Neural Networks for Web Page Classification Based on Augmented PCA

Ali Selamat and Sigeru Omatu
Division of Computer and Systems Sciences, Graduate School of Engineering,
Osaka Prefecture University, Sakai, Osaka 599-8531, Japan.
Telephone: +81-722-54-9278
Fax: +81-722-57-1788
Email: aselamat@sig.cs.osakafu-u.ac.jp, omatu@cs.osakafu-u.ac.jp

Abstract— **Automatic categorization is the only viable method to deal with the scaling problem of the World Wide Web (WWW). In this paper, we propose a news web page classification method (WPCM). The WPCM uses a neural network with inputs obtained by both the principal components and class profile-based features (CPBF). Each news web page is represented by the term-weighting scheme. As the number of unique words in the collection set is big, the principal component analysis (PCA) has been used to select the most relevant features for the classification. Then the final output of the PCA is augmented with the feature vectors from the class-profile which contains the most regular words in each class before feeding them to the neural networks. We have manually selected the most regular words that exist in each class and weighted them using an entropy weighting scheme. The fixed number of regular words from each class will be used as a feature vectors together with the reduced principal components from the PCA. These feature vectors are then used as the input to the neural networks for classification. The experimental evaluation demonstrates that the WPCM method provides acceptable classification accuracy with the sports news datasets.**

I. Introduction

Neural networks have been widely applied by many researchers to classify the text documents with different types of feature vectors. Wermeter [1] has used the document title as the vectors to be used for a document categorization. Lam et al. [2] have used the principal component analysis (PCA) method as a feature reduction technique of the input data to the neural networks. However, if some original terms are particularly good when discriminating a class category, the discrimination power may be lost in the new vector space after using the Latent Semantic Indexing (LSI) as described by Sebastini [3].

Here, we propose a web page classification method (WPCM), which is base on the PCA and class profile-based features (CPBF). Each web page is represented by the term frequency-weighting scheme. As the dimensionality of a feature vector in the collection set is big, the PCA has been used to reduce it into a small number of principal components. Then we augment the feature vectors generated from the PCA with the feature vectors from the class-profile which contains the most regular words in each class before feeding them to the neural networks for classification. We have manually selected the most regular words that exist in each class and weighted them using an entropy weighting scheme [4]. *The CNN* [5] and *the Japan Times* [6] English sports news web pages have been used for the classification purpose. The Bayesian, TF-IDF, and WPCM methods have been used as a benchmark test for the classification accuracy. The experimental evaluation demonstrates that the proposed method provides an acceptable classification accuracy with the sports news datasets.

The organization of this paper is as follows: The news classification using the WPCM is described in Section II. The preprocessing of web pages is explained in Section III. The comparisons of web pages classification using the Bayesian, TF-IDF, and WPCM methods are discussed in Section IV. The discussions on the web pages classification results using the WPCM, Bayesian, and TF-IDF apporaches are described in Section V. In Section VI, we will conclude the classification accuracy by using the WPCM compared with other methods.

II. News Classification Using the WPCM

The news web pages have different characteristics where the text length for each of them is variable. Also the structures of the pages are different in the tags usage (i.e., XML, html, SGML tags, etc.). Furthermore, a huge number of distinct words exist in those pages as there is no restriction on a word usage in the news web pages discussed by Hisao et al.[7]. The high dimensionality of *the CNN* and *the Japan Times* news web pages dataset has made the classification process difficult. This is because there are many categories of news in the web news pages such as sports, weathers, politics, economy, etc. In each category there are many different classes. For example, the classes that exist in the business category are stock market, financial investment, personal finance, etc. Our approach is based on the sports news category of web pages. In order to classify the news web pages, we propose the WPCM which uses the PCA and the CPBF as the input to the neural networks. Firstly, we have used the PCA algorithm [8] to reduce the original data vectors to a small number of relevant features. Then we combine these features to the CPBF before inputting them to the neural networks for classification as shown in Fig. 1.

