Automated Web Pages Classification with Independent Component Analysis

Lee Zhi Sam[1] , Mohd Aizaini bin Maarof[2], Ali Selamat[3]
Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia,
81300 Sukdai, Johor.
Email: samleecomp@gmail.com[1], maarofma@fsksm.utm.my[2], aselamat@fsksm.utm.my[3]

ABSTRACT
Automated web pages classification becomes an essential tool to reduce the manpower and time consuming for manually categorize web pages. In this paper, we propose a new method of web page classification which uses the output vector of Independent Component Analysis (ICA) and Class Profile Based Feature (CPBF) as input for ANN to do classification.

KEYWORDS
Text Categorization, Independent Component Analysis, Neural Network

1.0 Introduction

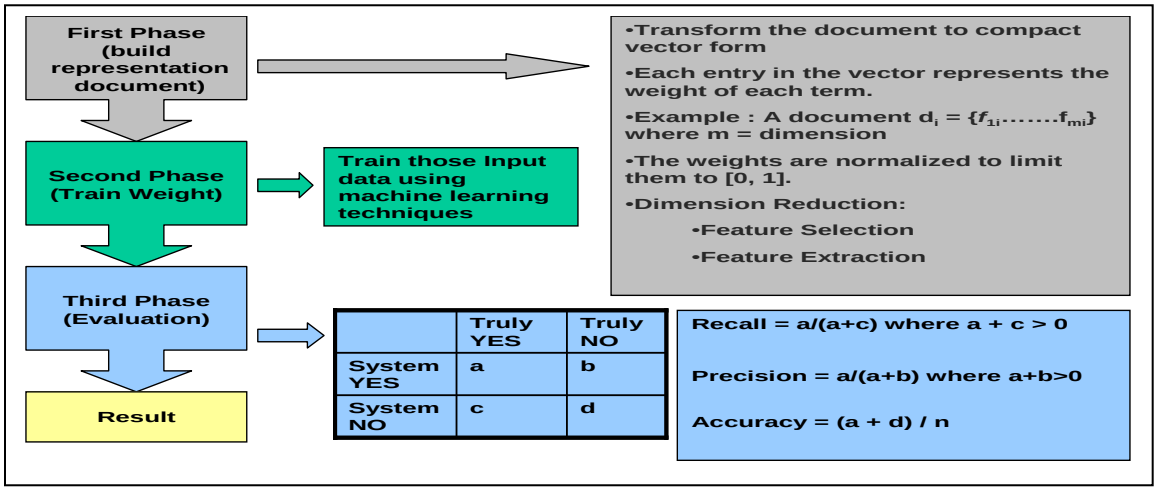
With the explosive growth of internet, web pages classification becomes an essential issue in order to provide an efficient information search for internet users. By using web pages classification, it allows web visitors navigate a web site quickly and efficiently. Without professional classification a website becomes a jumble yard of content that is confusing and time wasting. Presently, there are two approaches that commonly used by web users which are using search directory [9],[10] or search engine [11],[12] to find useful information in the web. Both approaches have its own advantages, for example search directories are useful when browsing general topics, while search engines work well when searching for specific information.

In web directories, web pages are classified in 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 still classifying web pages manually or semi-automated (huge teams of human editors) [1], automated web pages classification is highly demanded in order to replace expensive manpower and extreme time consuming.

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. [3] use Positive Example Based Learning (PEBL) for web pages classification, Ali et al. [2] propose a news web page classification method (WPCM) which using 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. The average for precision, recall and F1 measures for [2] produce quite a good result which are 90.31%, 93.81% and 91.65%. We believe that it still be able to perform better if modify the model of WPCM. In this paper we propose another approach that implements Independent Component Analysis (ICA) in [2]. This paper is constructed as follow: section 2 discuss the basic overview of web page classification and its related work, section 3 discuss our approach, section 4 discus the future work and section 5 give the conclusion of this paper.

2.0 Web page classification

Text categorization plays an important role in web pages classification. It is a task that automatically sorting a set of document into different categories according to its predefined set [6]. Figure 2 shows that text categorization is implemented as a part of web classification.



