International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:1, 2015

Feature Selection for Web Page Classification Using Swarm Optimization

B. Leela Devi, A. Sankar

***Abstract*—**The web's increased popularity has included a huge amount of information, due to which automated web page classification systems are essential to improve search engines' performance. Web pages have many features like HTML or XML tags, hyperlinks, URLs and text contents which can be considered during an automated classification process. It is known that Web-page classification is enhanced by hyperlinks as it reflects Web page linkages. The aim of this study is to reduce the number of features to be used to improve the accuracy of the classification of web pages. In this paper, a novel feature selection method using an improved Particle Swarm Optimization (PSO) using principle of evolution is proposed. The extracted features were tested on the WebKB dataset using a parallel Neural Network to reduce the computational cost.

***Keywords*—**Web page classification, WebKB Dataset, Term Frequency-Inverse Document Frequency (TF-IDF), Particle Swarm Optimization (PSO).

I. Introduction

THE increase in usage of Web and its growth are well known. Textual data on the Web is estimated at one terra byte, in addition to audio and video images which imposes new challenges to Web directories. Web directories enable users to search the Web, by classifying Web documents into subjects. Web pages manual classification suffers as Web documents increase [1].Text classification aims to categorize documents into a specific number of predefined classes using document features. Text classification has a crucial role in retrieval and management tasks like information extraction, information retrieval, document filtering, and building hierarchical directories [2]. When text classification focuses on web pages, it is called web classification or web page classification.

Classification assigns predefined class labels to unseen or test data. For this, a set of labelled data trains a classifier which then labels unseen data. Classification is supervised learning [3]. The process is not different in web page classification so that there are one or more predefined class labels. Classification model assigns class labels to web pages which are hypertext with many features like textual tokens, markup tags, URLs and host names in URLs. As web pages have additional properties, this classification differs from traditional text classification.

Web page classification has subfields like subject classification and functional classification. In the former, the classifier is concerned with web page content and determines the web page "subject". For example, online newspapers categories like finance, sport, and technology are examples of subject classification. Functional classification deals with function or type of a web page. For example, determining whether a web page is a "personal homepage" or a "course page" is functional classification. Subject and functional classification are popular classification types [2].

A HTML document's individual component is an HTML element made up of a tree of HTML elements and nodes like text nodes with all elements having specified attributes. Elements have content and text. HTML represents semantics or meaning [4]. HTML markup has key components, including character references, character-based data types, elements (and attributes), and entity references. Another component is document type declaration, triggering standards mode rendering. Semantic HTML is writing HTML emphasizing encoded information's meaning over presentation (looks). HTML includes semantic markup from inception, and presentational markup like <font>, <i> and tags. There are also semantically neutral tags.

HTML and associated protocols from inception were accepted quickly. But, there were no standards in the early years of language. Though HTML was conceived as a semantic language without presentation details, practical use pushed presentational elements and attributes to it, driven by varied browser vendors. Latest HTML standards are efforts to overcome chaotic language development and to create a rational foundation to build meaningful and well-presented documents.

Feature selection is an important classification step. Web pages are in HTML format meaning that web pages are semi-structured data, with HTML tags and hyperlinks in addition to pure text. Due to this web pages property, feature selection in classification is different from traditional classification. Feature selection reduces data dimension with tens or hundreds or thousands of features which cannot be processed further.

A major problem of web page classification is the high dimensionality of the feature space. Best feature subsets have least features that most contribute to classification accuracy. To improve web page classification performance, many approaches imported from feature selection or text classification were applied. Information gain [3], mutual information [5], document frequency [6], and term strength [7] are popular feature selection techniques. Information gain (IG) measures information in bits about the class prediction, when the only information available is a feature and corresponding

Leela Devi B is with Professional Group of Institutions, Palladam, Tamilnadu, India (e-mail: leeladevi_2008@rediffmail.com).
Sankar A is with PSG College of Technology, Coimbatore, Tamilnadu, India.

class distribution.

This study proposes a new feature selection technique using PSO algorithm for web page classification. Section II reviews related work in literature. Section III describes methods used and Section IV discusses the results of experiments. Section V concludes the paper.

