Automated Web Pages Classification with Integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as Feature Reduction

Lee Zhi Sam[1] , Mohd Aizaini Maarof[2], Ali Selamat[3]

Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia,
81300 Skudai, Johor.
Email: samleecomp@gmail.com[1], maarofma@fsksm.utm.my[2], aselamat@fsksm.utm.my[3]

Abstract

With the explosive growth of internet, web pages classification has become an essential issue. This is because web pages classification will provide an efficient information search to internet users. Without professional classification, a website would become a jumble yard of content which is confusing and time wasting. By using web pages classification, it allows web visitors to navigate a web site quickly and efficiently. However, presently most of the web directories are still being classified manually or using semi-automated (huge teams of human editors)[1]. Automated web pages classification is highly in demand in order to replace expensive manpower and reduce the time consumed. In this paper we analyze the concept of a new model, which uses an integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as feature reduction for web pages classification. This model consists of several modules, which are web page retrieval process, stemming, stop-word filtering, feature reduction, feature selection, classification and evaluation.

Keywords

Web page classification, Web page retrieval, Principal Component Analysis, Independent Component Analysis, Class profile based feature, Neural Networks.

1. Introduction

There is an estimated of 1 billion pages accessible on the web with 1.5 million pages being added daily [2] .With the explosive growth of internet, web pages classification has become an essential issue. This is because web pages classification will provide an efficient information search for internet users. By using web pages classification, it allows web visitors to navigate a web site quickly and efficiently. Without professional classification, a website becomes a jumble yard of content that is confusing and time consuming. Presently, there are two methods that are commonly used by web users in finding useful information on the web. The two approaches are using search directory such as *yahoo*[1], *altavista*[2] or search engine like *google*[3]. Each approach has its own advantages. As an example internet users may find search directories as useful when browsing for general topics, while they may find search engines work well when searching for specific information.

In web directories, web pages are classified into hierarchical categories according to their content and stored in database. This allows the web users to browse desired information according to its category. However at present, most of the web directories are still being classified manually or semi-automatically (huge teams of human editors) [1]. Automated web pages classification is highly in demand in order to replace expensive manpower and to cut down the time consumed.

Currently no perfect web classification design has been found, as each classification design is highly dependable on the content of the web sites. Web site classification is an ongoing process prone to error. Each time a new document of content is published on the website, it needs to be classified. If the document is classified wrongly, then it undermines the entire classification design [3]. Therefore, various approaches have been applied to automated web pages classification in order to improve its performance. For example Qi et al. [1] use genetic K-means approach for automated web page classification, Yu et al. [4] use Positive Example Based Learning (PEBL) for web pages classification and Ali et al.[5, 6] propose a new web page classification method (WPCM)

Independent component analysis (ICA) is a well-known method of finding latent structure in data. ICA is a statistical method that expresses a set of multidimensional observations as a combination of unknown latent variables. ICA was originally developed for signal processing purposes, in particular for continuously distributed signals [7, 8]. Text documents is a very different application area in implementing ICA. However in statistical natural language processing, it has been observed that if text documents are presented in a numerical format, many numerical and computational methods can be utilized to analyze the textual data [8].

In this paper we analyze the concept of a new model using an integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as feature reduction for web pages classification. This model comprises several modules. The modules are web page retrieval process, stemming, stop-word filtering,

[1] http://www.yahoo.com/
[2] http://www.altavista.com/
[3] http://www.google.com/

