云环境下 NB 算法的垃圾邮件过滤研究

刘月峰，张亚斌，苑江浩

（内蒙古科技大学 信息工程学院，内蒙古 包头 014010）

摘要： 朴素贝叶斯算法在解决垃圾邮件分类领域内具有较高的准确性，能够很好的将邮件区分开来，但是在分类前期的训练阶段却会大量耗用系统和网络资源，严重影响分类效率.为此引入 spark 平台.以并行的思想去解决邮件分类问题，利用 spark 计算平台 RDD 的血缘关系合理的安排 NB 邮件分类的各个过程.实验结果表明,与其他传统的分类方法对比而言，朴素贝叶斯在精确率，召回率等方面具有很好的效果，并且与传统单机下的邮件分类，本次实验因引入分布式的思想，利用 spark 集群的优势大大加快了分类的速率.

关键词： 垃圾邮件；朴素贝叶斯；spark 计算平台；分布式

中图分类号： TP393

文献标识码： A

文章编号： 1000-7180(2018)08-0060-04

Research of Spam Filtering Based on NB Algorithm in Cloud Environment

LIU Yue-feng, ZHANG Ya-bin, YUAN Jiang-hao

(School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou 014010,China)

Abstract：Na 汇 ve Bayes algorithm has high accuracy in solving the spam classification field and can distinguish the mail very well. However, in the pre-classification training phase, it consumes a lot of system and network resources and seriously affects the classification efficiency . Spark platform for this introduction. With parallel thinking to solve the problem of mail classification, the use of spark computing platform RDD kinship rational arrangement of NB mail classification of the various processes. The experimental results show that, compared with other traditional classification methods, Na 汇 ve Bayes has a good effect on the Precision and Recall rate, etc., and with the traditional mail classification under single machine, this experiment because of the introduction of distributed thinking , The use of spark clusters greatly accelerate the classification speed.

Key words： spam email；naive bayes；spark computing platform；distributed

作者简介：

刘月峰男，（1977-），博士研究生，副教授.研究方向为机器学习、文本分类.

张亚斌（通讯作者）男，（1992-），硕士研究生.研究方向为大数据、文本分类.E-mail:1551255025@163.com.

苑江浩男，（1992-），硕士研究生.研究方向为机器学习、数据挖掘.