II. Related Works

Automatic web page classification was emphasized by [8] through minimum features. They also proposed a procedure to generate optimum features for web pages. The optimum features model, machine learning classifiers. Experiment with a bench marking data set with such machine learning classifiers improved classification accuracy. To improve classification results with resources use, a full web page is not needed. As web pages contain high dimensions data, they are preprocessed to identify best representative features reflecting categories. A multilevel feature selection without bias to frequent terms was proposed. Results reveal the new feature selection process identifies lesser features with high information gain ensuring classification accuracy.

Previous works reveal a web page being partitioned into many segments or blocks. Blocks importance to a page is not equal. It was proved that differentiating noisy and unimportant blocks from pages helped web mining, search, and accessibility. No uniform approach was presented, in the same works, to measure important web page portions. A user study by [9] found people having consistent views on web page blocks importance. The new work investigates how to locate a model to automatically assign importance values to web page blocks. Block importance estimation as a learning problem is defined. First, Vision-based Page Segmentation (VIPS) algorithm partitions a web page into semantic blocks with hierarchy structure. Then spatial features like position, size, and content features including many images and links were extracted for feature vector construction in each block. Learning algorithms like SVM and NN train block importance models based on this. In the proposed experiments, best models achieve 79% performances with Micro-F1 and 85.9%, with Micro-Accuracy.

Web classification was tried via different technologies. Xhemali et al. [10] compared NN, NB, and DT classifiers for automatic analysis and classification of training course web pages attribute data. The study introduced an enhanced NB classifier and ran the same data sample through DT and NN classifiers to determine classifiers success rate in training courses domain. Research revealed that enhanced NB classifier outperformed traditional NB classifier. They performed well if not better than popular, rival techniques. The new study revealed that NB classifier is the best choice for training courses domain, achieving a F-Measure value of over 97%, despite being trained with fewer samples than classification systems encountered.

A graph-based semi-supervised learning algorithm applied to the Web page classification was proposed by [11]. The algorithm used a similarity measure between Web pages to construct a k-nearest neighbor graph. Preliminary experiments on a WebKB dataset showed that the new algorithm exploited unlabeled data as also labeled ones to get higher Web page classification accuracy.

The effect of considering named entities as web page classification features was investigated by [12]. The tests were in five different domains “baseball, football, health, politics and science "with web pages from online news providers. Results showed that incorporating named entities leads to slight gains in classifier performance for narrow domains, but is not true for all domains. Results showed that classification based on named entities can be good for some domains (baseball) but is lower than lexical terms based representation.

Saraç and Ozel [13] aimed to apply a recent optimization technique called Firefly Algorithm (FA), to choose best features for Web page classification. FA selected a features subset and to evaluate selected features fitness, the J48 classifier of Weka data mining tool was used. WebKB and Conference datasets evaluated the proposed feature selection system’s effectiveness. The result showed that when a features subset was selected by using FA, WebKB and Conference datasets were classified without accuracy loss, also as features decreased the time needed to classify the new Web pages, reduced.

A GA to select best features for Web page classification problem to improve accuracy and classifiers run time performance was proposed by [14]. The increased information on the Web has raised a need for accurate automated classifiers for Web pages to maintain Web directories and increase search engines' performance. To decrease feature space, a GA that determines best features for a set of Web pages was developed. It was found that when GA proposed features were used, and a kNN classifier was employed, accuracy went up to 96%.

A new centroid-based approach to classify web pages by genre using character n-grams from different information sources like title, URL, headings and anchors was proposed by [15]. To deal with web pages complexity and web genres rapid evolution, the new approach implemented a multi-label and adaptive classification scheme where web pages were classified singly, and each affected more than one genre. According to similarity between a new page and every genre centroid, the new approach either adapted the genre centroid under consideration or considered the new page as a noise page and discarded it. The results showed better results than current multi-label classifiers.

An approach of Web page classification using NB classifier based on ICA was proposed by [16]. To perform classification, a Web page was first represented by a features vector with different weights, and a weight calculated method was improved. As features were big, PCA selected relevant features from a preprocessing section as input for improved ICA algorithm (MFICA). Finally, MFICA output was sent to a NB classifier for classification to boost classifier performance. Evaluation proved that the ICA model based NB classifier ensured acceptable classification accuracy.

