Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 260–272
Expert Systems with Applications
www.elsevier.com/locate/eswa

Two novel feature selection approaches for web page classification

Chih-Ming Chen [a,*], Hahn-Ming Lee [b], Yu-Jung Chang [c]

[a] Graduate Institute of Library, Information and Archival Studies, National Chengchi University, No. 64, Sec. 2, ZhiNan Road, Wenshan District, Taipei 116, Taiwan, ROC
[b] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, ROC
[c] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, ROC

Abstract

To help the growing qualitative and quantitative demands for information from the WWW, efficient automatic Web page classifiers are urgently needed. However, a classifier applied to the WWW faces a huge-scale dimensionality problem since it must handle millions of Web pages, tens of thousands of features, and hundreds of categories. When it comes to practical implementation, reducing the dimensionality is a critically important challenge. In this paper, we propose a *fuzzy ranking analysis* paradigm together with a novel relevance measure, *discriminating power measure* (DPM), to effectively reduce the input dimensionality from tens of thousands to a few hundred with zero rejection rate and small decrease in accuracy. The two-level promotion method based on fuzzy ranking analysis is proposed to improve the behavior of each relevance measure and combine those measures to produce a better evaluation of features. Additionally, the DPM measure has low computation cost and emphasizes on both positive and negative discriminating features. Also, it emphasizes classification in parallel order, rather than classification in serial order. In our experimental results, the *fuzzy ranking analysis* is useful for validating the uncertain behavior of each relevance measure. Moreover, the DPM reduces input dimensionality from 10,427 to 200 with zero rejection rate and with less than 5% decline (from 84.5% to 80.4%) in the test accuracy. Furthermore, to consider the impacts on classification accuracy for the proposed DPM, the experimental results of China Time and Reuter-21578 datasets have demonstrated that the DPM provides major benefit to promote document classification accuracy rate. The results also show that the DPM indeed can reduce both redundancy and noise features to set up a better classifier.

Keywords: Feature selection; Fuzzy decision making; Web page classification; Discriminating power measure

1. Introduction

With the growing popularity of the World Wide Web, and the maturity and availability of related tools and techniques (like Web Servers, browsers, visual tools for Web page makers, dynamic HTML, Web-based databases and so on), more and more heterogeneous information is being "published" and added to the Web. The explosive growth in the number of Web pages has, in turn, contributed to the popularity of search tools. However, those search tools suffer from some problems. Search robots (like Openfind (Openfind), AltaVista (AltaVista)) often make users feel lost in *irrelevant search results*. Search tools based on manually maintained classified directories (like Yam (Yam), Yahoo! (Yahoo)) provide high-quality results but are hampered by low *production rates*. Since classification does improve search results but is time-consuming when done manually, *automatic* Web page classification should be considered to remedy the information-overloading problem.

An Automatic Web Page Classifier (AWPC) not only can relieve the slowness of manual classification, but could also guide the users of search tools through the various kinds of ambiguity by providing a list of topic paths. In order to achieve high-quality classification performance, both selection of effective features and selection of a

* Corresponding author. Tel.: +886 2 29393091x88024; fax: +886 2 29384704.
E-mail address: chencm@nccu.edu.tw (C.-M. Chen).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.09.008

classifier that can make good use of those features with limited training data, memory, and computing power are essential (Lippmann, 1989). In (Lippmann, 1987, 1989; Zurada, 1992; Nadler & Smith, 1993; Holmstom, 1997; Joshi, 1997), numerous pattern classification techniques (including statistical pattern recognition, neural networks, machine learning, neuro-biological, and neuro-fuzzy) are introduced, classified and compared. Importantly, many scholars were conscious of the subject of applying pattern classification techniques to Web page classification, and thus a growing number of classification models and machine-learning techniques have been applied to Web page classification in recent years, including multivariate regression models (Yang & Chute, 1994), nearest neighbor classification (Yang, 1994), Bayesian probabilistic approaches (Friedman, Geiger, & Goldszmidt, 1997), decision trees, neural networks (Musavi, Ahmed, Chan, Faris, & Hummels, 1992), symbolic rule learning (Cohen William & Singer, 1996), and inductive learning algorithms (Lewis, Schapire, Callan, & Papka, 1996). Moreover, lots of techniques were proposed to focus on the subject of Chinese Web page classification, such as the linear-based classifier (Chen, Liu, & Lee, 2001) (i.e. vector space model (VSM) classifier), neural network models (Chen, Lee, & Hwang, 2005), fuzzy theory (Yang & Hou, 1998), and so on. Besides, to improve Web page classification techniques, a novel proposed self-organizing HCMAC neural network classifier (Chen, 2003; Lee, Chen, & Lu, 2003) has been demonstrated its good performance for Web page classification. Actually, among the various kinds of classifiers, determining which ones are more appropriate for Web page classification is difficult and complex job.

