Journal of Intelligent Learning Systems and Applications, 2016, 8, 63-76
http://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

# Improved Term Weighting Technique for Automatic Web Page Classification

**Kathirvalavakumar Thangairulappan[1], Aruna Devi Kanagavel[2]**

[1]Department of Computer Science, V. H. N. S. N. College, Virudhunagar, India
[2]Department of Computer Science (PG), Kristu Jayanti College, Bengaluru, India
Email: kathirvalavakumar@vhnsn.edu.in, k.arunadeviselvi@gmail.com

**How to cite this paper:** Thangairulappan, K. and Kanagavel, A.D. (2016) Improved Term Weighting Technique for Automatic Web Page Classification. *Journal of Intelligent Learning Systems and Applications*, **8**, 63-76.
http://dx.doi.org/10.4236/jilsa.2016.84006

**Received:** July 7, 2016
**Accepted:** September 24, 2016
**Published:** September 27, 2016



## Abstract

Automatic web page classification has become inevitable for web directories due to the multitude of web pages in the World Wide Web. In this paper an improved Term Weighting technique is proposed for automatic and effective classification of web pages. The web documents are represented as set of features. The proposed method selects and extracts the most prominent features reducing the high dimensionality problem of classifier. The proper selection of features among the large set improves the performance of the classifier. The proposed algorithm is implemented and tested on a benchmarked dataset. The results show the better performance than most of the existing term weighting techniques.

## Keywords

Web Page Classification, Term-Weighting Scheme, Feature Selection, Feature Extraction, Artificial Neural Network, Back Propagation

## 1. Introduction

The rapid development of technology leads human beings and the devices to connect to internet and share the data. Thus the information is accumulating in WWW at a very high rate. In this scenario it is necessary to categorize the web contents in an organized way. Automatic classification of web pages into relevant categories helps the search engines to give quicker and better results. Web page classification [1] is the task of deciding whether a page belongs to a set of predefined category of document which is relevant to the topic. The web pages can be structured or unstructured or semi-structured and also they have heterogeneous contents which include text, images, audios, videos, links etc. The standard way of representing a web document for classification is Vector

space model that include documents and index terms as its components. Since a web page has no restrictions on the occurrences of a word it may have any number of words. The words in the web pages can have different length and different spelling also. The high dimensionality of web pages has its impact on performance of the classifier. Intensive researches are going on to reduce the dimension of web pages and to improve the classifier performance.

The goal of web page categorization is to classify the information on WWW into certain number of predefined categories. Categorization is an active research area in IR and machine learning. Several text categorization methods such as Naive Bayes [2], Rocchio [3], and Nearest Neighbor [4] and Back Propagation [5] have been proposed. The whole process can be done in three successive stages. In the first stage, the information is retrieved from the web and the features are to be extracted by analyzing the source of the web pages. In the second stage fix the input values of the neural network. Finally the third stage should determine the class of a certain web page out of predefined classes.

Ali Selamat and Sigeru Omatu [6] have proposed automatic categorization method that deals with the scaling problem of the World Wide Web. A news web page classification method (WPCM) uses a neural network with inputs obtained by both the principal components and class profile-based features. Each news web page is represented feature vectors and weighted using TF-IDF scheme. The principal component analysis (PCA) has been used to select the most relevant features out of many features for the classification. The final output of the PCA is combined with the feature vectors from the class-profile which contains the most regular words in each class. These feature vectors are then used as the input to the neural networks for classification.

Thabit Sabbah *et al.* [7] proposed Hybridized term weighting method for web contents classification using SVM. The hybridized method combines feature sets generated by the term weighting schemes TF, DF, TF-IDF, Glasgow and Entropy into one feature set for effective classification. Improved Web Page Identification Method using Neural Networks proposed by Ali Selamat *et al.* [8] is based on the improvement of feature selection of the web pages using Class Based Feature Vectors. The approach has been examined using the modified term weighting scheme. Gowri Shanthi and Antony Selvadoss Thangamani [9] have proposed an Enhanced Approach on Web Page Classification Using Machine Learning Technique that clean the dataset and transform it in to the pattern for classification. Then the feature extraction is performed to extract only minimum number of representative features or terms extracted from it without using the entire Web page. Then FP-Growth algorithm is used to classify the dataset into one of the seven classes. Alamelu Mangai, Santhosh Kumar and Appavu Balamurugan [10] have proposed a new classification model for web page classification called a probabilistic web page classifier (PWPC) which is based on a probabilistic framework and Attribute-Value Similarity measure (AVS).