III. Methodology

WebKB dataset is used here for evaluation. Four different feature selection techniques were described and a new feature selection method using PSO algorithm was proposed.

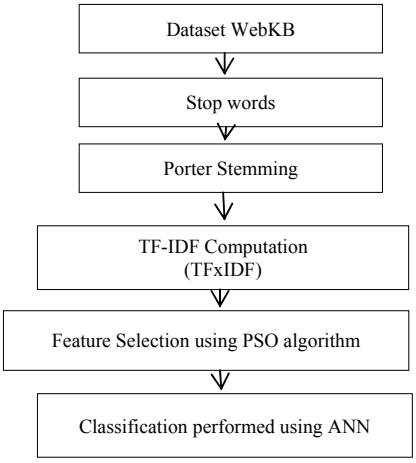

Fig. 1 Flowchart for the proposed method

WebKB Data Set

WebKB dataset [17] is a set of webpages collated by the World Wide Knowledge Base (WebKB) project of the CMU text learning group downloaded from The 4 Universities Dataset Homepage [18]. The pages are collected from various universities computer science departments in 1997 and manually classified to 7 different classes, including student, faculty, staff, department, course, project, and others. For every class, the collection has web pages from four universities; i.e. Cornell, Wisconsin Texas, and Washington universities, and miscellaneous pages from other universities.

All 8,282 web pages are classified manually into the seven categories so that the student category has 1641 pages, faculty 1124, staff 137, department 182, course 930, project 504, and others have 3764 pages. The class *other* is a collection of pages not deemed the "main page" and do not represent any instance of the earlier six classes. WebKB dataset includes 867 web pages from Cornell University, 827 pages from Texas University, 1205 from Washington University, 1263 from Wisconsin University, and finally 4120 miscellaneous pages from other universities. This study uses *Project, Faculty, Student,* and *Course* classes from the WebKB dataset. As *Staff* and *Department* classes have limited positive examples, they are not considered.

Stopwords

Stop words are filtered prior to, or after, natural language data processing. Stop words are commonly used words frequently filtered from text in information retrieval tasks. When removing stop words, noise are get ridded of, and space is saved to store documents. For example, consider an instance ""I am a student of computer science at Wisconsin University." The stopwords "I", "am", "a", "of", and "at" are left out of the full-text index. Thus on removal of the stopwords the instance is represented by "student computer science Wisconsin University".

Stemming

Stemming is a tool used in vocabulary mismatch problem, where query words do not match document words. Stemmers *conflate* certain variant forms of same word like (*paper, papers*) and (*hold, holds, holding…*) [19]. After removing high frequency words, indexing conflates word variants into same stem or root using a stemming algorithm. For example, words "engineering", "engineers" or "engineered" are reduced to the stem "engineer". Grouping words in information retrieval, with the same root under same stem (or indexing term) increases success rate when matching documents to a query [20].

Porter's algorithm is based on steps that each step removes a type of suffix by substitution rules. These rules only apply when specific conditions hold; the resulting stem must have a minimal length. Most rules have a condition based on a so-called measure. A measure is a number of vowel-consonant sequences (where consecutive vowels/consonants are counted as one) present in a resulting stem. This condition must prevent letters which resemble a suffix, but only a part of the stem is removed [21]. For example, "student, students" on Porter stemming is "student". Similarly, for "studied, studies, study, studying" is "studi".

Term Frequency-Inverse Document Frequency (TF-IDF)

Inverse Document Frequency (IDF) represents scaling factor. If term t occurs frequently in documents, its IDF value is less as term has lower discriminative power [22]. *IDF(t)* is defined as (1):

$$IDF(t) = \log \frac{1+|d|}{d_t} \tag{1}$$

d_t is a set of documents with term *t*. Similar documents have similar relative term frequencies. Similarity is measured among document sets or between a document and query. Cosine measure locates documents [23]; the cosine measure is got by (2)

$$sim(v_1, v_2) = \frac{v_1 . v_2}{|v_1| \ |v_2|} \tag{2}$$

where v_1 and v_2 are two document vectors, v_1. v_2 defined as $\sum_{i=1}^{t} v_{1i} v_{2i}$ and $v_1| = \sqrt{v_1 . v_1}$.