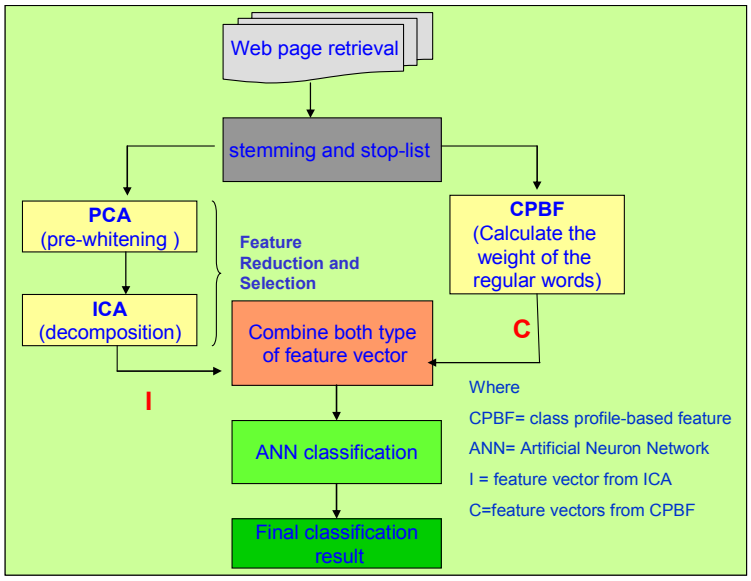

Figure 1.0 - The process of ICA and CPBF as input for ANN to do classification.

feature reduction, feature selection, classification and evaluation. This paper is constructed as follows: section 2 discusses our new web page classification model and section 3 discusses the factors of choosing feature extraction algorithm for this model. Finally, section 4 gives the conclusion of this paper.

2. Web Page Classification Model

Selamat et al. [5, 6] propose a sport news web page classification method (WPCM) which uses output vector of Principal Component Analysis (PCA) and class profile based feature (CPBF) as training input vector for Artificial Neural Network (ANN) to perform web pages classification. However the WPCM could achieve a higher classification rate if the integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) is implemented as feature reduction.

In this paper we would like to accentuate the concept of a new model using an integration of PCA and ICA as feature reduction for web pages classification. This model consists of several modules, which are web page retrieval process, stemming, stop-word filtering, feature reduction, feature selection, classification and evaluation. In order to classify the web pages after the preprocessing of web pages, first we will use the PCA algorithm [9, 10] to reduce the original data vectors to small number of relevant features. Features will be separated better and more independent during decomposition process using ICA algorithm [7, 11-14]. Then we will combine these features with the CPBF before keying in the input to neural networks for classification. This is as shown in Figure 1.0.

2.1 Preprocessing of Web Pages

Web page retrieval is a process that retrieves collections of web documents to database from online internet with the help of web crawler. Those retrieved web pages will be stored in local database for further process. Stop-list is a dictionary that contains the most common and frequent words such as 'I', 'You', 'and' and etc. Stopping is a process, which filters those common words that exist in web document by using stop-list. Stemming plays an important role in reducing the occurrence of term frequency that has similar meaning in the same document. It is a process of extracting each word from a web document by reducing it to a possible root word. For example, 'beauty' and 'beautiful' have the similar meanings. As a result, the stemming algorithm will stem it to its root word 'beauty'.

The data will be represented as the document-term frequency matrix (Doc_j x TF_{jk}) after the stemming and stopping process. Calculation based on term frequency inverse document frequency (*tfidf*) will be defined before feature selection and reduction. The calculation of the terms weight x_{jk} of each word w_k will be done by using a method as used by Salton[15] where

$$x_{jk} = TF_{jk} \text{ x } idf_k \quad (1)$$

Table 1 – Explanation of index for calculation based on term frequency inverse document frequency

Index	Explanation
j	Variable, j= 1,2,…,*n*
k	Variable, k=1,2,…,*m*
x_{jk}	Terms weight
Doc_j	Each web page document that exists in local database
TF_{jk}	Number of how many times the distinct word w_n occurs in document Doc_j
df_k	Total number of documents in the database that contains the word w_k
idf_k	Equal to log (n/df_k) where n is the total number of documents in database

2.2 Feature Reduction using PCA

In this model, Principal Component Analysis (PCA) will be performed in the preprocessing section as input for Independent Component Analysis (ICA) algorithm. Using PCA, the dimension reduction process will reduce the original data vectors into small numbers of relevant features. Meanwhile it also performs as pre-whitening process for ICA[8, 14].

