruning
v. Stopping the process

The terms Information gain and Entropy, on which the ID3 algorithm revolves around are explained below.

Entropy

Given a set of probabilites $p_1, p_2, ...p_s$ where $\sum p_i = 1$, Entropy is defined as $H(p_1, p_{2,........}p_s) = \sum -(p_i \log p_i)$. Entropy calculates the quantity of order in a given dataset. When H = 0 means that a site is classified perfectly. The higher entropy value means that the classification process can be improved with the high potentiality.

Information Gain

Information gain *G(D, S)* is defined based on how much information is needed after the split *S*. By calculating the differences of the entropy of the original dataset *D* and the weighted sum of the entropies from each of the subdivided datasets Information gain is obtained. Information Gain is calculated using

$$G(D,S) = H(D) - \sum P(D_i).\ H(D_i) \qquad (1)$$

3.2. Classification Based on Firefly- Naive Bayes Algorithm

Firefly Algorithm

This new algorithm was first proposed by Xin-She Yang in 2008 and it adopted the behaviour of fireflies and the flashing patterns of them. It is a recent technique to find the best features for Web pages to make fast and accurate classification. It follows a population oriented iterative steps with collective agents (perceived as fireflies) and are used by different researchers to select from Web pages [13]. The basic principle of a firefly algorithm is described as follows [14].

> i. Fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex
>
> ii. The attractiveness is proportional to the brightness, and they both decreases as their distance increases. Thus, for any two flashing fireflies, the less brighter one will move towards the brighter one. If there is no brighter one than a particular firefly, it will move randomly.
>
> iii. The brightness of a firefly is determined by the landscape of the objective function.

As a firefly's attractiveness is proportional to the light intensity seen by adjacent fireflies, the variation of attractiveness can be defined as $\beta = \beta_o e^{-\gamma r^2}$, where β_0 is the brightness at distance r = 0 and γ is the light absorption coefficient.

Solution Representation

For optimal attribute selection in the decision tree construction, one of the most significant issues is how to symbolize a solution. The solution representation ties up with the firefly algorithm performance. We define one firefly (solution) as a possible solution in the population. The initial population of fireflies is constructed randomly for firefly algorithm. The initial population of size 'Y' is defined as

$$Y = A_d \tag{2}$$

Where, $(d = 1, 2, ..., N)$ and 'N' is the number of fire flies. The initialized continuous position values are generated by the following formula

$$u_k = u_{min} + (u_{max} - u_{min}) * r_e \tag{3}$$

Where, $u_{min} = 0$ and $u_{max} = 1$ and 'r' is a uniform random number between 0 and 1.

Fitness evaluation

The fitness function is defined based on our objective. In our work, an optimization formula is derived based on the minimizing the objective function.

$$W(y) = \min \sum_{i=1}^{m} w(y_i).H_x(y_i) \tag{4}$$

Where,

$H_x(y_i) \rightarrow$ Entropy .

$W(y_i) \rightarrow$ Weight of the Entropy for each attribute.

Firefly updating

The movement of the firefly p, when attracted to another more attractive (brighter) firefly q, is determined by

$$u'_p = u_p + \gamma(r) * (u_p - u_q) + \phi\,(\text{rand} - \tfrac{1}{2}) \tag{5}$$

The second term in the above equation is due to attraction, the third term introduces randomization with ϕ being the randomization parameter and "*rand*" is a random number generated uniformly distributed between 0 and 1.

$$\text{Attractiveness} \quad \gamma(r) = \gamma_o e^{-\theta r^m} \quad m \geq 1 \tag{6}$$

Where, r is the distance between two fireflies, γ_0 is the an initial attractiveness of firefly and θ is a absorption coefficient.

$$\text{Distance,} \quad r_{pq} = \left\| u_p - u_q \right\| = \sqrt{\Sigma_{k=1}^{d} (u_{p,s} - u_{q,s})^2} \tag{7}$$

Where, $u_{p,s}$ is the s^{th} component of the spatial coordinate of the $G(D,S) = H(D) - \sum P(D_i).H(D_i)$ firefly and 'd' is the total number of dimensions. Also, $q \in \{1, 2, ...F_n\}$ is randomly chosen index. Although q is determined randomly, it has to be different from p . Here, F_n is the number of fireflies. The procedure for an optimal attribute selection in decision tree using the firefly algorithm as follows.

i. An initial population of fireflies is randomly generated.
ii. The fitness of each firefly in the initial population is evaluated.
iii. Create a new population by replacing the firefly updating, as in (4) until the new population is complete.
iv. Using the newly generated population for the further sum of the algorithm.
v. If the test condition is satisfied, stop and return the best solution in the current population.
vi. Repeat step in (iii) until the target is met.
vii. Finally, obtain the optimal split point O_S to decision tree.