TF-IDF function weights each vector component (each relating to a word of a vocabulary) of every document as follows. First, it incorporates word frequency in a document. Thus, the more a word appears in a document (its TF, term frequency is high) the more it is thought to be significant in the document. Also, IDF measures how infrequent a word is in a collection. This is estimated using an entire training text collection at hand.TF–IDF combines weights of TF and IDF by multiplying them. TF gives more weight to a frequent term

in an essay while IDF downscales the weight, if a term occurs in many essays [24].

Feature Selection Using PSO

PSO originated from the simulation of birds social behavior in a flock. In PSO, a particle flies in search space with a velocity adjusted by its flying memory and companion's flying experience. A particle has its objective function value decided by a fitness function (3):

$$v_{id}^{t} = w \times v_{id}^{t-1} + c_1 \times r_1 \left(p_{id}^{t} - x_{id}^{t} \right) + c_2 \times r_2 \left(p_{gd}^{t} - x_{id}^{t} \right) \quad (3)$$

where i represents ith particle and d is a solution space dimension, c1 denotes a cognition learning factor, and c2 indicates a social learning factor, r1 and r2 are random numbers uniformly distributed in (0,1), p_{id}^{t} and p_{gd}^{t} stand for position with best fitness found so far for the *i*th and the best position in the neighborhood, v_{id}^{t} and v_{id}^{t-1} are velocities at time t and time t − 1, and x_{id}^{t} is position of *i*th particle at time *t*. Every particle moves to a new potential solution based on (4):

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t}, \quad d = 1, 2, ..., D, \quad (4)$$

A binary PSO where a particle moves in a state space restricted to 0 and 1 on each dimension, regarding changes in probabilities that a bit will be in one state or the other was proposed by [25] as in (5), (6):

$$x_{id} = \begin{cases} 1, & rand() < S(v_{i,d}) \\ 0 \end{cases}, \quad (5)$$

$$S(v) = \frac{1}{1+e^{-v}}. \quad (6)$$

Function S(v) is a sigmoid limiting transformation and rand ( ) is a random number from a uniform distribution in [0.0, 1.0].

This study uses a binary PSO algorithm version for PSO. Each particle's position is given in a binary string form representing a feature selection situation.

Feature Selection Using Proposed PSO

The process for proposed PSO is given by

1. Start PSO
2. Find gbest and Pbest
3. Each particle is updated according to (7)

$$v_{id}^{t} = w \times v_{id}^{t-1} + c_1 \times r_1 \left(p_{id}^{t} - x_{id}^{t} \right) + c_2 \times r_2 \left(p_{gd}^{t} - x_{id}^{t} \right) \quad (7)$$

4. Start random mutation hill climbing for gbest as follows:

Random mutation hill climbing is a local search method that has a stochastic component.

i. Choose a binary string at random. Call this string best_evaluated.
ii. Mutate a bit chosen at random in best_evaluated.
iii. Compute the fitness of the mutated string. If the fitness is greater than the fitness of best_evaluated, then set best_evaluated to the mutated string.
iv. If maximum number of iterations has been performed return best_evaluated_ otherwise, go to Step ii.

5. Start parallel hill climbing on the new pbest values.

Artificial Neural Network (ANN)

Realistic cognitive simulation and Computer Assisted Learning (CAL) inspired search for optimal learning and teaching methodology and classical teaching performance. ANN learning model use led to fair assessments performance of the suggested learning and teaching topics. So, optimal tutoring method is reached after analysis and evaluation of simulation results. Fig. 2 depicts an ANN learning and teaching model's block diagram which presents 2 diverse learning paradigms simulation. Both are related to interactive tutoring and learning process and self-organized learning. The first is related to classical (tutor supervised) learning seen in classrooms (face to face tutoring). This moves interactively through a bidirectional communication process between tutor and learner(s). The second paradigm undertakes self-organized (unsupervised) tutoring process.

