山东大学学报(工学版) 2012, 42(5) 87-90 DOI: ISSN: 0412-1961 CN: 21-1139

机器学习与数据挖掘

基于词贡献度的垃圾短信分类方法

张永军1,刘金岭2,于长辉3

淮阴工学院计算机工程学院, 江苏 淮安 223003

摘要： 针对垃圾短信分类问题, 提出了一种以词贡献度为基础的分类方法。该方法引入词贡献度的概念表达词在不同短信分类中的权重差别,通过构建词贡献度——分类矩阵和计算矩阵行均方差来实现降维,以词贡献度为基础计算短信隶属于短信分类的隶属度,并通过比较隶属度密度的方法解决分类冲突问题。实验结果表明,该方法在分类效果和实时性方面优于其他常用垃圾短信分类方法。

关键词： 垃圾短信 文本分类 词贡献度 方差 降维

A spam short message classification method based on word contribution

ZHANG Yong-jun1, LIU Jin-ling2, YU Chang-hui3

Faculty of Computer Engineering, Huaiyin Institute of Technology, Huai′an 223003, China

Abstract: A classification method based on word contribution was proposed to classify spam short messages. The concept of word contribution was introduced for representing weight difference of a word in different categories, the word contribution classification matrix was constructed, then the mean square deviation of each row in the matrix was computed to reduce dimensionalities. To determine the classification a short message belongs to, short message category membership degrees were calculated based on word contribution. Furthermore if category candidates were more than one, the classification conflict problem could be resolved by comparing the densities of short message category membership degree. The experimental results showed that the proposed method was superior to other classification methods in the classification result and real time.

Keywords: spam short message text classification word contribution variance dimensionality reduction

收稿日期 2012-05-20 修回日期 网络版发布日期

DOI:

基金项目:

江苏省教育厅高校哲学社会研究资助项目(2012SJD87001)

通讯作者:

作者简介: 张永军(1978-),男,江苏扬州人,讲师,硕士,主要研究方向为文本数据挖掘. E-mail:13511543380@139.com

作者Email:

PDF Preview

参考文献：

本刊中的类似文章

1．李可,刘常春,李同磊 .一种改进的最大互信息医学图像配准算法[J]. 山东大学学报(工学版), 2006,36(2): 107-110

2．任敬喜,耿金花,高齐圣 .多因素多指标产品的质量优化[J]. 山东大学学报(工学版), 2007,37(3): 114-117

3．张道强.知识保持的嵌入方法[J]. 山东大学学报(工学版), 2010,40(2): 1-10

4．廖伙木,董增川, 束龙仓,贠汝安 .地下水位预报中的组合时间序列分析法[J]. 山东大学学报(工学版), 2008,38(2): 96-100

5．曾雪强1,李国正2.基于偏最小二乘降维的分类模型比较[J]. 山东大学学报(工学版), 2010,40(5): 41-47

6．张丽梅1,2 ,乔立山1,2 ,陈松灿1 .基于张量模式的特征提取及分类器设计综述[J]. 山东大学学报(工学版), 2009,39(1): 6-14

扩展功能
本文信息
Supporting info
PDF(1003KB)
参考文献[PDF]
参考文献
服务与反馈
把本文推荐给朋友
加入我的书架
加入引用管理器
引用本文
Email Alert
文章反馈
浏览反馈信息
本文关键词相关文章
垃圾短信
文本分类
词贡献度
方差
降维
本文作者相关文章
PubMed

7. 王熙照,白丽杰*,花强,刘玉超.null[J]. 山东大学学报(工学版), 2011,41(4): 1-6

8. 王洪元,封磊,冯燕,程起才.流形学习算法在中文文本分类中的应用[J]. 山东大学学报(工学版), 2012,42(4): 8-12

9. 贺广南，杨育彬*.基于流形学习的图像检索算法研究[J]. 山东大学学报(工学版), 2010,40(5): 129-136

10. 王法波,许信顺.文本分类中一种新的特征选择方法[J]. 山东大学学报(工学版), 2010,40(4): 8-11

Copyright by 山东大学学报(工学版)