计算机工程与应用 2009 45 (36): 129-130 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

一种改进的文本分类特征选择方法

黄秀丽，王 蔚

南京师范大学 教育科学学院 机器学习与认知实验室，南京 210097

收稿日期 2008-12-30 修回日期 2009-3-2 网络版发布日期 2009-12-30 接受日期

摘要 文本分类中特征空间的高维问题是文本分类的主要障碍之一。特征选择（Feature Selection）是一种有效的特征降维方法。现有的特征选择函数主要有文档频率（DF），信息增益（IG），互信息（MI）等。基于特征的基本约束条件以及高性能特征选择方法的设计步骤，提出了一种改进的特征选择方法SIG。该特征选择方法在保证分类效果的同时，提高了对中低频特征的偏向。在语料集Reuters-21578上的实验证明，该方法能够获得较好的分类效果，同时有效提高了对具有强分类能力的中低频特征的利用。

关键词 文本分类 特征选择 信息增益

分类号 TP181

Improved feature selection method for text categorization

HUANG Xiu-li，WANG Wei

School of Education，Nanjing Normal University，Nanjing 210097，China

Abstract

High dimensionality is one of the main problems in text categorization.Feature selection methods can be regarded as an effective way.Main feature selection methods are document frequency，information gain，mutual information，and so on.This paper improves a new feature selection method SIG based on TTC and a universal method for developing feature selection functions.This method emphasizes the terms with middle and low frequencies and gets a good classification performance.Experiments on Reuters-21578 collection imply that this method is effective and can make better use of the terms with middle and low frequencies.

Key words text categorization feature selection information gain

DOI: 10.3778/j.issn.1002-8331.2009.36.038

通讯作者 黄秀丽 jing21000@yahoo.com.cn