3.3. Naïve Bayes Method

The Naive Bayes method represents an easy probabilistic classifier dependent on the famous Bayesian theorem. It classifier is based on the configuration of a feature, independent probability model. It presumes that the existence of a specific feature of a class is not related to the existence of any other feature, for a specified class variable. One of its outstanding merits is that it does not require a huge size of samples for effective training. This classifier is the most widely used classifier in various web mining applications like web crawling, Web page classification and so on.

This Naive Bayesian classifier simplifies the calculation and provides a fast and accurate result in most of the classification methods. The formula to find the probabilistic Naïve Bayesian theorem is as follows

$$P(A \mid B) = \frac{P(B \mid A).P(A)}{P(B)} \tag{8}$$

Where, A and B are to stochastic events [15].

4. Experimental Results and Analysis

4.1. Dataset Description

The proposed system, is a hybrid of Firefly-Naive Bayes classifier and implemented using Java, NetBeans IDE 8.0 and JDK 1.7 and the experiments are done in an i7 processor with 5 GB RAM, by using the benchmarking WebKB data sets.

4.2. Performance Evaluation Metrics

The standard performance evaluation metrics generally used in web mining, such as Precision, Recall, F-measure and Accuracy are considered in our study and that are determined as the following formula.

Precision: It is defined as is the ratio of True Positive to the number of True and False positive value

$$\text{Precision(p)} = \frac{TP}{TP + FP} \tag{9}$$

Recall: It is defined as the number of correctly classified true positive instances divided by the number of positive instances in the data

$$\text{Recall (r)} = \frac{TP}{TP + FN} \tag{10}$$

F-measure: It can be employed in Information Retrieval to test the performance of the classifier. It is the harmonic mean of precision and recall. It can be calculated by the formula.

$$\text{F - Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} \times \text{Recall}} \tag{11}$$

Table 1. Confusion Matrix

	Confusion Matrix	**Predicted Positive**	**Predicted Negative**
1	Actual Positive	TP-True Positive	FN - False Negative
2	Actual Negative	FP-False Positive	TN - True Negative

Accuracy: It is the ratio of the total number of TP and TN of the total number of data.

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{12}$$

A confusion matrix is listed in Table-1, which can be used as the performance evaluation metrics discussed so far.

4.3. Results of the Proposed Method

The outcome of the proposed work helps us to examine the efficiency of the new classification method. Table-2 gives the results of proposed method.

Table 2. Results of the proposed Web page classification

Iteration	**Precision**	**Recall**	**F-measure**	**Accuracy**
50	0.93	0.72	0.857	84.05
100	0.94	0.75	0.872	88.51
150	0.96	0.76	0.879	91.63
200	1.0	0.79	0.919	94.84

The performance measures, namely, Precision, Recall, F-measures and Accuracy are graphically represented in figures 2 to 5 respectively.

4.3.1. Comparative Analysis

The existing methods from literature review are compared with the proposed work to show that the new hybrid of firefly-Naive Bayes classification work is better than the earlier works. It helps us to show better accuracy of the given Web page dataset. KNN classifier results are used to compare with the proposed method results. The results are summarized in the Table 3.