Ruma Dutta, Anirban Kundu and Debajyoti Mukhopadhyay [11] have proposed a web page prediction model giving significant importance to the user's interest using the

clustering technique and the navigational behavior of the user through Markov model. The clustering algorithms considered are K-means and K-mediods, where K is determined by HITS algorithm. Finally, the predicted web pages are stored in form of cellular automata to make the system more memory efficient. Qiming Luo, Enhong Chenand HuiXiong [12] proposed a term weighting scheme by exploiting the semantics of categories and indexing terms. The semantics of categories are represented by senses of terms appearing in the category labels. The weight of a term is correlated to its semantic similarity with a category. Man Lan *et al.* [13] investigated several widely-used unsupervised and supervised term weighting methods on benchmark data collections in combination with SVM and kappa NN algorithms. They have proposed a new simple supervised term weighting method, *i.e.* tf.rf, to improve the terms' discriminating power for text categorization task.

Chris Buckley [14] concluded in his research article that Good weighting methods are more important than the feature selection process and it is suggested that the two need to go hand-in-hand in order to be effective. Kusum Kumari Bharti and Pramod Kumar Singh [15] used the feature selection methods term variance (TV) and document frequency (DF) for features' relevance score computation. Principal component analysis (PCA) is applied to further reduce dimensions in the feature space without losing much information. Behzad *et al.* [16] investigated two different kinds of feature selection metrics (one-sided and two-sided) as a global component of term weighting schemes (called as tffs) in scenarios where different complexities and imbalance ratios are available. They concluded that supervised term weighting methods based on one-sided term selection metrics are the best choice for SVM in the imbalanced datasets and k-NN algorithm usually perform well with tfidf. The construction of a text classifier [17] usually involves (i) a phase of term selection, in which the most relevant terms for the classification task are identified, (ii) a phase of term weighting, in which document weights for the selected terms are computed, and (iii) a phase of classifier learning, in which a classifier is generated from the weighted representations of the training documents. This process involves an activity of supervised learning.

In this paper feature vectors of the web pages are classified using a new term weighting scheme. As the weight of a term drives the classifier for efficient and accurate categorization, the weighted features are ranked and the top ranked features are selected for classification. Neural network classifier is used to classify the 20 Newsgroups dataset [18], a benchmarked dataset. The paper is organized as follows. The traditional term weighting schemes and the new term weighting scheme are introduced in Section 2. Classifier and its training method are discussed in Section 3. The Web page classification method is presented in Section 4. The results of the proposed method are compared with the standard weighting schemes TF, DF, TF-IDF, Glasgow and Entropy in Section 5.

## 2. Traditional and Proposed Term Weighting Scheme

In web page classification all the terms in the document will not be helping to identify

the class the web page belongs to. Hence it is necessary to select only the most relevant terms from a document. The weight [19] of the term represents how much the term contributes to the semantics of the document. The widespread term weighting methods are TF, DF, TF-IDF, Glasgow and Entropy. The Vector Space Model (VSM) is a very common and effective way to represent the collection of documents. In VSM model each document $d$ is represented as vector in the term spaces, $d=(t_1,t_2,\cdots,t_n)$, where $n$ is the size of vocabulary and the corresponding weight vector is $w=(w_1,w_2,\cdots,w_n)$. The weights $w_1,w_2,\cdots,w_n$ are the weights of the respective terms $t_1,t_2,\cdots,t_n$ computed using any term weighting scheme. The VSM model is a two dimensional matrix where rows represent the documents and columns represent the terms. The computed weight can be in the range of 0 to 1. The various weighting schemes are discussed and their formulae are tabulated in Table 1.

a) Term Frequency (TF): TF is based on the normalized frequency of a certain term. If a term has more occurrences then it has more implication.

b) Document Frequency (DF): Document frequency states the number of documents in the collection has the term t.

c) Term Frequency—Inverse Document Frequency (TF-IDF): In IDF the more occurred term in the collection is considered as least significant term. It is a global term weighting scheme in which the weight is computed with respect to its occurrence in the entire collection. TF-IDF is an eventual ranking measure in which the less frequent term in the collection but at the same time it is most occurred in a document is considered significant and vice versa.

