基于卡方统计的情感文本分类

周爱武，马那那，刘慧婷

(安徽大学 计算机科学与技术学院，安徽 合肥 230601)

摘 要：通过对情感文本与 n-gram 特征的研究与分析，提出了一种基于卡方统计的特征词提取方法.方法中，n-gram 特征作为文本特征，在传统卡方统计的基础上选取共现或单独出现的特征，因为共现与单独出现的特征在不同类别中可能存在区别性.然后，根据多元特征与类别的相关性判别去除 n-gram 中冗余的特征，从而选取高类别相关而低冗余的 n-gram 特征.对上述方法利用 SVM 算法在不同语料中进行测试，通过实验对比分析，验证了该方法的有效性.

关键词：情感分析；卡方统计；n-gram；特征选择；相关性

Sentiment Text Classification Based on Chi-square Statistics

ZHOU Ai-wu，MA Na-na，LIU Hui-ting

(College of Computer Science and Technology， Anhui University， Hefei 230601， China)

Abstract：Because of the short sentiment text length， the lack of information，and the sparseness of features. When use the n-gram approach， the redundancy and relevance between words are ignored. This paper proposes n-gram features selection method based on Chi-square statistics. Firstly， each feature is evaluated by taking into account the simultaneous or individual occurrence of features within the feature set. Based on the idea that the occurrence of one feature but not the other may also convey valuable information for discrimination. Then the redundancy between words is reduced by chi-square statistic algorithm calculate the relevance between features and categories. So that we can extract n-gram features of high categories relevance and low redundancy. Finally， using Support Vector Machine classifier to identify the text orientation in different corpus， the experimental results show that this method improves the accuracy of text classification.

Key words： sentiment analysis;Chi-square statistics； n-gram;feature selection; relevance

作者简介：

周爱武 女，(1965-)，硕士，副教授，硕士生导师.研究方向为数据挖掘与 WEB 技术.

马那那(通讯作者) 女，(1988-)，硕士研究生.研究方向为数据挖掘. E-mail：1538238620@qq.com.

刘慧婷 女，(1978-)，博士，副教授，硕士生导师.研究方向为数据挖掘.