A. Preprocessing of Web Pages

The classification process of a news web page using the WPCM method is shown in Fig. 1. It consists of a web news retrieval process, stemming and stopping processes, a feature reduction process using our proposed method, and a web classification process using error back-propagation neural networks. The retrieving process of sports news web pages has been done by our software agent during night-time [9]. Only the latest sports news web pages category will be retrieved

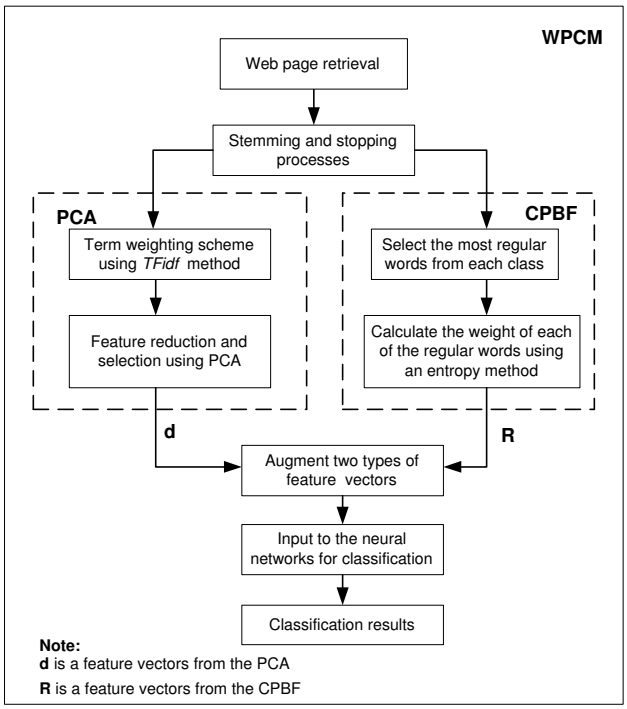

Fig. 1. The process of classification a news web page using the WPCM method.

TABLE I
THE DOCUMENT-TERM FREQUENCY DATA MATRIX AFTER THE STEMMING AND STOPPING PROCESSES.

Doc_j	TF_1	TF_2	...	TF_m
Doc_1	2	4	...	5
Doc_2	2	3	...	2
Doc_3	2	3	...	2
Doc_4	2	6	...	1
Doc_5	4	3	...	3
⋮	⋮	⋮	⋮	⋮
Doc_n	1	3	...	7

from *the CNN* and *the Japan Times* web servers from the WWW. Then these web pages will be stored in the local news database. Stopping is a process of removing the most frequent word that exists in a web page document such as 'to', 'and', 'it', etc. Removing these words will save spaces for storing document contents and reduce time taken during the search process. Stemming is a process of extracting each word from a web page document by reducing it to a possible root word. For example, the words 'compares', 'compared', and 'comparing' have similar meaning with a word 'compare'. We have used the Porter stemming algorithm [10] to select only 'compare' to be used as a root word in a web page document. After the stemming and stopping processes of the terms in each document, we will represent them as the document-term frequency matrix $(Doc_j \times TF_{jk})$ as shown in Table I. Doc_j is referring to each web page document that exists in the news database where $j = 1, \ldots, n$. Term frequency TF_{jk} is the number of how many times the distinct word w_k occurs in document Doc_j where $k = 1, \ldots, m$. The calculation of the terms weight x_{jk} of each word w_k is done by using a method that has been used by Salton [11] which is given by

$$x_{jk} = TF_{jk} \times idf_k \tag{1}$$

where the document frequency df_k is the total number of documents in the database that contains the word w_k. The inverse document frequency $idf_k = log(\frac{n}{df_k})$ where n is the total number of documents in the database.

1) Feature reduction using the PCA: Suppose that we have A, which is a matrix with document-terms weight as below

$$\mathbf{A} = \begin{pmatrix} x_{11} & x_{12} & x_{1k} & \cdots & x_{1m} \\ x_{21} & x_{22} & x_{2k} & \cdots & x_{2m} \\ x_{31} & x_{32} & x_{3k} & \cdots & \ldots \\ \vdots & \vdots & \vdots & \ddots & \ldots \\ x_{n1} & x_{n2} & x_{nk} & \cdots & x_{nm} \end{pmatrix}$$

where x_{jk} is the terms weight that exist in the collection of documents. The definitions of $j, k, m,$ and n have been described in the previous paragraph. There are a few steps to be followed in order to calculate the principal components of data matrix A. The mean of m variables in data matrix A will be calculated as follows

$$\overline{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}. \tag{2}$$

After that the covariance of matrix $S = \{s_{jk}\}$ is calculated. The variance, s_k^2, is given by