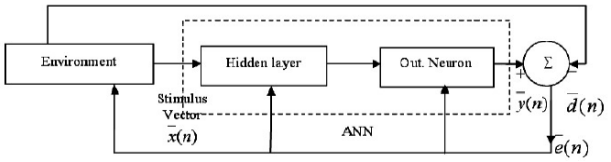

Fig. 2 Generalized ANN block diagram

Error vector referring to Fig. 2, at a time instant (n) observed during learning process is (8)

$$\bar{e}(n) = \bar{y}(n) - \bar{d}(n) \quad (8)$$

where e(n) is error correcting signal controlling learning process, adaptively, x(n) is input stimulus, y(n) is output response vector, and d(n) desired numeric value(s). Equation (9) is deduced:

$$\begin{aligned} V_k(n) &= X_j(n) W_{kj}^{T}(n) \\ Y_k(n) &= \varphi(V_k(n)) = (1 - e^{-\lambda V_k(n)}) / (1 + e^{-\lambda V_k(n)}) \\ e_k(n) &= |d_k(n) - y_k(n)| \\ W_{kj}(n+1) &= W_{kj}(n) + \Delta W_{kj}(n) \end{aligned} \quad (9)$$

where X is input vector, W weight vector, ϕ an activation (odd sigmoid) function characterized by λ as gain factor and Y as output. e_k is error value and dk the desired output. Noting that $\Delta W_{kj}(n)$ is a dynamic change of weight vector value connecting k th and ith neurons. The weight vector value dynamic changes for supervised phase are (10):

$$\Delta W_{kj} = \eta e_k(n) X_j(n) \quad (10)$$

where, η is learning rate value during learning process. But for unsupervised paradigm, dynamic weight vector value change is given by (11):

$$\Delta W_{kj} = \eta Y_k(n) X_j(n) \quad (11)$$

Note that $e_k(n)$ in is substituted by $y_k(n)$ at arbitrary time instant (n) during learning. The proposed parallel NN has 2 sub-networks. Every network has 2 hidden layers, with differing transfer functions. This study uses transfer function sigmoid and tanh. Different functions advantage is that mutual interference is reduced during complex tasks simultaneous processing and execution. Table I gives the proposed NN parameters.

TABLE I
PARAMETERS FOR THE PROPOSED MLP NN

Parameter	Value
Input Neuron	50
Output Neuron	4
Number of Hidden Layer	2
Number of processing elements upper	4
Number of processing elements lower	4
Transfer function of hidden layer upper	Tanh
Transfer function of hidden layer lower	Sigmoid
Learning Rule of hidden layer	Momentum
Step size	0.1
Momentum	0.7
Transfer function of output layer	Tanh
Learning Rule of output layer	Momentum
Step size	0.1
Momentum	0.7

TABLE II
THE TOP 15 WORDS SELECTED BY VARIOUS FEATURE SELECTION TECHNIQUES

CFS	MI	PSO	Proposed PSO
annual	professor	annual	annual
associate	science	aspects	assignment
compute	thu	assignment	associate
course	cornell	associate	component
define	universal	component	define
develop	return	define	direct
direct	compute	direct	people
geometric	annual	note	professor
hall	departmental	people	research
hour	research	professor	return
note	public	research	phd
overview	lecture	resume	universal
page	people	return	public
people	professional	phd	lecture
phd	phd	universal	note
annual	professor	annual	annual
associate	science	aspects	assignment
compute	thu	assignment	associate
course	cornell	associate	component

IV. EXPERIMENTAL RESULTS

The proposed PSO based feature selection for web page classification using HTML tags is calculated using the 4 Universities Dataset and compared with Correlation Feature Selection (CFS), Mutual Information (MI) and PSO feature extraction method. Five classes are classified (Student, Course, Faculty, Project, and Others). The accuracy, precision, recall and f measure are computed:

The top 15 words selected by various feature selection technique is tabulated in Table II.

The Neural Network classifies the web pages based on the keywords. The following tables and figures show the experimental results in detail. Tables III-V show the average precision, recall and F measure obtained for various feature extraction. The results are shown graphically from Figs. 3-6.