Table 3. Comparison of Precision, Recall and F-measures and Accuracy

Iteration	Precision		Recall		F-measure		Accuracy	
	KNN	FA - NBC	KNN	FA - NBC	KNN	FA - NBC	KNN	FA - NBC
50	0.88	0.93	0.68	0.72	0.837	0.857	70.29	84.05
100	0.89	0.94	0.72	0.75	0.843	0.872	71.82	88.51
150	0.90	0.96	0.73	0.76	0.867	0.879	76.94	91.63
200	0.93	1.0	0.74	0.79	0.872	0.919	78.84	94.84

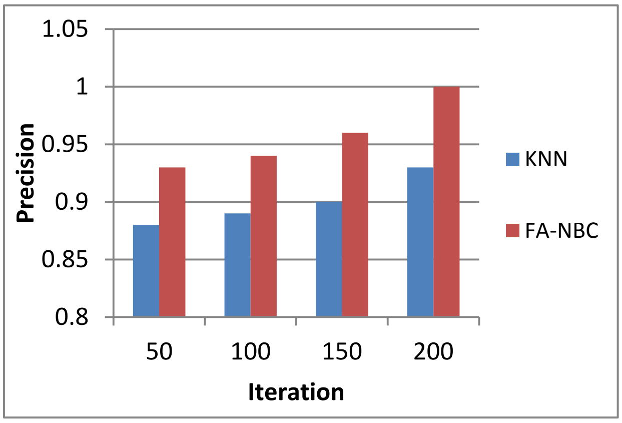

Figure 2. Comparison of KNN Vs FA-NBC in Precision

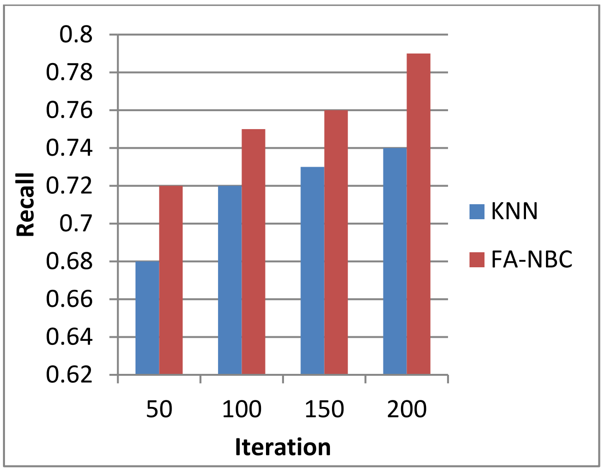

Figure 3. Comparison of KNN Vs FA-NBC in Recall

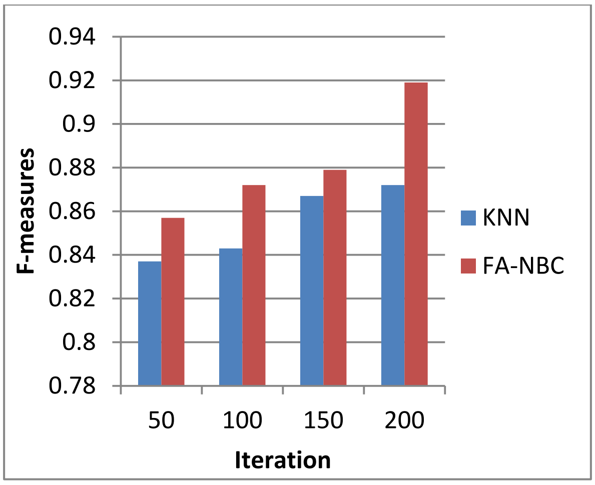

Figure 4. Comparison of KNN Vs FA-NBC in F-measures

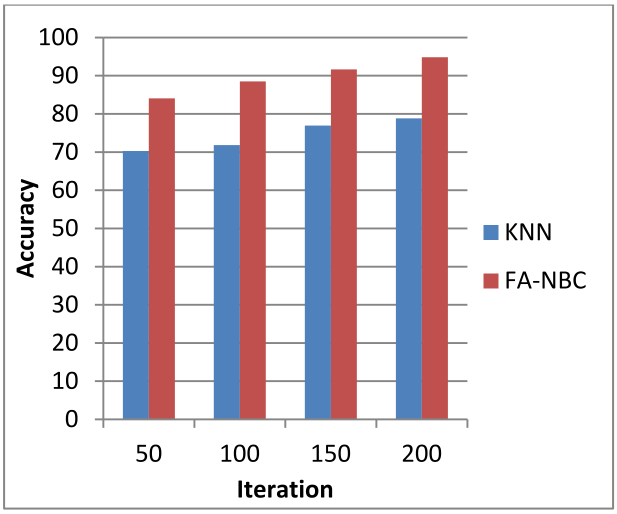

Figure 5. Comparison of KNN Vs FA-NBC in Accuracy

5. Conclusions and Feature Work

The hybrid Firefly-Naive Bayes classification method presented in this paper has three phases – feature extraction, information learning and classification. It first extracts object based features and utilizes them to extract informative contents. This method takes both terms and HTML tags at the same time on a Web page as features to extract the informative contents from the Web pages. The decision tree learning method is applied to extract the rules from the features calculated. Based on the rules extracted, the Web pages are classified using optimal FA based Naive Bayes Classifier (FA-NBC). The performance measures of precision, recall, f-measures were evaluated. The performance of the proposed classification model is high in terms of accuracy and web pages are classified effectively. From the consolidated results, we conclude that the hybrid of firefly-naive Bayes classification model provides a better performance than the other existing classifiers. As a sequel, more classification algorithms may be used to augment the FA-NBC method to obtain better results.

REFERENCES

[1] S. Weiming Hu Ou Wu, Zhouyaochen, Recognition of Web pages by Classifying Texts and Images, In Proceedings of IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 29, No. 6, 2007.

[2] Priya Venkateshan, “Clustering Web People Search Results using Fuzzy Ant- Based Clustering,” ELESVIER Sciences, vol. 20, pp. 569–571, Nov. 1999.

[3] Hanan M. Alghamdi, “Arabic Web pages clustering and an notation using semantic class feature,” Journal of king Saud University-Computer and Information Sciences , vol. 26,No.4. pp. 388–397, Dec. 2014.