Let M to be the matrix document-terms weight as below:

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

The definition of x,j,k,m and n have been explained in Table 1. The mean of m variables in data matrix M will be calculated by

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

Then the covariance matrix, C={c_{jk}} is calculated. The variance c^2_{kk} is given by

$$c^2{}_{kk} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{jk} - \bar{x}_k\right)^2 \tag{3}$$

The covariance is given by

$$c_{ik} = \frac{1}{n}\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) \tag{4}$$

where i =1,2,..,m. An eigenvalue λ and eigenvector **e** can be found by C**e** = λ**e** where C is covariance matrix. If C is an *m* x *m* matrix of full rank, *m* eigenvalues and all corresponding eigenvectors can be found by using

$$\left(C - \lambda_i I\right)e_i = 0 \tag{5}$$

We will sort the eigenvalues and eigenvectors so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$. A square matrix E can be constructed from the eigenvector columns where E=[$\mathbf{e}_1$ $\mathbf{e}_2$ **…** $\mathbf{e}_m$]. Let matrix B be denoted as

$$B = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & 0 & \lambda_m \end{pmatrix}$$

We will perform eigenvalue decomposition to get the principal component of matrix C by using

$$E^T C E = B \tag{6}$$

Next, we will select the first $d \leq m$ eigenvectors where *d* is the desired value such as 300,400,etc. The set of principal components is represented as $Z_1=\mathbf{e}_1{}^T c$, $Z_2=\mathbf{e}_2{}^T c$, …, $Z_p=\mathbf{e}_p{}^T c$. An *n* x *p* matrix R is represented as

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{np} \end{pmatrix}$$

where r_{ij} is a reduced feature vectors from the *m* x *m* original data size to *n* x *p* size.

2.3 Feature Reduction using ICA

In matrix form, the ICA framework is usually defined as linear noise-free generative model

$$X = AS \tag{7}$$

where the latent independent component is represented as random variable vector $S=(s_1,s_2,..,s_n)^T$. X is the observed mixture signals which $X=(x_1,x_2,\ldots,x_n)^T$ that are generated by multiplying A where matrix $A=(a_1,a_2,..,a_n)$ is a constant *n* x *n* mixing square matrix. It can be expressed by

$$X = a_1 s_1 + a_2 s_2 + \ldots + a_n s_n \tag{8}$$

which can also written as

$$X = \sum_{k=1}^{n} a_k s_k \tag{9}$$

where column a_k of the mixing matrix A, give the basis where the observations are represented. The goal of ICA is given X, find S and A where both A and S are statistically independent.

Independent component analysis relies on the concept of statistical independence. Statistical independence can be expressed as follows:

Random variables A and B are independent if the conditional probability of A with respect to A is just the probability of A. This means, knowing the value of B tells us nothing about A. Following is the equation:

$$P(A|B) = P(A) \tag{10}$$

Since P(A|B)= P(A,B), where P(A,B) is the joint density function of A and B. Then

$$P(A,B) = P(A)P(B) \tag{11}$$

The sources s can be estimated if the mixing matrix A is known. It can be done by multiplying the observed signals X with the inverse of the estimated mixing matrix $W=A^{-1}$:

$$S = y = A^{-1}y = WX \tag{12}$$

The first step in ICA algorithms is to whiten (sphere) the data. This means that it removes any correlations in the data, i.e. the feature data are forced to be uncorrelated and also statistically independent. Uncorrelatedness could be expressed as follow:

For two random vectors x and y to be uncorrelated, their cross-covariance matrix Cxy must be a zero matrix

$$C_{xy} = E\left\{(x - m_x)(y - m_y)^T\right\} = 0 \tag{13}$$

where $m_x = E\{x\}$ and $m_y = E\{y\}$ are the mean vectors of x and y, respectively. For zero-mean variables, zero covariance is equivalent to zero correlation. This means both are uncorrelated. However both matrices are real if the covariance matrix is symmetrical which is

$$\text{cov}(x,y) = \text{cov}(y,x) \tag{14}$$

If random vectors have zero-mean and the covariance matrix is an identity matrix, this is called white or sphered. As a result, the requirements for whiteness are

$$m_x = 0,\ A_x = C_x = I \tag{15}$$

where I is an identity matrix.

The whitened data has the form of

$$X = D^{-1/2}E^T V \tag{16}$$

where X is the whitened data vector, D is a diagonal matrix containing the eigenvalues of the correlation matrix and E contains the corresponding eigenvectors of the correlation matrix as its columns.

Dimensionality reduction is performed by PCA simply by choosing the number of retained dimensions, *d* and projecting the m-dimensional observed vector V to a lower dimensional space spanned by the d ($d < m$) dominant eigenvectors of the correlation matrix that was discussed in section 3.2. Now the matrix E in the equation (16) has only d columns instead of *m*, and similarity D is of size *d* x *d* instead of *m* x *m*, if whitening is desired.

