计算机工程与应用 2010 46 (2): 129-131 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

一种不良文本识别特征选择方法

张永奎[1，2]，高 峰[1]

1.山西大学 计算机与信息技术学院，太原 030006
2.计算智能与中文信息处理教育部重点实验室，太原 030006

收稿日期 2008-7-29 修回日期 2008-10-20 网络版发布日期 2010-1-20 接受日期

摘要 针对不良文本的特殊性，提出一种两步特征选择方法。首先对训练文本用有限自动机识别其中的特殊词作为特征加入特征集，同时将原文还原为不含特殊词的文本。对还原后文本用“组合特征选择方法”选择特征加入特征集。实验结果表明利用两步特征选择方法能有效提高非法文本识别精度。

关键词 特殊词 有限自动机 特征选择 不良文本识别

分类号 TP391

Feature selection for illegitimate contents recognition

ZHANG Yong-kui[1，2]，GAO Feng[1]

1.Faculty of Computer & Information Technology，Shanxi University，Taiyuan 030006，China
2.Key Laboratory of Ministry of Education for Computation Intelligence and Chinese Information Processing，Taiyuan 030006，China

Abstract

To describe a two-steps fearture selection method.Firstly，recognise all the special words from the training texts by finite accepter and add it to the final feature set，recover the original text as well.Then select features from the processed texts and add them to the feature set by the way of‘conbination feature selection method’.The experiment result shows that it can improve the precision of the illegitimate contents recognition

Key words special words finite accepter feature selection illegitimate contents recognition

DOI: 10.3778/j.issn.1002-8331.2010.02.039

通讯作者 张永奎 zyk@sxu.edu.cn