Web page feature selection and classification using neural networks

Ali Selamat and Sigeru Omatu
Computer and System Sciences, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
Received 20 June 2002; revised 30 December 2002; accepted 18 March 2003.
Available online 4 September 2003.

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. 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 combined with the feature vectors from the class-profile which contains the most regular words in each class. 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.

Author Keywords: Text categorization; WWW; Profile-based neural networks; Principal component analysis