计算机工程与应用 2009 45 (4): 156-158 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

基于边界可信度相似的快速文本分类方法

杨林波[1], 王士同[1, 2]

1.江南大学 信息工程学院，江苏 无锡 214122
2.江南大学 创新多媒体中心，江苏 无锡 214122

收稿日期 2008-1-9 修回日期 2008-4-2 网络版发布日期 2009-1-24 接受日期

摘要 类别的中心和边界是类别的重要特征. 利用训练样本的中心和边界作为分类准则，提出了一种基于边界可信度相似的快速文本分类算法。通过类别边界可信度调整文本与类别的相似性，克服了数据集类别间样本分布不均衡和类别中样本密度不均的缺点，提高了分类性能。实验结果表明该算法提高了文本分类的效果，显示出了较好的鲁棒性，并显著提高了文本分类效率。

关键词 文本分类 相似度 快速分类

分类号

Fast text categorization approach based on similarities between text boundaries

YANG Lin-bo[1],WANG Shi-tong[1,2]

1.School of Information，Jiangnan University，Wuxi，Jiangsu 214122，China
2.Creative Multimedia Center，Jiangnan University，Wuxi，Jiangsu 214122，China

Abstract

Center and boundaries are important characters of a class in text analysis.Using the center and boundaries as the criterion for text categorization，a fast text categorization approach based on the similarities between boundaries had been presented in this paper.By adjusting the similarity of a text to its class based on the similarity of the boundaries，the disadvantages of the imbalance of the classes and the distribution of the samples can be overcome such that the performance of text categorization may be enhanced.The experimental results demonstrate the advantage of the proposed approach in accuracy and robustness，especially in speed.

Key words text categorization similarity fast categorization

DOI: 10.3778/j.issn.1002-8331.2009.04.044

通讯作者 杨林波 seekingyang@163.com