Table 1. Various term weighting schemes and their mathematical formulae.

| Term Weighting Scheme | Formula |
|---|---|
| Term Frequency (*TF*) | $TF_{t,d}=\dfrac{fr_{t,d}}{\sqrt{\sum_{t=1}^{n} fr_{t,d}^2}}$ |
| Document Frequency (*DF*) | $DF_t=\sum_{d=1}^{N}\begin{cases}1 & t\in d\\ 0 & t\notin d\end{cases}$ |
| Term Frequency–Inverse Document Frequency (*TF-IDF*) | $TF\text{-}IDF_{t,d}=TF_{t,d}\cdot IDF_t$ where $IDF_t=\log\left(\dfrac{N}{DF_t}\right)+1$ |
| Glasgow | $w_{t,d}=\dfrac{\log(fr_{t,d}+1)}{\log(length_d)}x\left(\log\left(\dfrac{N}{DF_t}\right)+1\right)$ |
| Entropy | $w_{t,d}=L_{t,d}xG_t$ where $G_t=\dfrac{1+\sum_{j=1}^{N}\dfrac{fr_{t,d}}{F_t}\log\left(\dfrac{fr_{t,d}}{F_t}+1\right)}{\log N}$ and $L_{t,d}=\begin{cases}1+\log fr_{t,d}, & fr_{t,d}>0\\ 0, & fr_{t,d}=0\end{cases}$ |

d) Glasgow term weighting: This weighting scheme is introduced to avoid favoring the longer documents with lot of irrelevant words than the small documents.

e) Entropy term weighting: It is based on a probabilistic analysis of the text. The more frequent the term is contained in most documents is considered as more significant. It computes the term weight from two aspects, which are local term-weighting and global term-weighting. This means that, once every term receives a weight, it will be composed of local and global weights. It is calculated at the range of 0 - 1, hence values are normalized.

f) Proposed Term Weighting Scheme

The new term—weighting scheme is originally improved from traditional TF-IDF. The improved term-weighting considers three main factors—term frequency factor, collection frequency factor, and document length factor. The proposed term weighting formula is expressed as,

$$w_{t,d} = \frac{\log\left(TF_{t,d}+1\right)}{\log\left(\frac{DF_t}{N}+1\right)} \tag{1}$$

where $TF_{t,d}$ represents the Term Frequency and $DF_t$ represents the Document Frequency of the term.

The new term weighting technique is focused on the word occurrence in a single document and also its occurrence in the entire collection. The term value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the collection, which helps to adjust for the fact that some words appear more frequently in general.

## 3. Classifier and Its Training Method

Artificial neural network is used as the classifier to classify the web documents. ANN is composed of collection of neurons and layers. It usually consists of three layers which are input layer, hidden layer and output layer. A Single hidden layer feed-forward neural network in Figure 1 is used. Let $X = (x_i)$ be the input vector, $Y = (y_i)$ be the output vector, $H = (h_i)$ be the hidden neuron vector, $W = (w_{ij})$ be the weight matrix between input layer and hidden layer, and $V = (v_{ij})$ be the weight matrix between hidden

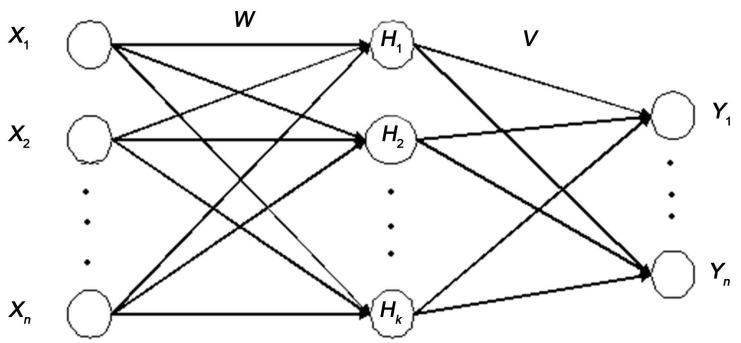


Figure 1. Feed forward neural network.

layer and output layer. In the input layer a bias node is included which with the constant value of 1.0. Bias node is added to increase the flexibility of the model to fit the data. It allows the network to fit the data when all input features are equal to 0.The weighted sum for neurons in hidden layer and output layer can be calculated by,

$$net_j^h = \sum_{i=1}^{n1} w_{ij} \cdot x_i \tag{2}$$

$$net_j^o = \sum_{i=1}^{n2} v_{ij} \cdot h_i \tag{3}$$

where n1 represents number of input neurons,

n2 represents the number of hidden neurons,

$net^o$ represents net value for output layer neurons and

$net^h$ represents net value for hidden layer neurons.

The training patterns are given as input to the network and the respective outputs of the output layer and hidden layer are computed. The Sigmoidal function is used as the activation function for hidden and output layer. It is given as follows,

$$f\left(net_i^l\right) = \frac{1}{1+\mathrm{e}^{-net_i^l}} \tag{4}$$

where l represents the layer.