In addition, when constructing an automatic Web page classifier, one still has to deal with the problem of huge-scale datasets. Namely, the AWPC must handle millions of Web pages (huge amount of instances), tens of thousands of features (extremely high input dimensionality), and hundreds or thousands of categories (high output dimensionality). Unfortunately, this situation is getting worse due to the explosive growth of Web pages. Consequently, effective feature selection mechanisms are critically important. Moreover, the VSM model (Chen et al., 2001; Salton, 1983, 1989) is simple and generally used for automatic document classification. To demonstrate the performance of the proposed feature selection approach for Web page classification, we focus on the combination of VSM classifier with the proposed feature selection method because an effective feature selection approach can generally promote classification accuracy rate for any classification models. To determine an appropriate criterion for feature selection, we emphasize that the given threshold value for extracting the informative feature terms has practically the uncertainty behavior. Therefore, we propose a *fuzzy ranking analysis* paradigm, which consists of *ranking analysis* steps to analyze and evaluate the uncertain behavior. Additionally, we also propose a *two-level promotion* technique to promote the performance of existing relevance measures, and present a novel relevance measure, named *discriminating power measure* (DPM), to obtain higher quality feature terms for document classification.

According to our experimental results, the *fuzzy ranking analysis* is useful for validating the uncertain behavior of each relevance measure. The *two-level promotion* techniques, which work under the restriction that relevance measures are limited and the perfect relevance measure is difficult to acquire by available sensing techniques, can show the trade-off between the rejection rate and the accuracy rate. Also, the experimental results for the DPM are very encouraging. The DPM greatly reduces input dimensionality, with zero rejection rate, while maintaining high classification accuracy. The DPM can reduce both redundancy and noise features.

2. Feature selection

In pattern classification, the so-called feature engineering process can be divided into three stages: feature generation stage, feature refinement stage, and feature utilization stage. In the feature generation stage, candidate features (i.e., the original feature set) are generated by pre-determined kinds of sensing techniques from the training set. For greater efficiency and even accuracy, the original feature set can be refined by feature selection and/or feature extraction. In feature selection, it is assumed that there are sufficient relevant features in the original feature set to discriminate clearly between categories, and that some irrelevant features can be eliminated to improve efficiency and even accuracy. For instance, elimination of redundancy will improve efficiency without losing accuracy, and elimination of noise will improve both efficiency and accuracy. In the feature extraction approach, it is supposed that the features in the original set are not all appropriate; nevertheless, sufficient information for classification is already captured by them. Feature extraction, which generates new features and measurements from the original features and measurements, is designed to handle this kind of situation. Examples include feature clustering, which may be the simplest way to implement feature extraction, and factor analysis by latent semantic indexing using singular value decomposition (Deerwester, 1990). The best way to test whether a representation is useful or not is simply to utilize it. In the feature utilization stage, features in the refined set are first used to represent each instance in the dataset. Then, an appropriate classification model is selected to make good use of these features.

2.1. Consideration of feature set quality

In pattern classification applications, when accuracy and/or efficiency are unacceptable, one tries to find the possible reasons and solve the problems. In Lewis (1992), Lewis enumerates six situations where feature set are of poor-quality so that obtaining a useful classifier is difficult or impossible: