Knowledge-Based Systems 24 (2011) 1024–1032

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm

Harun Uğuz *

Department of Computer Engineering, Selçuk University, Konya, Turkey

ARTICLE INFO

Article history:
Received 8 September 2010
Received in revised form 23 April 2011
Accepted 23 April 2011
Available online 29 April 2011

Keywords:
Text categorization
Feature selection
Genetic algorithm
Principal component analysis
Information gain

ABSTRACT

Text categorization is widely used when organizing documents in a digital form. Due to the increasing number of documents in digital form, automated text categorization has become more promising in the last ten years. A major problem of text categorization is its large number of features. Most of those are irrelevant noise that can mislead the classifier. Therefore, feature selection is often used in text categorization to reduce the dimensionality of the feature space and to improve performance. In this study, two-stage feature selection and feature extraction is used to improve the performance of text categorization. In the first stage, each term within the document is ranked depending on their importance for classification using the information gain (IG) method. In the second stage, genetic algorithm (GA) and principal component analysis (PCA) feature selection and feature extraction methods are applied separately to the terms which are ranked in decreasing order of importance, and a dimension reduction is carried out. Thereby, during text categorization, terms of less importance are ignored, and feature selection and extraction methods are applied to the terms of highest importance; thus, the computational time and complexity of categorization is reduced. To evaluate the effectiveness of dimension reduction methods on our purposed model, experiments are conducted using the *k*-nearest neighbour (KNN) and C4.5 decision tree algorithm on Reuters-21,578 and Classic3 datasets collection for text categorization. The experimental results show that the proposed model is able to achieve high categorization effectiveness as measured by precision, recall and *F*-measure.

1. Introduction

The number of text documents in digital format is progressively increasing and text categorization becomes the key technology to organize text data. Text categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns [2,29]. Although many information retrieval applications [3] such as filtering and searching for relevant information can benefit from text categorization research, a major problem of text categorization is the high dimensionality of the feature space due to a large number of terms. This problem may cause the computational complexity of machine learning methods used for text categorization to be increased and may bring about inefficiency and results of low accuracy due to redundant or irrelevant terms in the feature space [20,41,46]. For a solution to this problem, two techniques are used: feature extraction and feature selection.

Feature extraction is a process that extracts a set of new features from the original features into a distinct feature space [38]. Some feature extraction methods have been successfully used in text categorization, such as principal component analysis (PCA) [16,30], latent semantic indexing [33], clustering methods [31], etc. Among the many methods that are used for feature extraction, PCA has attracted a lot of attention. PCA [15] is a statistical technique for reduction of dimensionality that aims at minimizing loss in variance in the original data. It can be viewed as a domain independent technique for feature extraction, which is applicable to a wide variety of data [16].

Feature selection is a process that selects a subset from the original feature set according to some criteria of feature importance [22]. A number of feature selection methods are successfully used in a wide range of text categorizations. Yang and Pedersen [40] compared five feature selection methods for text categorization including information gain (IG), χ^2 statistic document frequency, term strength, and mutual information. They reported that IG is the most effective method among the compared feature selection methods. In addition to these feature selection methods, biologically inspired algorithms such as genetic algorithm (GA) [7,32,45] and ant colony optimization algorithm [1] have been successfully used in the literature for text categorization.

Genetic algorithm is an optimization method mimicking the evolution mechanism of natural selection. GA performs a search in complex and large landscapes and provides near-optimal solutions for optimization problems [32].

* Tel.: +90 332 223 19 26; fax: +90 332 241 06 35.
E-mail addresses: harun_uguz@selcuk.edu.tr, harun_uguz@hotmail.com

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.04.014

Text categorization is the task of classifying a document into predefined categories based on the contents of the document [4]. In recent years, more and more methods have been applied to the text categorization task based on statistical theories and machine learning, such as KNN [21,34,39], Naive Bayes [4,23], Rocchio [13], decision tree [6,9], support vector machine (SVM) [14,21,44], neural network [19,42], and so on. In this study, the C4.5 decision tree and KNN methods, which are used for text categorization, are used as classifiers.