TABLE III
PRECISION

	CFS	MI	PSO	Proposed PSO
Student	0.8747	0.879	0.8823	0.8848
Faculty	0.883	0.8871	0.8924	0.8876
Course	0.7672	0.7525	0.8077	0.8402
Project	0.669	0.6596	0.7063	0.7809
Other	0.9128	0.8973	0.9237	0.9438

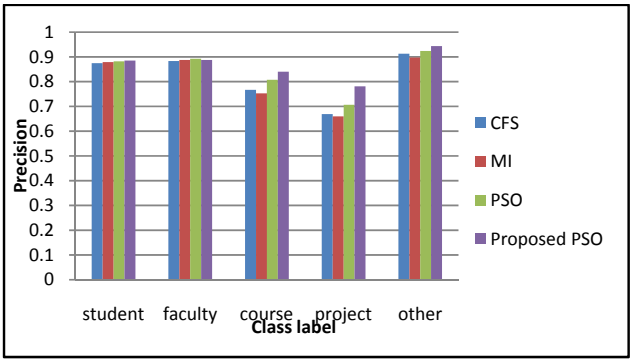

Fig. 3 Precision

From Table III and Fig. 3, it is observed that Precision is improved for Proposed PSO when compared to CFS, MI, and PSO. On an average, Precision increases for Proposed PSO by 2.13% when compared to PSO, by 6.23% when compared to MI and by 5.47% when compared to CFS. For class label course, precision of Proposed PSO increases by 3.94% when compared to PSO.

TABLE IV
RECALL

	CFS	MI	PSO	Proposed PSO
Student	0.7773	0.7404	0.8034	0.8492
Faculty	0.8749	0.8625	0.9025	0.909
Course	0.84	0.8203	0.8628	0.8785
Project	0.759	0.7294	0.7864	0.7982
Other	0.9218	0.9346	0.9275	0.9397

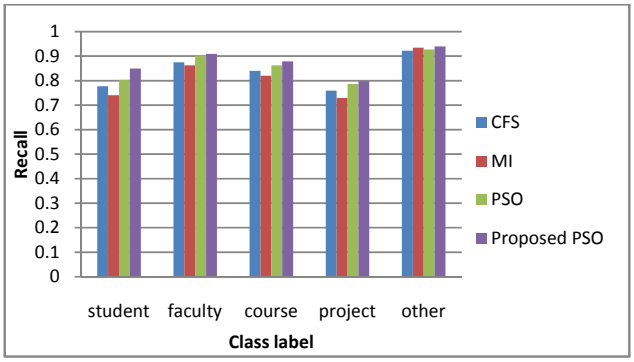

Fig. 4 Recall

From Table IV and Fig. 4 it is observed that Recall is improved for Proposed PSO when compared to CFS, MI, and PSO. On an average, Recall increases for Proposed PSO by 2.94% when compared to PSO, by 6.79% when compared to MI and by 4.71% when compared to CFS. For class label faculty, recall of Proposed PSO increases by 3.82% when compared to CFS.

TABLE V
F MEASURE

	CFS	MI	PSO	Proposed PSO
Student	0.8225	0.8037	0.8402	0.8667
Faculty	0.8784	0.8745	0.8975	0.8982
Course	0.8013	0.7856	0.8343	0.8595
Project	0.7116	0.6924	0.7448	0.7898
Other	0.9175	0.9162	0.925	0.9416

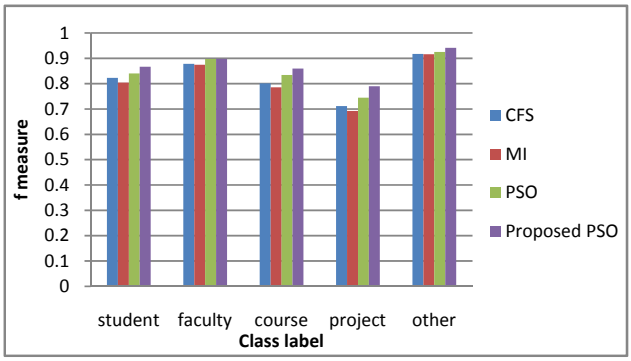

Fig. 5 f measure

From Table V and Fig. 5 it is observed that f measure is improved for Proposed PSO when compared to CFS, MI, and PSO. On an average, f measure increases for Proposed PSO by 2.65% when compared to PSO, by 6.72% when compared to MI and by 5.30% when compared to CFS. For class label student, f measure of Proposed PSO increases by 7.54% when compared to MI.

V. Conclusion

Automatic Web-page classification by using hypertext is a big approach to categorize large Webpage quantities. Two major approaches were studied for Web-page classification: content-based and context-based approaches. Content-based classification methods use words or phrases of a target document to build the classifier and achieve limited accuracy. This study proposed a new feature selection method using PSO algorithm. Results from experiments showed that the new method outperformed other feature selection methods and ensured good classification accuracy.