Firgure 1: Overview process of text categorization

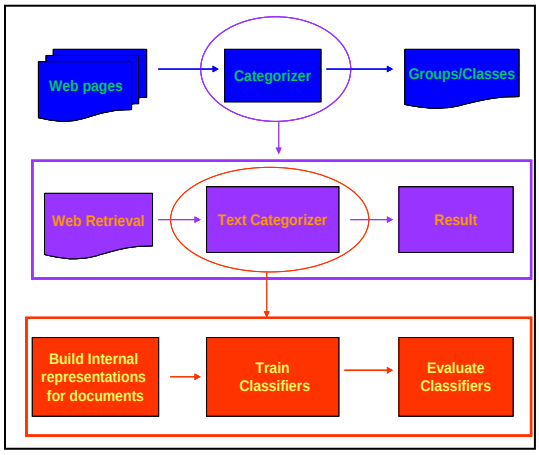


Figure 2: Simple structure of web classification.

Text categorization composes of three phases: building internal representations for documentations (phase-1), train classifiers with machine learning algorithm (phase-2), and evaluate classifiers using certain standard information retrieval measures such as precision, recall and F1 (phase-3, refer to figure 1) .

In phase-1, documents are transformed to compact vector form where each entry in the vector represents the weight of each term. For example: A document $d_i = \{f_{1i}.....f_{mi}\}$ where m= dimension. The vector will be the input to *tfdif* function (*td.dif*) so that the weight are normalize to limit them to [0,1]. Next, the approach of feature selection and extraction will be use for dimension reduction. WPCM approach [2] implements PCA algorithm as feature reduction and selection and combine the feature vector from CPBF as input for ANN training (phase-2).

Bingham et al. [4] and Kolenda et al. [5] implement ICA algorithm in text categorization. Their research shows that the dataset that using ICA may able to perform more independent result than PCA. The classification result from machine learning algorithm is highly dependent with the input of training data. Therefore, we assume that the classification result may perform better if the input data are more independent. As a result, we propose ICA algorithm as an extension of PCA in this paper.

In phase-2, those input data will be trained using machine learning techniques. Different approach of text classification method will use different design of machine learning techniques. For example [1] using Genetic K-means, [2] using ANN and [3] using Support Vector Machine (SVM) as machine learning algorithm.

In phase-3, the classification result will be evaluated using certain information retrieval. In this paper, we will only discuss recall, precision and accuracy which commonly use by most classifiers. Precision mean the percentage of positive predictions that are correct. Recall is the percentage of positive labeled instances that