$$s_k^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \overline{x}_k)^2 \tag{3}$$

where $k = 1, 2, \ldots, m$. The covariance, s_{ik}, is given by

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n} (x_{ji} - \overline{x}_i)(x_{jk} - \overline{x}_k) \tag{4}$$

where $i = 1, \ldots, m$. Then we determine the eigenvalues and eigenvectors of the covariance matrix S which is a real symmetric positive matrix. An eigenvalue λ and a nonzero vector $\mathbf{e}$ can be found such that, $S\mathbf{e} = \lambda\mathbf{e}$ where $\mathbf{e}$ is an eigenvector of S.

In order to find a nonzero vector $\mathbf{e}$ the characteristic equation $|S - \lambda I| = 0$ must be solved. If S is an $m \times m$ matrix of full rank, m eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_m)$ can be found. By using $(S - \lambda I)\mathbf{e} = 0$, all corresponding eigenvectors can be found. The eigenvalues and corresponding eigenvectors will be sorted so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. The eigenvector matrix is represented as $\mathbf{e} = [u_1\ u_2\ u_3\ u_4\ \ldots\ u_m]$. A diagonal nonzero eigenvalue matrix is represented as

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \ldots \\ 0 & 0 & 0 & \lambda_m \end{pmatrix}.$$

In order to get the principal components of matrix S, we will perform eigenvalue decomposition which is given by

$$S = \mathbf{e}\mathbf{\Lambda}\mathbf{e}^{\mathbf{T}}. \tag{5}$$

Then we select the first $d \leq m$ eigenvectors where d is the desired value, e.g., 100, 200, 400, etc. The set of principal components is represented as $Y_1 = \mathbf{e}_1^T x$, $Y_2 = \mathbf{e}_2^T x$, ..., $Y_d = \mathbf{e}_d^T x$. An $n \times d$ matrix M is represented as

$$\mathbf{M} = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1d} \\ f_{21} & f_{22} & \cdots & f_{2d} \\ f_{31} & f_{32} & \cdots & \cdots \\ & & \ddots & \\ f_{41} & f_{42} & & \cdots \\ f_{n1} & f_{n2} & \cdots & f_{nd} \end{pmatrix}$$

where f_{ij} is a reduced feature vectors from the $m \times m$ original data size to $n \times d$ size.

2) Feature selection using the CPBF: For the feature selection using the class profile-based approach, we have manually identified the most regular words that exist in each category and weighted them using an entropy weighting scheme [4] before adding them to the feature vectors that have been selected from the PCA. For example, the words that exist regularly in a baseball class are 'Ichiro', 'baseball', 'league', 'baseman', etc. Then a fixed number of regular words from each class will be used as a feature vectors together with the reduced principal components from the PCA. These feature vectors are then used as the input to the neural networks for classification. The entropy weighting scheme on each term is calculated as $L_{jk} \times G_k$ where L_{jk} is the local weighting of the term k and G_k is the global weighting of the term k. The L_{jk} and G_k are given by

$$L_{jk} = \begin{cases} 1 + \log TF_{jk} & (TF_{jk} > 0) \\ 0 & (TF_{jk} = 0) \end{cases} \tag{6}$$

and

$$G_k = \frac{1 + \sum_{k=1}^{n} \frac{TF_{jk}}{F_k} \log \frac{TF_{jk}}{F_k}}{\log n} \tag{7}$$

where n is the number of documents in a collection and TF_{jk} is the term frequency of each word in Doc_j as mentioned previously. The F_k is a frequency of the term k in the entire document collection. We have selected $R = 50$ words that have the highest entropy value to be added to the first d components from the PCA as an input to the neural networks for classification.

3) Input data to the neural networks : After the preprocessing of news web pages, a vocabulary that contains all the unique words in the news database has been created. We have limited the number of unique words in the vocabulary to $1,800$ as the number of distinct words is big. Each of the words in the vocabulary represents one feature vector. Each feature vector contains the document-terms weight. The high dimensionality of feature vectors to be as an input to the neural networks is not practical due to poor scalability and performance. Therefore, the PCA has been used to reduce the original feature vectors $m = 1,800$ into a small number of principal components. In our case, we have selected the value of $d = 400$ together with $R = 50$ features selected from the CPBF approach since this parameter performs better for web news classification compared to other parameters to be input to the neural networks. The loading factor graph for the accumulated proportion of eigenvalues are shown in Fig. 2. The value of d contribute 81.6% of proportions from the original feature vectors.