References

[1] Mangai, J. A., & Kumar, V. S. (2011). A Novel Approach for Web Page Classification using Optimum. IJCSNS, 11(5), 252.
[2] X. Qi and B. D. Davison, "Web page classification: features and algorithms," ACM Computing Surveys, vol. 41, no. 2, article 12, 2009.
[3] T. M. Mitchell, Machine Learning, McGraw-Hill, NewYork, NY, USA, 1st edition, 1997.
[4] Golub, K. and A. Ardo (2005, September). Importance of HTML structural elements and metadata in automated subject classification. In Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Volume 3652 of LNCS, Berlin, pp. 368–378. Springer.
[5] C. E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, vol. 27, pp. 379–423, 1948.
[6] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in Proceedings of the 14th International Conference on Machine Learning (ICML '97), pp. 412–420, Nashville, Tenn, USA, July 1997.
[7] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," Journal of Information Science, vol. 18,no. 1, pp. 45–55, 1992..
[8] Mangai, J. A., & Kumar, V. S. (2011). A Novel Approach for Web Page Classification using Optimum. IJCSNS, 11(5), 252.
[9] Song, R., Liu, H., Wen, J. R., & Ma, W. Y. (2004, May). Learning block importance models for web pages. In Proceedings of the 13th international conference on World Wide Web (pp. 203-211). ACM.
[10] Xhemali, D., Hinde, C. J., & Stone, R. G. (2009). Naive bayes vs. decision trees vs. neural networks in the classification of training web pages.
[11] Liu, R., Zhou, J., & Liu, M. (2006, October). Graph-based semi-supervised learning algorithm for web page classification. In Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on (Vol. 2, pp. 856-860). IEEE.
[12] Samarawickrama, S., & Jayaratne, L. (2012, September). Effect of Named Entities in Web Page Classification. In Computational Intelligence, Modelling and Simulation (CIMSiM), 2012 Fourth International Conference on (pp. 38-42). IEEE.
[13] Saraç, E., & Ozel, S. A. (2013, June). Web page classification using firefly optimization. In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on (pp. 1-5). IEEE.
[14] Ozel, S. A. (2011, June). A genetic algorithm based optimal feature selection for web page classification. In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 282-286). IEEE.
[15] Jebari, C., & Wani, M. A. (2012, December). A Multi-label and Adaptive Genre Classification of Web Pages. In Machine Learning and Applications (ICMLA), 2012 11th International Conference on (Vol. 1, pp. 578-581). IEEE.
[16] He, Z., & Liu, Z. (2008, October). A Novel Approach to Naïve Bayes Web Page Automatic Classification. In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on (Vol. 2, pp. 361-365). IEEE.
[17] Sun, A., Lim, E. P., & Ng, W. K. (2002, November). Web classification using support vector machine. In Proceedings of the 4th international workshop on Web information and data management (pp. 96-99). ACM.
[18] Kan, M. Y., &Thi, H. O. N. (2005, October). Fast webpage classification using URL features. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 325-326). ACM.
[19] Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002, August). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. In Proceedings of the 25th annual

international ACM SIGIR conference on Research and development in information retrieval (pp. 275-282). ACM.
[20] Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. JASIS, 50(10), 944-952.
[21] Kraaij, W., & Pohlmann, R. (1994). Porter's stemming algorithm for Dutch. Informatiewetenschap, 167-180.
[22] Papineni, K. (2001, June). Why inverse document frequency?. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-8). Association for Computational Linguistics.
[23] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2), 103-134.
[24] Soucy, P., & Mineau, G. W. (2005, July). Beyond TFIDF weighting for text categorization in the vector space model. In IJCAI (Vol. 5, pp. 1130-1135).
[25] Kennedy, J.; Eberhart, R.C., "A discrete binary version of the particle swarm algorithm", Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on Volume 5, 12-15 Oct. 1997 Page(s):4104 - 4108 vol.5.

Leela Devi B is with the School of Computer Applications at Professional Group of Institutions, Palladam. She is currently pursuing his doctorate in India.
Sankar A is currently working in PSG College of Technology, Coimbatore, India.