The network is learned by minimizing the Mean Square Error (MSE). MSE is defined as

$$\mathrm{MSE} = \frac{1}{p}\sum_{p=1}^{p}\sum_{j=1}^{n}\left(d_j - o_j\right)^2 \tag{5}$$

where "$p$" represents patterns, "$j$" represents $j^{\text{th}}$ neuron of the output layer, "$d$" represents desired value and "$o$" represents obtained value.

**Standard Backpropagation**

The basic idea of the Standard backpropagation algorithm is the repeated application of the chain rule to compute the influence of each weight in the network with respect to an arbitrary error function. This algorithm is used on hidden layer to find optimal set of hidden layer weights and thresholds that minimizes error. The sum of squared error of the network is

$$E_p = \frac{1}{2}\sum_{j=1}^{nl}\left(e_{1j}\right)^2 \tag{6}$$

where the nonlinear error signal $e_1$ is, $e_{1j} = d_j - y_j$, $d_j$ and $y_j$ represent desired and obtained outputs for $j^{\text{th}}$ unit in the output layer.

$$e_{2i} = \sum_{j=1}^{n} e_{1j} f'\left(net_j^o\right) v_{ij} \tag{7}$$

Now the weight update rule for hidden layer is as follows,

$$\Delta w_{ij}^h = -\mu \frac{\partial E_P}{\partial w_{ij}} = \mu x_{ij} f'\left(net_i^h\right) e_{2i} \tag{8}$$

where $\mu$ is the learning rate, $h$ represents hidden layer.

Then the hidden layer weight will be updated by the following equation. $w_{ij} = w_{ij} + \Delta w_{ij}$.

## 4. Web Page Classification Method

The news web pages can be of different length, different structure and different vocabulary. The high dimensionality of the web pages diminishes the performance of the classifier. There are many categories of news in the web news pages such as sports, weather, politics, economy, computer etc. In each category there can be many different classes. The objective of this paper is to classify the web pages according to their category. The principal function of the web page classification is represented in **Figure 2.** The news web pages which are retrieved from www have text information, multimedia contents, html tags etc. The preprocessing phase removes the stop words like and, for, is etc. and applies stemming to reduce derived words to their root word like playing, played to play. After the stemming and stopping processes of the terms in each document, the web documents have only unique words and they are represented as the document-term frequency matrix as shown in **Table 2**. Each row in the table indicates the documents and columns represent terms. $Doc_j$ refers to each web page document that exists in the collection where $j = 1, 2, \cdots, m$ and $t_i$ refers to the unique terms in the collection where $i = 1, 2, \cdots, n$. Here $m$ is the total number of documents and $n$ is the total number of terms in the collection. The values in the table give the number of times the term $t_i$ occurred in $Doc_j$.

Since the numbers of terms are more in number it is necessary to recognize and extract the significant terms. To identify the significant terms, their term weights are computed using the proposed Formula (1). The term weights are sorted in ascending order and the top ranked $p$ numbers of features are selected. The top ranked $p$ feature profiles are stored in the Feature profiles. Feature selection is significant in supervised learning tasks such as binary or multiclass classification in order to improve the classifier efficiency and training time of the supervised classifier. In feature extraction phase the selected feature profiles are retrieved from the collection. The selected features are weighted using the new term weighting scheme. The selected features undergo normalization so that the data sets are scaled within the range of [1, −1]. To perform the normalization task, $E$ is represented as $\{e_1, e_2, \cdots, e_m\}$ as a $m \times p$ matrix, where $p$ is the number of term features and $m$ is the number of documents. $e_j$ denotes the $p \times 1$ vector containing p term features for each $e_j$. The normalized feature vector is given to ANN for classification. The feed forward back propagation artificial neural network categorizes the given vector and in the final stage the classifier performance is measured.

**Classification Algorithm**

The steps involved in the web page classification are given as follows. Given the Vector space model of web documents in matrix format to the algorithm and the output is the class label of the web documents.