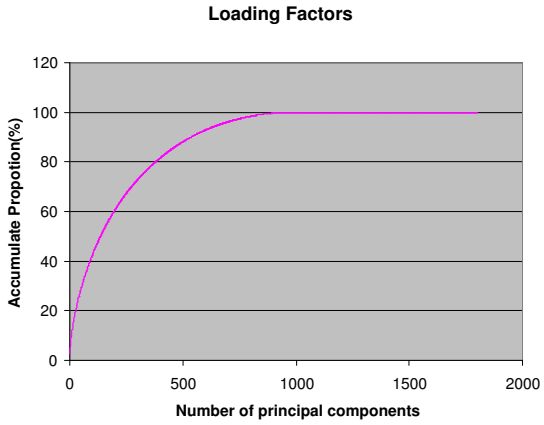

Fig. 2. Accumulated proportion of principal components generated by the PCA.

4) Characterization of the neural networks: The architecture of neural networks used for the classification process is shown in Fig. 3. The number of input layers (p) is 450 where principal components (d=400) and R (=50). The number of hidden layers (q) is 25. The trial and error approach has been used to find a suitable number of hidden layers that provide good classification accuracy based on the input data to the neural networks. The number of output layers (r) is 11 which is based on the number of classes in the sports news category as shown in Table II in the next page.

We have defined t as the iteration number, η is a learning rate, α is a momentum rate, θ_q is a bias on hidden unit q, θ_r is a bias on output unit r, δ_q is the generalized error through a layer q, and δ_r is the generalized error between layers q and r. The input values to the neural network are represented by $f_1, f_2, \ldots, f_{450}$. Adaptation of the weights between hidden (q) and input (p) layers is given by

$$W_{qp}(t+1) = W_{qp}(t) + \Delta W_{qp}(t+1) \tag{8}$$

where

$$\Delta W_{qp}(t+1) = \eta \delta_q O_p + \alpha \Delta W_{qp}(t) \tag{9}$$

$$\delta_q = O_q(1 - O_q) \sum_r \delta_r W_{rq}. \tag{10}$$

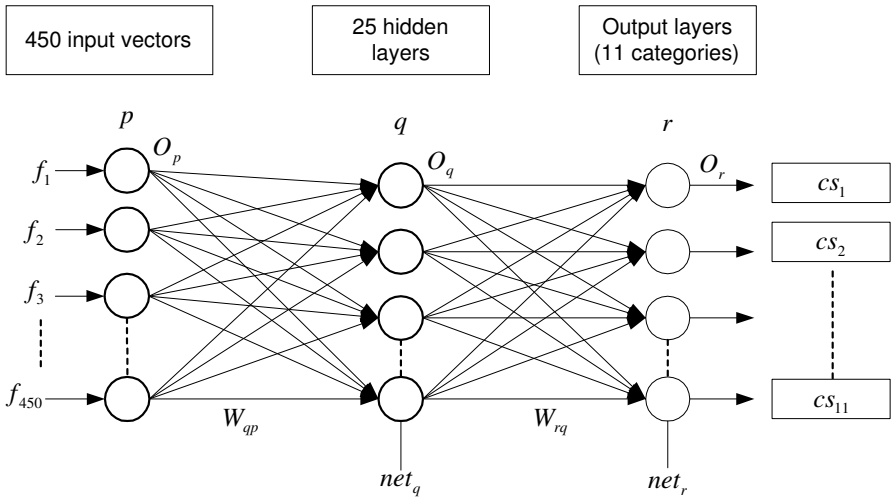

Fig. 3. The feature vectors from the PCA and CPBF are fed to the neural networks for classification.

TABLE II

THE NUMBER OF TRAINING AND TESTING DOCUMENTS THAT ARE STORED IN THE NEWS DATABASE. THESE ARE THE CLASSES THAT EXIST IN SPORTS CATEGORY.

