计算机工程与应用 2009 45 (23): 140-143 ISSN: 1002-8331 CN: 11-2127/TP

数据库、信息处理

由Logistic回归识别Web社区的垃圾评论

何海江，凌 云

长沙学院 计算机中心，长沙 410003

收稿日期 2008-5-29 修回日期 2008-8-4 网络版发布日期 接受日期

摘要 针对Web社区垃圾信息泛滥的问题，采用基于Logistic回归（LR）的分类器区分合法评论和垃圾评论，并和支持向量机（SVM）的性能对比。提出了相关度向量空间模型cVSM作为评论的文档表示模型，讨论了信息增益IG、互信息MI、χ^2统计CHI、文档频率DF等不同特征抽取方法对模型的影响。实验结果表明，LR的训练时间不到SVM的1/10；DF和IG比MI和CHI表现更好；与传统的向量空间模型相比，使用cVSM显著提高垃圾评论识别能力。

关键词 Logistic回归 向量空间模型 博客 垃圾评论 相关度

分类号 TP391

dentifying comment spams of Web forums by classifier based Logistic regression

HE Hai-jiang，LING Yun

Computer Teaching Center，Changsha University，Changsha 410003，China

Abstract

A classifier based on Logistic Regression（LR） is employed to identify comment spams which have flooded in Web forums.Comparative study on performances of LR and Support Vector Machine（SVM） is presented.It is introduced that a relevancy coefficient vector space model named cVSM which is used to express comment archives.Some feature extractive methods are discussed，including Information Gain（IG），Mutual Information（MI），χ^2 statistic（CHI）and Document Frequency（DF）.The experiments show that：The learn time of LR is less than 1/10 of SVM's.DF and IG have better performances than MI and CHI.To be compared with vector space model，cVSM has improved comment spam cognitive capability of classifier.

Key words Logistic Regression（LR） vector space model blog comment spam relevancy coefficient

DOI: 10.3778/j.issn.1002-8331.2009.23.039

通讯作者 何海江 haijianghe@sohu.com