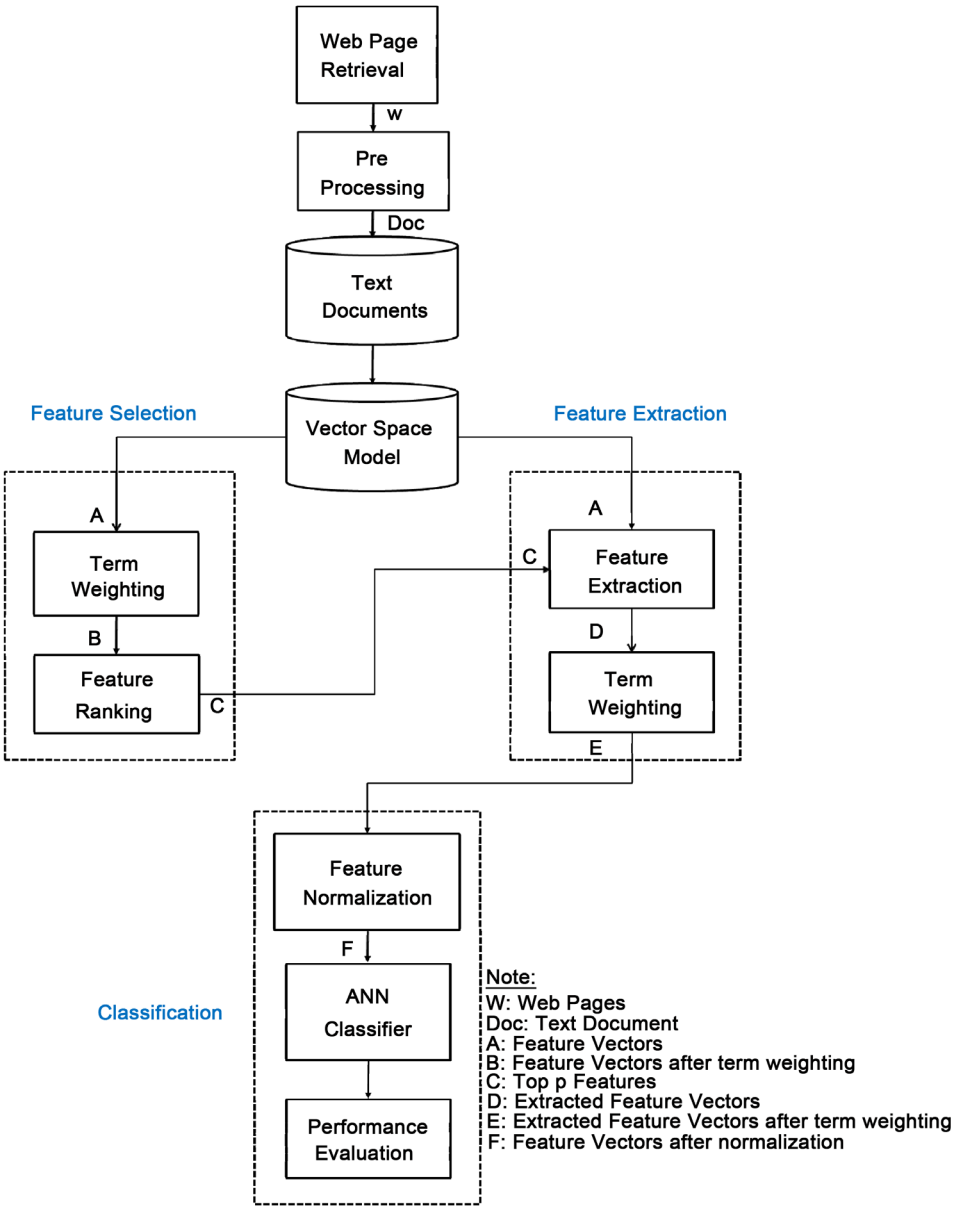


**Figure 2.** Web classification process

**Table 2.** The document-term frequency data matrix.

| Doc/Term | $t_1$ | $t_2$ | $t_3$ | … | $t_n$ |
|---|---|---|---|---|---|
| $Doc_1$ | 4 | 1 | 2 | … | 2 |
| $Doc_2$ | 2 | 3 | 3 | … | 1 |
| $Doc_3$ | 0 | 2 | 1 | … | 4 |
| : | : | : | : | : | : |
| $Doc_m$ | 3 | 4 | 4 | … | 0 |

Step 1: Start the training.

Step 2: Fix the desired number of features p to be given to classifier.

Step 3: Read the VSM of the web documents in matrix *A* of dimension m x n, where n is the no. of terms in the collection; m is the no. of documents in the collection.

Step 4: Compute term weight matrix *B* by the new term weighting scheme for each element in matrix *A*.

Step 5: Rank the terms by sorting the value in *B* from highest to lowest according to sum of term weight values.

Step 6: Select p number of highest value from B and store it in feature profile matrix *C*.

Step 7: Extract the term values from the collection corresponding to feature profile matrix *C* and VSM model and store significant features in matrix *D* of dimension m x p.

Step 8: Compute the term weight of matrix *D* using proposed formula and store the result in *E*.

Step 9: Normalize the term matrix *E* and results can be stored in matrix F.

Step 10: The normalized output matrix *F* of dimension $m$ x $p$ is given as input to the neural network classifier.

Step 11: Initialize the parameters for neural network training.

Step 12: For each input pattern compute output of the network applying standard back propagation approach

Step 13: Update the weights for the network.

