计算机工程与应用 2007 43 (35): 126-129 ISSN: 1002-8331 CN: 11-2127/TP

网络、通信与安全

最大熵模型在邮件分类中的应用

李军辉, 李培峰, 朱巧明, 钱培德

苏州大学 计算机科学与技术学院，江苏 苏州 215006

收稿日期 修回日期 网络版发布日期 2007-11-29 接受日期

摘要 邮件分类是指在给定的分类体系下，根据邮件的内容和属性，确定其类别标签的过程。将最大熵模型应用于邮件分类中，给出了邮件的预处理过程，介绍了邮件信头特征，分析比较了特征数量和迭代次数、邮件特征字段对分类结果的影响，以及对层次分类和平面分类的效果进行了比较。实验表明，特征数量和迭代次数分别取2 000和250时为宜；充分利用邮件各字段信息，取得的总体分类效果最好，但对合法邮件，利用邮件头及邮件标题却取得了最好结果，并在层次分类中验证了这点，层次分类效果要优于平面分类。最后进行了总结和展望。

关键词 最大熵模型 邮件分类 特征 层次分类

分类号

Email categorization with maximum entropy model

LI Jun-hui,LI Pei-feng,ZHU Qiao-ming,QIAN Pei-de

School of Computer Science and Technology，Suzhou University，Suzhou，Jiangsu 215006，China

Abstract

Email categorization assigns new emails to pre-defined categories based on their contents and properties.In this paper，the maximum entropy model is applied to email categorization.The pre-process of email is discussed firstly，and features extracted from email header are presented.No only the effects of categorization performance coursed by Email feature fields，the numbers of features and iteration are presented and discussed，but also the performance of hierarchy categorization and direct categorization is compared.The results of experiments show that the appropriate numbers of features and iteration are 2 000，250 respectively and utilizing all email fields gets the best whole performance than others，and that the performance of legitimate by using email header and subject is best though the whole performance is worst，which is also verified in the hierarchy categorization experiments.The results also illuminate that the effect of hierarchy categorization is better than that of direct categorization.The summarization and future work are presented in the end.

Key words maximum entropy model e-mail classification feature hierarchy categorization

DOI:

通讯作者 李军辉 210313060@suda.edu.cn

扩展功能

本文信息

Supporting info

PDF(770KB)

[HTML全文](0KB)

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

复制索引

Email Alert

文章反馈

浏览反馈信息

相关信息

本刊中 包含“最大熵模型”的相关文章

本文作者相关文章

- 李军辉
- 李培峰
- 朱巧明
- 钱培德