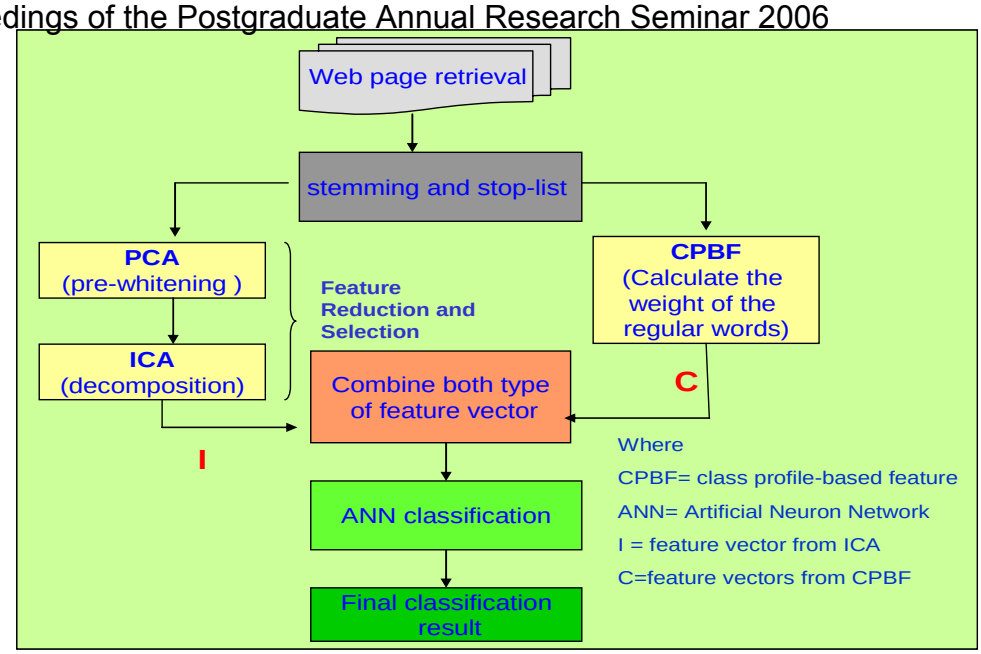


Figure 3: The process of ICA and CPBF as input for ANN to do classification.

were predicted as positive. However accuracy represents the percentage of predictions that are correct.

3.0 Our Approach

This paper proposed ICA and CPBF as input for ANN to perform classification. The process is shown in Figure 3. This approach consists of web page retrieval process, stemming, stop-word filtering, feature reduction and selection by PCA and ICA and CPBF.

Stop-list contains the most common and frequent words that exist in web document. Stopping is a process that will filter those common words such as 'I', 'You', 'and' and etc by using stop-list. However stemming is a process of extract each word from a web page 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 This will go through the process building representation document that had been discuss at section 2. In this model, PCA will perform in preprocessing section as input for ICA algorithm. This dimension reduction process will be able to reduce the original data vectors into small numbers of relevant features.

In the CPBF process, we will identify those most regular words in each class or category and calculate the weights of them by implement entropy method. The output of feature vectors for both CPBF and ICA will be combined as the input for ANN classifier. The final result of classifier will be evaluated using certain standard information retrieval measures that have been discussed in section 2.

4.0 Discussion

This approach is an extension of [2] in order to perform a better classification result. Experiments will be conducted to verify the performance of this model. Further modifications will be done in future (etc web page filtering) if the evaluation process brings positive results.

5.0 Conclusion

Automated web page classification is important to reduce the manpower and time in categorizing web pages manually. Moreover, it plays an important role to establish the semantic web [1]. This paper proposes a new method which using the output from ICA and CPBF algorithms as the input for ANN classifier. Further modifications will be done (etc web page filtering) in future if the evaluation process brings positive results.

Reference:

[1] Dehu Qi, Bo Sun , *A Genetic K-means Approaches for Automated Web Page Classification*, 0-7803-881 9-04 IEEE, 2004.

[2] Ali Selamat *, Sigeru Omatu , *Web Page Feature Selection and Classification Using Neural Networks*, 0020-0255, doi:10.1016/j.ins.2003.03.003, Elsevier Inc, 18 March 2003

[3] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang (2004), *PBL:Web Page Classification without Negative Examples*, IEEE Transaction On Knowledge And Data Engineering, Volume 16 No. 1,1041-4347 IEEE, January 2004.

[4] Ella Bingham, *Advance in Independent Component Analysis with Application to Data Mining*, Dissertation for degree of Doctoral of Science in Technology at Helsinki University of Technology (Espoo Finland), 2003.

[5] Thomas Kolenda, *Adaptive Tools in Virtual Enviroment, Independent Component Analysis for Multimedia*, Doctoral Thesis at Technical University of Denmark, ISSN 0909-3192 2002.

[6] Fabrizio Sebastiani , *Machine Learning In Automated Text Categorization*, 0360-0300/02/0300-0001 ACM,2002

[7] Paul Bennett, *Introduction to Text Categorization*, (Available at http://www.softlab.ntua.gr/facilities/public/AD/Text%20Categorization/Introduction%20to%20Text%20Categorization.pptT), 2002.

[8] Muluwork Geremew, *Machine Learning in Text Categorization*, (Available at http://www.umiacs.umd.edu/~joseph/tex-categorization.ppt)

[9] http://www.yahoo.com

[10] http://www.LookSmart.com

[11] http:// www.Google.com

[12] http://www.AltaVista.com