Step 14: Calculate network error for the network with updated weights.

Step 15: Repeat the training, until the MSE reaches 0.001.

Step 16: Test the trained network with the testing documents.

Step 17: Measure the classifier performance using Accuracy, Precision, Recall and F1.

## 5. Experiments and Results

### 5.1. Data Sets

Experiments were done on benchmark dataset called 20 News groups [18]. This dataset is a collection of approximately 20,000 newsgroup documents with 60,000 terms, partitioned across 20 different newsgroups. For the analysis of the proposed work 3300 documents (m = 3300) of 5 different newsgroups are considered. **Table 3** depicts the details of various web page collections taken for the experiments. The number of features taken are 1200 (n = 1200).

### 5.2. Feature Extraction and Feature Selection

The best representative features are selected using new term weighting and ranking procedure. The features are weighted using the proposed formula then sorted and ranked according to the sum of weights. After feature ranking only 100 ($p$ = 100) significant features are extracted for classification. The experiments were run under the hardware and software configurations specified in **Table 4**.

Table 3. 20 Newsgroup dataset.

| Class | No. of documents taken for Training | No. of documents taken for Testing | Total no. of documents |
|---|---|---|---|
| Alt.atheism | 440 | 310 | 750 |
| Comp.graphics | 520 | 360 | 880 |
| Misc.forsale | 250 | 200 | 450 |
| Sci.Crypt | 370 | 180 | 550 |
| Rec.motorcycles | 410 | 260 | 670 |
| | Total | | 3300 |

Table 4. Hardware and software specifications.

| Hardware | Software |
|---|---|
| Processor: Intel Core Duo 2.1 GHz | Platform: MS Windows 8 |
| Memory: 3GB RAM; 32 bit OS | Software: Matlab R2014a |

### 5.3. Neural Network Classifier

The extracted p number features are then given as input to the neural network classifier. The feed-forward neural network with an input layer, a hidden layer and an output layer is used. The Back-propagation method is used for training the network. The network parameters are given in Table 5.

The learning rate is chosen based on trial basis for minimum cost. The weights are initialized randomly. The weight adjustments are drastic initially and then get stabilized. Thus the network learning is found to be smooth and is shown in Figure 3.

### 5.4. Evaluation

After the neural network is trained the test input patterns are given and the results of classifier are evaluated using standard information retrieval measurement tools. We evaluated using precision ($P$), recall ($R$), Accuracy ($Acc$) and $F1$. They can be expressed as follows:

$$P = \frac{a}{a+b} \tag{8}$$

$$R = \frac{a}{a+c} \tag{9}$$

$$F1 = \frac{2PR}{P+R} \tag{10}$$

$$Acc = \left( \frac{\text{Total Correct}}{\text{Total No.of Documents}} \right) \times 100\% \tag{11}$$

The values of $a$, $b$, $c$ and $d$ are explained in Table 6. The relationship between the system classification and the expert judgment is expressed using four values as shown in the table.

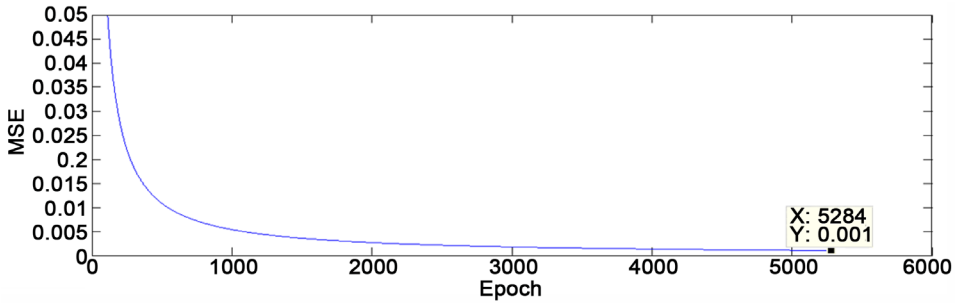


Figure 3. Epoch Vs MSE learning curve.