Class (cs)	**Training documents** (T)	**Test documents**
1. baseball	100	101
2. boxing	70	12
3. cycling	84	40
4. football	50	10
5. golf	100	120
6. motor-sports	50	11
7. hockey	70	20
8. rugby	50	13
9. skiing	70	26
10. swimming	70	12
11. tennis	70	17
Total	784	382

Note that the first transfer function at hidden layer (q) is given by

$$net_q = \sum_q W_{qp} O_p + \theta_q \tag{11}$$

$$O_q = f(net_q) = 1/(1 + e^{-net_q}). \tag{12}$$

Adaptation of the weights between output (r) and hidden (q) layers is given by

$$W_{rq}(t+1) = W_{rq}(t) + \Delta W_{rq}(t+1) \tag{13}$$

where

$$\Delta W_{rq}(t+1) = \eta \delta_r O_q + \alpha \Delta W_{rq}(t) \tag{14}$$

$$\delta_r = O_r(1 - O_r)(t_r - O_r). \tag{15}$$

Then the output function at the output layer (k) is given by

$$net_r = \sum_r W_{rq} O_q + \theta_r \tag{16}$$

$$O_r = f(net_r) = 1/(1 + e^{-net_r}). \tag{17}$$

The error back-propagation neural networks parameters are set as in Table III.

TABLE III

THE ERROR BACK-PROPAGATION NEURAL NETWORKS PARAMETERS FOR TEST RUN1 AND RUN2.

NN Parameters	**Run1**	**Run2**
1. Learning rate (η)	0.05	0.005
2. Momentum rate (α)	0.01	0.001
3. Number of iteration (t)	1200	1000
4. Mean square error (MSE)	0.05	0.005

III. EXPERIMENTS

For the experiments, we have used a set of documents and classes as shown in Table II. There are 784 documents belonging to 1 or more classes. To select a set of documents for training, we have chosen 200 documents randomly to be a positive training set. For the negative examples, another 200 documents are selected randomly from another set of documents, which do not belong to the 11 classes set. The total number of training documents is 400 including from the negative examples and positives examples. We have used the same test as being done by Yang and Honavar [12], which includes the Bayesian and TF-IDF classifiers. Also we include

the WPCM approach as a comparison to the methods that are used in order to examine the applicability of the classification system. The description of WPCM has been mentioned in Section II. The TF-IDF and Bayesian methods for the tests are described as below.

A. TF-IDF measures

TF-IDF classifier is based on the relevant feedback algorithm by Rocchio using the vector space model [13]. The algorithm represents documents as vectors so that the documents with similar contents have similar vectors. Each component of a vector corresponds to a term in the document, typically a word.

The weight of each component is calculated using the term frequency inverse document frequency weighting scheme (TF-IDF) which tries to reward words that occur many times but in few documents. To classify a new document Doc', the cosines of the prototype vectors with corresponding document vectors are calculated for each class. Doc' is assigned to the class which its document vector has the highest cosine. Further description on the TF-IDF measure is described by Joachim [14].

B. Bayesian classifier

For statistical classification, we have used a standard Bayesian classifier [14]. When using the Bayesian classifier, we have assumed that term's occurrence is independent of other terms. We want to find a class cs that gives the highest conditional probability given a document Doc'. Let $w_k^m = \{w_1, w_2, \ldots, w_m\}$ is the words representing the textual content of the document Doc' and let k denote the term number where $k = 1, .., m$. The classification score is measured by

$$P(cs) = \prod_{k=1}^{m} P(w_k|cs) \tag{18}$$

where $P(cs)$ is the prior probability of a class cs, and $P(w_k|cs)$ is the likelihood of a word w_k in the class cs that is estimated on a labeled training document. A given web page document is then classified in a class that maximizes the classification score. If the scores for all the classes in the sports category are less than a given threshold, then the document is considered unclassified. The Rainbow toolkit [15] has been used for the classification of the training and test news web pages using the Bayesian and TF-IDF methods.