Features will be separated better and more independent during decomposition process by implementing FastICA algorithm[11, 14]. The algorithm is an iterative fixed–point algorithm with the following update for W:

$$w \leftarrow E\left\{Xg\left(w^T X\right)\right\} - E\left\{g'\left(w^T X\right)\right\}w \tag{17}$$

where *w* is one of the rows of the unmixing matrix W. X is the data vector that has been centered and whitened in equation (16). The nonlinear function *g* is chosen so that it is the derivative of the non-quadratic contrast function g that measures non-Gaussianity, kurtosis or whatever our objective is. The choice of *g* is important to optimize the performance of the algorithm in some way. An initial unit norm vector w is chosen randomly. *w* is again normalized to have unit norm after each iteration step (17). The iteration is continued until the direction of *w* does not change significantly.

Alternatively, all of the vectors can be calculated at once by accomplishing

$$w \leftarrow \left(ww^T\right)^{-1/2} w \tag{18}$$

or iterative by [11, 14]

1. $w \leftarrow w/\|w\|$ (19)
2. $w \leftarrow \frac{3}{2}w - \frac{1}{2}ww^T w$ (20)
3. Repeat step 2 until convergence (21)

2.4 Feature Selection using CPBF

Class profile-based feature (CPBF) process is a process that identifies those most regular words in each class or category and calculate the weights of them by utilizing entropy method [5, 6, 16]. In feature selection using CPBF approach, we will manually identify those most regular words and weight them using the entropy weighting scheme before adding them to feature vectors that have been selected from the PCA and ICA. A fixed number of regular words from each class or category together with reduced independent component from the ICA will be used as a feature vector. This feature vectors will then be used as the input to the neural network for classification.

The entropy weighting scheme on each term is calculated as G_k x L_{jk} where G_k is the global weighting scheme of term k and L_{jk} is the local weighting of term k. The G_k and L_{jk} can expressed as

$$G_k = \frac{1 + \sum_{k=1}^{n} \frac{TF_{jk}}{F_k} \log \frac{TF_{jk}}{F_k}}{\log n} \tag{22}$$

and

$$L_{jk} = \begin{cases} 1 + \log TF_{jk} & \left(TF_{jk} > 0\right) \\ 0 & \left(TF_{jk} = 0\right) \end{cases} \tag{23}$$

where F_k is a frequency of term *k* in the entire document collection and TF_{jk} is the term frequency of each word in Doc_j which was mentioned previously in section 3.1. The *n* is the number of document in a collection.

2.5 Neural Network as Classifier

The output of feature vectors for both CPBF and ICA will be combined as the input for classifier. In this model, the artificial feed forward-back propagation neural network (ANN) is adopted as the classifiers. For classifying a test document d_j , its term weight w_{jk} is loaded into the input units. The activation of these units will propagate forward through the network, and finally the value of the output

unit(s) determines the categorization decision(s). The backpropagation neural network is used because, if a misclassification occurs, the error is "backpropagated" so as to change the parameters of the network and minimize or eliminate the error.

2.6 Measures of Classification Effectiveness

The classification effectiveness of this model will be measured using standard information retrieval measurement that are precision, recall, and F1 [2, 5, 15, 17]. These can be expressed as

$$P = \frac{a}{a+b} \tag{24}$$

$$R = \frac{a}{a+c} \tag{25}$$

$$F1 = \frac{2PR}{P+R} \tag{26}$$

where the values of *a, b* and *c* are explained in table 2. The F1 measure combines precision and recall with equal importance into a single parameter for optimization.

Table 2 – Explanation of parameter a, b and c

Category set $C=\{c_1,c_2,\ldots,c_n\}$		**Expert Judgment**	
		Yes	**No**
System Judgment	**Yes**	a	b
	No	c	d

3. Discussion

This model implements ICA as an extension of PCA because PCA does not perform as good as ICA in data separation (data independent)[8, 13, 18]. In addition, while ICA can also be seen as method of dimensionality reduction, we interpret dimensionality reduction as finding a parsimonious representation of the data. However dimensionality reduction is not the primary aim of ICA and in fact most of the ICA algorithms favor moderate (a few dozens compared to a few hundreds or more) dimensionalities of data [8].In other words, estimating ICA in the original high-dimensional space may lead to poor result. However the integration of PCA and ICA tends to handle badly conditioned separations better, hence being more stable [8]. On the other hand, using PCA to reduce the dimension prior to the ICA would aid the ICA decomposition in getting faster convergence[7, 8]. By using the integration of PCA and ICA as its feature reduction, we expect this model will yield a better classification result than the previous WPCM [5].