Table 5. Parameters for back-propagation neural network used.

| Parameter | Value |
|---|---|
| Learning Rate | 0.005 |
| Mean Squared Error | 0.001 |
| Input Neurons | n (No. of terms) |
| Hidden Neurons | 10 |
| Output Neurons | 5 |

Table 6. Definition of parameters a, b, c and d.

| Category Set | | Expert Judgment | |
|---|---|---|---|
| | | Yes | No |
| System Judgement | Yes | a | b |
| | No | c | d |

The classification accuracy of the term weighting methods TF, TF-IDF, DF, Glasgow, Entropy and the proposed method is tabulated in Table 7. On comparing the classifier accuracy of all the methods (Figure 4) the proposed scheme is found to be better. The overall classifier performance is tabulated in Table 8. With the new term weighting technique 97.69 percent web pages are classified accurately. The presented results show that the proposed solutions are reasonably accurate and fast. Through the proposed term weighting technique, we have succeeded in identifying the significant terms that are essentially needed for the classification. The classifier accuracy varies with respect to the class of the page also. In Rec. motorcycles class and Misc. for sale class the accuracy is comparatively low. The reason may be the page consists of more graphical contents than the text information.

## 6. Conclusion

With the volatile growth of the web pages it is necessary to identify the category of the pages. With the similarity between the pages and its different attributes, the classifiers have a tough time to make the decision about the category of the web pages. It is the terms in a page, which describes the class of a web page. All the terms may not be essential

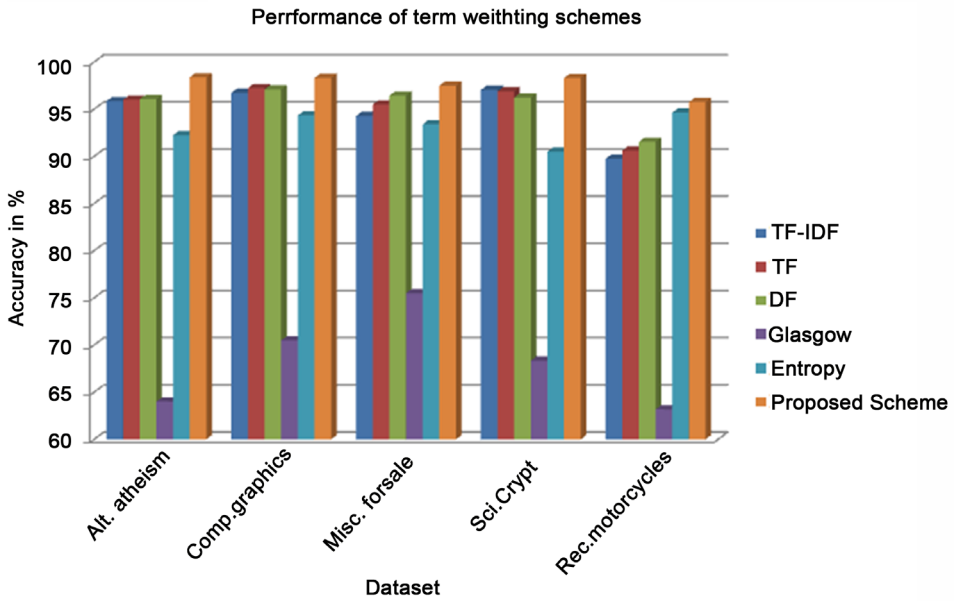


Figure 4. Comparison of classification accuracy.

Table 7. Classification accuracy of various term weighting methods.

| Category | TF-IDF | TF | DF | Glasgow | Entropy | Proposed Scheme |
|---|---|---|---|---|---|---|
| Alt.atheism | 95.86 | 96.01 | 96.08 | 64.01 | 92.25 | 98.3871 |
| Comp.graphics | 96.74 | 97.25 | 97.10 | 70.5 | 94.35 | 98.3333 |
| Misc.forsale | 94.29 | 95.5 | 96.45 | 75.5 | 93.4 | 97.5 |
| Sci.Crypt | 97.06 | 96.9 | 96.24 | 68.35 | 90.5 | 98.3146 |
| Rec.motorcycles | 89.75 | 90.62 | 91.55 | 63.18 | 94.65 | 95.7692 |
| **Average** | **94.74** | **95.256** | **95.484** | **68.308** | **93.03** | **97.68524** |

Table 8. Overview of performance of the proposed method.

| Category | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Alt.atheism | 98.35 | 100 | 95.81 | 97.86 |
| Comp.graphics | 98.1450 | 99.4 | 97.13 | 98.25 |
| Misc.forsale | 97.9126 | 100 | 96.93 | 98.44 |
| Sci.Crypt | 98.2121 | 99.5 | 96.33 | 97.89 |
| Rec.motorcycles | 95.8065 | 96.7 | 95.65 | 96.17 |

during the classification hence the significant term selection and retrieval is mandatory. In this paper a new term weighting method is proposed that identifies the important and unique term in a web page. As only few significant terms fed to the classifier, the results are very accurate and efficient. The experiments were conducted on different classes of 20 News group dataset and the comparative results show the average results are better than the existing methods. The performance of this method is comparatively

good for text based web documents, but when the page has more graphical contents than text contents the system performance goes low.