IV. Simulation Results

The results of individual simulations using the Bayesian, TF-IDF, and WPCM are shown in Table IV. The accuracies using the Bayesian, TF-IDF, and WPCM are 81.00%, 83.94%, and 84.10%, respectively. Also we have found that if the number of training and test documents are low, i.e., 10-80 pages, the accuracy of classification is less than 85%. The classification accuracies of the golf, rugby, skiing, swimming, and tennis classes using the WPCM approach is better compared to the Bayesian and TF-IDF approaches, which are 95.90%, 75.26%, 89.90%, 85.90%, 96.02%, respectively, as shown in Table IV. The CPBF feature selection process applied in the WPCM approach is based on the highest entropy of the keywords calculated as in (6) and (7). The first 50 keywords with the highest entropy values are selected and combined with the features taken from the PCA approach. The suitability of the keywords selection belonging to the particular classes has not been carefully considered in the WPCM approach while doing the experiments as described in Section III. For example, the entropies for the keywords 'famous', 'field', 'skills', and 'manager' from the class soccer are selected although these keywords also exist in the other classes such as the basketball and baseball classes. However, identifying keywords alone are not enough. The associated weights related to each of the keywords are also important to improve the classification performance using the proposed approach as the same word may occur in different classes. This is the main reason why the classification accuracies of the boxing and cycling classes (86.80% and 80.00%) are better when using the Bayesian approach compared with the WPCM approach as shown in Table IV. Furthermore, the stemming and stopping processes are also degrading the classification performance using the WPCM approach compared with the Bayesian approach. Also, the classification accuracies of the football and hockey classes (85.30% and 85.50%) are better when using the TF-IDF approach compared with the WPCM approach as shown in Table IV. This is because the contents of training documents belong to the football and hockey classes contain many sparse keywords. A better document selection approach needs to be used for selecting the candidate documents from each class in order to increase the classification results by using the WPCM approach. The results of the experiments for the error back-propagation neural networks parameters as in Table III are shown in Figs. 4 and 5. For the momentum rate $\alpha = 0.01$, we have found that the local minima exist. But if the value of $\alpha = 0.001$ is used, we have found that the MSE is smooth. For the rest of the experiments we have used the parameter $\alpha = 0.001$ as it indicates a stable MSE value to

TABLE IV

The percentage of the news web pages classification accuracy using the Bayesian, TF-IDF, and WPCM methods.

Docs. Class	Bayesian (%)	TF-IDF (%)	WPCM (%)
1. baseball	99.01	96.04	96.04
2. boxing	86.80	65.80	74.20
3. cycling	80.00	68.00	74.20
4. football	71.40	85.30	76.00
5. golf	96.67	95.00	95.90
6. hockey	80.21	85.50	78.86
7. motor sports	85.65	89.90	82.60
8. rugby	63.56	72.60	75.26
9. skiing	63.60	84.08	89.90
10. swimming	75.21	88.00	85.90
11. tennis	82.21	94.30	96.02
Average	**81.00**	**83.94**	**84.10**

be used for our classification process. We have also done an experiment on newsgroups datasets such as *alt.politics.people*, *alt.politics.mideast*, and *alt.politics.misc* [15] as a comparison. The results of classification using the WPCM on these datasets are shown in Table V. The average results of classifications are 87.82%. This indicates that, if the number of training datasets are more than 300 on each class, the possibility of getting a good classification performance is high. Although the WPCM approach provides an improvement of the classification results, the time taken to classify the web news pages is significantly long by using the the proposed approach compared to the other approaches.

TABLE V
THE CLASSIFICATION RESULT OF 3 NEWSGROUP DATASETS [15] USING THE WPCM METHOD.

Doc. Class	Training	Test	Result (%)
Alt.politics.people	300	1000	93.7
Alt.politics.mideast	300	1000	89.0
Alt.politics.misc	400	1000	80.8
Average	-	-	**87.82**

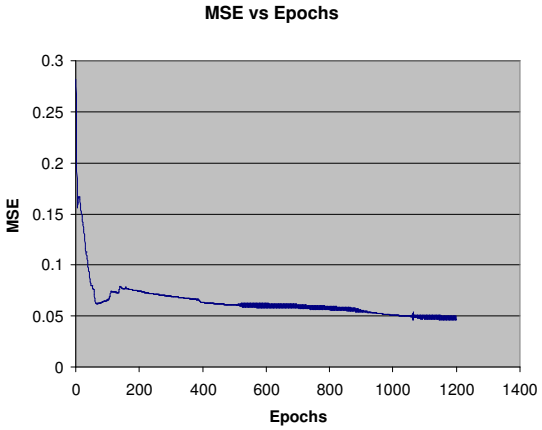

Fig. 4. The results of Run1 for MSE vs. Epochs for the news web pages classification using the neural networks (momentum rate, $\alpha = 0.01$).

