Int. J. Production Economics 165 (2015) 215–222

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

A discriminative and semantic feature selection method for text categorization

Wei Zong [a], Feng Wu [a], Lap-Keung Chu [b,*], Domenic Sculli [b]

[a] School of Management, Xi'an Jiaotong University, PR China, 28 Xianning West Road, Xi'an, Shaanxi, PR China
[b] Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong SAR, PR China

ARTICLE INFO

Article history:
Received 6 January 2014
Accepted 28 December 2014
Available online 6 January 2015

Keywords:
Feature selection
Big data
Discriminative power
Semantic similarity
Text categorization
Support vector machine (SVM)

ABSTRACT

Text categorization is an important and critical task in the current era of high volume data storage and handling. Feature selection is obviously one of the most important steps in text categorization. Traditional feature selection methods tend to only consider the correlation between features and categories, and have in the main ignored the semantic similarity between features and documents. To further explore this issue, this paper proposes a novel feature selection method that first selects features in documents with discriminative power and then computes the semantic similarity between features and documents. The proposed feature selection method is tested using a support vector machine (SVM) classifier upon two published datasets, viz. Reuters-21578 and 20-Newsgroups. The experimental results show that the proposed feature selection method generally outperforms the traditional feature selection methods for text categorization for both published datasets.

1. Introduction

Aided by the rapid development of Internet and information technologies, the amount of digital documents to be stored and classified has increased dramatically. The syntactic complexity of written languages and the explosion of text volume to be processed have made it necessary for users to be able to quickly manage, organize and obtain the desired information in an age of 'big data'. Automatic text categorization has therefore attracted an increasing number of researchers, with the focus on how information can be easily retrieved and browsed (Balakrishnan et al., 1995). In a general context, text categorization is defined as the task of automatically classifying unlabeled documents into predefined categories so as to effectively manage huge volumes of text information. This definition is well established and has been widely applied in various fields such as information storage, information organization and information retrieval.

Up to now, much research effort has been directed at the problem of text categorization and many approaches have been proposed to address the problem with varying degrees of success. These approaches include, among others, neural networks (Ruiz and Srinivasan, 2002), *k*-nearest neighbors (*k*NN) (Yang, 1999; Tam et al., 2002), naive Bayes (McCallum and Nigam, 1998) and support vector machines (SVM) (Joachims, 1998). In most of these techniques, a text document is usually modeled as a vector space model (VSM). In this type of model, a document is represented as a feature vector whose components are the term weights, $d_k = (w_{1,k}, w_{2,k}, \ldots, w_{i,k}, \ldots, w_{n,k})$, where $w_{i,k}$ is the weight of term t_i in document d_k. Representing documents in this way was a major step forward, and it significantly improved the performance of text categorization methods (Leopold and Kindermann, 2002). However, because of the large number of features often found in a text document, a major difficulty in improving text categorization methods is now the extremely high dimensionality of the feature space. It is not uncommon for a typical text domain to contain thousands of features. While the dimensionality may be very high, in typical applications a considerable number of terms or features are not relevant to the text, and can be regarded as random noise features. Feature selection is therefore usually used to simplify the feature vectors and in turn improve the accuracy and efficiency of text categorization.

Most of the popular feature selection methods only consider the correlation between features and the categories they belong to, ignoring the semantic similarity between features and documents. Semantic similarity research in text categorization relies on the use of a common thesaurus such as WordNet and Wikipedia. However, information in a thesaurus is usually of a general purpose nature, and will usually not be particularly helpful in identifying domain-specific features and their relationships. It is also time consuming and

* Correspondence to: Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, 8/F Haking Wong Building, Pokfulam Road, Hong Kong. Tel.: +852 28592590.
E-mail address: lkchu@hkucc.hku.hk (L.-K. Chu).

http://dx.doi.org/10.1016/j.ijpe.2014.12.035
0925-5273/