In the current study, a two-stage feature selection and feature extraction are used to reduce the high dimensionality of a feature space composed of a large number of terms, remove redundant and irrelevant features from the feature space and thereby decrease the computational complexity of the machine learning algorithms used in the text categorization and increase performances thereof. In the first stage, each term in the text is ranked depending on their importance for the classification in decreasing order using the IG method. Therefore, terms of high importance are assigned to the first ranks and terms of less importance are assigned to the following ranks. In the second stage, the PCA method selected for feature selection and the GA method selected for feature extraction are applied separately to the terms of highest importance, in accordance with IG methods, and a dimension reduction is carried out. In this way, during text categorization, terms of less importance are ignored, feature selection and feature extraction methods are applied to the terms of the highest importance, and the computational time and complexity of the category are reduced. To evaluate the effectiveness of dimension reduction methods, experiments are conducted on Reuters-21,578 and Classic3 datasets collection for text categorization. The experimental results show that the proposed model is able to achieve high categorization effectiveness as measured by precision, recall and *F*-measure.

The rest of this paper is organized as follows. Section 2 presents a brief overview of the research methodologies and the experimental setting used. The effectiveness of the purposed method and experimental results for the categorization of a text document are demonstrated in Section 3, and finally, the paper is concluded in Section 4.

2. Research methodologies

The parts of proposed text categorization structure are shown in Fig. 1. These parts are explained in the following subsections:

2.1. Datasets

In this section, Reuters-21,578 and the Classic3 datasets used in the experiments are described and analysed.

2.1.1. Reuters-21,578 dataset

There are some public datasets that can be used as test collections for text categorization. The most widely used is the Reuters collection, which contains documents collected from Reuters news agency. The Reuters-21,578 collection [18] is a set of economic news published by Reuters in 1987. This collection includes 21,578 documents that are organized in 135 categories. In this experiment, the six categories including a minimum of 500 terms are selected. There are 8158 documents belonging to the chosen categories. The distributions of the number of documents in the six categories are shown in Table 1. According to Table 1, the distribution of documents into the categories is unbalanced. Maximum and minimum categories occupy 45.88% and 6.13% of the dataset, respectively.

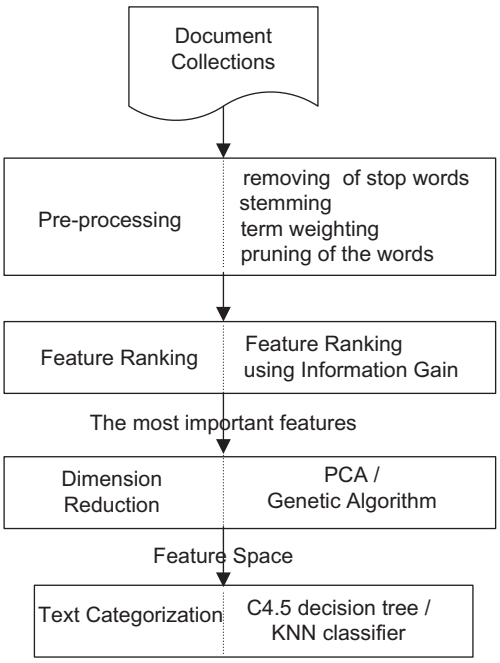

Fig. 1. Purposed text categorization structure.

Table 1
Distributions of the six categories for Reuters-21,578 Dataset.

Category name	Number of document
Earn	3743
Acquisition	2179
Money-fx	633
Crude	561
Grain	542
Trade	500

2.1.2. Classic3 dataset

We implemented the second experiment on the Classic3 dataset, a document collection from the SMART project at Cornell University (ftp://ftp.cs.cornell.edu/pub/smart). The Classic3 dataset is frequently used to evaluate performance of text categorization algorithms because it contains a known number of fairly well-separated groups. It contains three categories, i.e., 1398 CRANFIELD documents from aeronautical system papers, 1033 MEDLINE documents from medical papers, and 1460 CISI documents from information retrieval papers. The distribution of documents into the categories is balanced since all the categories are represented equally well in the dataset.

2.2. Pre-Processing

2.2.1. Removing of stop-words

Words such as conjunctions and pronouns that are not related to the concept of the text are called stop-words. This process involves removing certain common words such as 'a', 'an', 'the', etc., that occur commonly in all documents. It is important to removing these high-frequency words because they may misclassify the documents. In the study, stop words are removed in accordance with the existing stop word list (http://www.unine.ch/Info/clef/), which consists of 571 words.

2.2.2. Stemming

The stemming process leaves out the root forms of the words. Thereby, terms sharing the same root that seem like different