V. CONCLUSIONS

We have presented a new approach of web news classification using the WPCM. In our approach, the pre-processing work needs to be done in the selection and calculation of feature vectors before the news web pages can be classified. The result is not accurate enough if the type of terms, i.e., baseball terms, football terms, etc., are not carefully chosen. The stemming process during the feature selections also affects the performance of classifications. As a conclusion, the WPCM technique has been applied to classify *the CNN* and *the Japan Times* sports news web pages. The experimental evaluation with different classification algorithms demonstrates that this method provides acceptable classification accuracy with the sports news datasets. Future works will include a web news

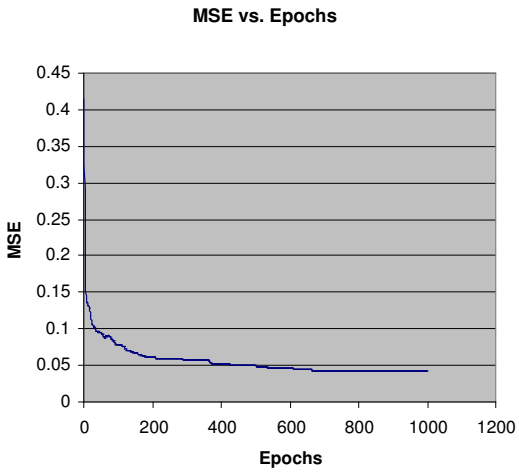

Fig. 5. The results of Run2 for MSE vs. Epochs for the news web pages classification using the neural networks (momentum rate, $\alpha = 0.001$).

classification using a support vector machine (SVM) with an automatic feature selection approach.

REFERENCES

[1] Stefan Wermeter, "Neural network agents for learning semantic text classification", *Information Retrieval*, Vol. 3. No. 2, pp. 87-103, 2000.

[2] Savio L.Y. Lam and Dik Lun Lee, "Feature reduction for neural network based text categorization", in *Proceedings of the 6th International Conference on Database Systems for Advanced Applications 19 - 22 April, Hsinchu, Taiwan*, 1999.

[3] Fabrizo Sebastini, "Machine learning in automated text categorization", *ACM Computing Surveys*, vol. 34, No.1, 2002.

[4] Dumais, S. T. , "Improving the retrieval of information from external sources", *Behavior Research Methods, Instruments and Computers*, Vol. 23, No. 2, pp. 229-236, 1991.

[5] The CNN web news page, *(http://www.cnn.com),* 2001.

[6] The Japan Times web news page, *(http://www.japantimes.co.jp),* 2001.

[7] Hisao Mase and Hiroshi Tsuji, "Experiments on automatic web page categorization for information retrieval system", *Journal of Information Processing, IPSJ Journal*, Feb. 2001, pp. 334-347, 2001.

[8] R. Calvo, M. Partridge, and M. Jabri, "A comparative study of principal components analysis techniques", in *Proc. Ninth Australian Conf. on Neural Networks*, Brisbane, QLD., pp. 276-281, 1998.

[9] A. Selamat, S. Omatu, H. Yanagimoto, "Information retrieval from the Internet using mobile agent search system", *International Conference of Information Technology and Multimedia 2001 (ICIMU),University Tenaga Nasional (UNITEN), Malaysia*, August 13-15, 2001.

[10] Sparck Jones, Karen, and Peter Willet, *Readings in information retrieval*, San Francisco: Morgan Kaufmann, USA, 1997.

[11] Salton & McGill, *Introduction to modern information retrieval*, New York, McGraw-Hill, USA, 1983.

[12] Yang, J., Pai, P., Honavar, V., and Miller, L. "Mobile intelligent agents for document classification and retrieval: a machine learning approach", In *Proceedings of the European Symposium on Cybernetics and Systems Research*, 1998.

[13] R. R. Korfhage, *Information storage and retrieval*, John Wiley and Sons, Inc, USA, 1997.

[14] Joachims, Thorsten, "Probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization", *Proceedings of International Conference on Machine Learning (ICML)*, 1997.

[15] McCallum, Andrew Kachites, "Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering", *(http://www.cs.cmu.edu/˜mccallum/bow)* , 1996.