[4] Chih-Ming Chen a, Hahn-Ming Lee, “Two Novel Feature selection approaches for Web page classification,” Journal of Expert Systems with Applications, vol. 26, No.4. pp. 260-272, Jan. 2009.

[5] Pir Abddul Rasool Qureshi and Nasrullah Memon, “Hybrid model of content extraction,” Journal of Computer and System sciences, vol. 78, No.4. pp. 1248–1257, July. 2012.

[6] Katerina Goseva and Popstojanova, “Characterization and Classification of malicious Web Traffic,” Journal of Computer and Security, vol. 42. pp. 92–115, May. 2014.

[7] Aidan Hogan and Andreas Harth, “Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine”, Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, No.4. pp. 365–401, Dec. 2011.

[8] Chulyun Kim and Kyuseok Shim, “TEXT: Auotmatic Template Extraction from Hetreogenous Web Pages”, in: Proc. IEEE Transaction and Data Engineering vol. 23, No.4. pp. 616–626, April. 2011.

[9] K. Pranitha Kumari and A. Venugopal Reddy, “Performance of Web page Genre classification”, International Journal of Computer applications, vol.53, No.10. pp. 24–27, Sept. 2012.

[10] A. Mangai, D. Kothari and V. S. Kumar, “A Navel Approach for Automatic Web page Classification using Feature Intervals”, International Journal of Computer Science Issue, vol.9, No.2. pp. 282–287, Sept. 2012.

[11] V. Vapnik, “The Nature of Statistical Learning Theory”, Springer-Verlag, NY, USA, 1995.

[12] Young Joong Ko and Jungyun Seo, “Text classification from unlabeled documents with bootstrapping and feature project techniques,” ELSEVIER Journal of Information Processing and Management, vol. 45, No.1. pp. 70–83, Jan. 2009.

[13] Shashank Dixit and R.K Gupta, “Layered Approach to classify Web Page using Firefly Feature selection by Support Vector Machin”, International Journal of u and e-Service, Science and Technology, vol. 8, No.5. pp. 355–364, Jan. 2015.

[14] X.S Yang, “Firefly algorithms for multimodal optimization”, in: Proc. 5th Symposium on Stochastic Algorithms Foundations and Applications, (Eds. O. Watanabe and T. Zeug mann) v Lecture Notes in Computer Science vol. 5792, pp. 169–178, April. 2009.

[15] Indira Mahadevan, Selvakumaran Karuppasamy and R. Rajaram, “Resource Optimization in Automatic Web page classification using integrated feature selection and machine learning,” International Arab Journal of e-Technology Computer applications, vol. 45, No.1. pp. 70–83, Jan. 2009.