4. Conclusion

With the explosive growth of internet, web pages classification has become an essential issue in order to provide an efficient information search for internet users. Currently no perfect web classification design is found because each classification design is highly dependable on the content of web sites. In this paper we analyze the concept of a new model, which uses an integration of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) as feature reduction for web pages classification. We expect the integration of PCA and ICA tends to handle badly conditioned separations better, hence being more stable[8]. In addition, by using the integration of PCA and ICA as its feature reduction, we also expect that this model will yield better classification result than the previous WPCM[5]

Acknowledgement

This research is based on the paper by Selamat et al. [5]. Part of the enhancement highlighted in this paper is done, by adapting a new feature extraction approach model proposed by Selamat et al. [5]. The authors wish to thank reviewers for helpful suggestions.

References

[1] D. Qi and B. Sun, 2004. A genetic K-means approaches for automated Web page classification, presented at IEEE.

[2] J. Pierre, 2000. Practical Issues for Automated Categorization of Web Sites.

[3] G. McGovern, 2001. Web Classification is Essential, available at http://www.gerrymcgovern.com/nt/2001/nt_2001_11_26_classify.htm

[4] Y. Hwanjo, H. Jiawei, and K. C. C. Chang, 2004. PEBL: Web page classification without negative examples, *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, pp. 70-81.

[5] S. Ali and O. Sigeru, 2004. Web page feature selection and classification using neural networks, *Inf. Sci. Inf. Comput. Sci.*, vol. 158, pp. 69-88.

[6] A. Selamat, 2003. Studies on Mobile Agents for Query Retrieval and Web Page Categorization Using Neural Networks, in *Division of Computer and Systems Sciences, Gradute School of Engineering*, vol. Doctoral. Osaka: Osaka Prefecture University , pp. 94.

[7] J. J. Väyrynen, 2005. Learning Linguistic features from natural text data by independent component analysis, in *Department of Computer Science and Engineering*, vol. Master. Espoo: Helsinki University of Technology, pp. 63.

[8] E. Bingham, 2003. Advances in Independent Component Analysis with Applicatios to Data Mining, in *Helsinki Graduate School of Computer Science and Engineering*, vol. Doctoral thesis. Espoo: Helsinki University of Technology, pp. 60.

[9] R. A. Calvo, M. Partridge, and M. A. Jabri, 1998. A Comparative Study of Principal Component

Analysis Techniques, presented at In Proc. Ninth Australian Conf. on Neural Networks, Brisbane.

[10] R. A. Johnson and W. D. Wichern, 2002. *Applied Multivariate Statistical Analysis*, Fifth Edition ed. USA: Prentice Hall.

[11] A. Hyvarinen, 1999. Fast and robust fixed-point algorithms for independent component analysis, *Neural Networks, IEEE Transactions on*, vol. 10, pp. 626-634.

[12] A. Hyvarinen, 1999. Survey on independent component analysis, *Neural Computing Surveys*, pp. 2:94--128.

[13] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther, 2002. Independent component analysis for understanding multimedia content, presented at Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII, Piscataway, New Jersey.

[14] E. Oja, 2002. Convergence of the symmetrical FastICA algorithm, presented at in 9th Int. Conf. on Neural Information Processing.

[15] G. G. Chowdhury, 1999. *Introduction to modern information retrieval*. London: Library Association Publishing.

[16] S. Dumais, 1991. Improving the retrieval of information from external sources, pp. 229--236.

[17] J. O. P. Yiming Yang, 1997. A comparative study on feature selection in text categorization, presented at Proceedings of ICML-97, 14th International Conference on Machine Learning.

[18] T. Kolenda, 2002. Adaptive tools in virtual environments: Independent component analysis for multimedia, in *Informatics and Mathematical Modelling*, vol. Ph.D Kgs. Lyngby: Technical University of Denmark, pp. 117.