## References


[1] Qi, X.G. and Davison, B.D. (2009) Web Page Classification: Features and Algorithms. *ACM Computing Surveys*, **41**, 12:1-12:31.

[2] McCallum, A. and Nigam, K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. *Proceedings in Workshop on Learning for Text Categorization*, AAAI'98, 41-48.

[3] Lewis, D.D., Schapire, R.E., Callan, J.P. and Papka, R. (1996) Training Algorithms for Linear Text Classifiers. *Proceedings of* 19*th International Conference on Research and development in Information Retrieval*, ACM, New York, 289-297. http://dx.doi.org/10.1145/243199.243277

[4] Yang, Y., Slattery, S. and Ghani, R. (2002) A Study of Approaches to Hypertext Categorization. *Journal of Information Systems*, **18**, 2-3.

[5] Kamruzzaman, S.M. (2006) Web Page Categorization Using Artificial Neural Networks. *Proceedings of the* 4*th Intl Conf. on Electrical Engg. &* 2*nd Annual Paper Meet*, Bangladesh, January 2006, 26-28.

[6] Selamat, A. and Omatu, S. (2004) Web Page Feature Selection and Classification Using Neural Networks. *Information Sciences*, **158**, 69-88. http://dx.doi.org/10.1016/j.ins.2003.03.003

[7] Selamat, A., Lee, Z.S., Maarof, M.A. and Shamsuddin, S.M. (2011) Improved Web Page Identification Method using Neural Networks. *International Journal of Computational Intelligence and Applications*, **10**, 87-114. http://dx.doi.org/10.1142/S1469026811003008

[8] Sabbah, T., Selamat, A., Selamat, M.H., Ibrahim, R. and Fujita, H. (2016) Hybridized Term Weighting Method for Web Contents Classification using SVM. *Neuro Computing*, **173**, 1908-1926. http://dx.doi.org/10.1016/j.neucom.2015.09.063

[9] Shanthi, S.G. and Thanamani, A.S. (2012)Enhanced Approach on Web Page Classification Using Machine Learning Technique. *International Journal of Advanced Research in Computer Engineering & Technology*, **1**, 278-282.

[10] Alamelu Mangai, J., Santhosh Kumar, V. and Appavu Balamurugan, S. (2013) A Novel Approach for Effective Web Page Classification. *International Journal of Data Mining Modelling and Management*, **5**, 233-245. http://dx.doi.org/10.1504/IJDMMM.2013.055860

[11] Dutta, R., Kundu, A. and Mukhopadhyay, D. (2011) Clustering-Based Web Page Prediction. *International Journal of Knowledge and Web Intelligence*, **2**, 257-271. http://dx.doi.org/10.1504/IJKWI.2011.045163

[12] Luo, Q.M., Chen, E.H. and Xiong, H. (2011) A Semantic Term Weighting Scheme for Text Categorization. *Expert Systems with Applications*, **38**, 12708-12716. http://dx.doi.org/10.1016/j.eswa.2011.04.058

[13] Lan, M., Tan, C.L., Su, J. and Lu, Y. (2009) Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 721-35. http://dx.doi.org/10.1109/TPAMI.2008.110

[14] Buckley, C. (1993) The Importance of Proper Weighting Methods. *Proceedings of the workshop on Human Language Technology—HLT*'93, Association for Computational Linguistics, Stroudsburg, 349-352. http://dx.doi.org/10.3115/1075671.1075753

[15] Bharti, K.K. and Singh, P.K. (2015) Hybrid Dimension Reduction by Integrating Feature

Selection with Feature Extraction Method for Text Clustering. *Expert Systems with Applications*, **42**, 3105-3114. http://dx.doi.org/10.1016/j.eswa.2014.11.038

[16] Naderalvojoud, B., Bozkir, A.S. and Sezer, E.A. (2014) Investigation of Term Weighting Schemes in Classification of Imbalanced Texts. *Proceedings of European Conference on Data Mining* (*ECDM*), Lisbon, 15-17 July 2014, 39-46.

[17] Debole, F. and Sebastiani, F. (2003) Supervised Term Weighting for Automated Text Categorization. *Proceedings of the* 18*th ACM Symposium on Applied Computing* (*SAC* 2003), Melbourne, 9-12 March 2003, 784-788. http://dx.doi.org/10.1145/952532.952688

[18] 20 Newsgroup Dataset. http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

[19] Ko, Y. (2012) A Study of Term Weighting Schemes Using Class Information for Text Classification. *Proceedings of SIGIR*'12, Portland, Oregon, 12-16 August 2012, 1029-1030.