计算机工程与应用 2009 45 (14): 136-137 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信号与信息处理

改进的χ^2统计文本特征选择方法

肖 婷, 唐 雁

西南大学 计算机与信息科学学院，重庆 400715

收稿日期 2008-11-20 修回日期 2009-2-13 网络版发布日期 2009-5-8 接受日期

摘要 特征选择是当今研究领域的一个热点，尤其是文本分类领域中的热点。针对χ^2统计方法的两个缺陷：降低了低频词的权重和提高了很少在指定类中出现但普遍存在于其他类的特征在该类中的权重，对χ^2统计方法进行改进，并通过做模拟和对比实验，对比改进前后的方法对文本分类的影响。在模拟和对比实验中，改进后方法的分类效果要好于传统的方法。

关键词 文本分类 特征选择 χ^2统计

分类号

Improved χ^2 statistics method for text feature selection

XIAO Ting,TANG Yan

School of Computer & Information Science，Southwest University，Chongqing 400715，China

Abstract

Feature selection is a hot topic in current search field，especially in the field of text categorization.In this paper，χ^2 statistical method has two defects.One is reducing the weight of the low-frequency words.The other is increasing the weight of the characteristics in the designated class.The characteristics little appear in designated class but other classes.Through simulation and comparison experiment，the result is better than before.

Key words text categorization feature selection χ^2 statistics

DOI: 10.3778/j.issn.1002-8331.2009.14.041

通讯作者 肖 婷 xiaott2006@163.com