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: Naïve 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, Naïve 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

1 引言

随着网络技术的越发成熟，垃圾邮件也猖獗异常，甚至携带病毒，为人们日常生活带来了不便[1-2]. 其中朴素贝叶斯因其分类能力和精确性表现优秀而备受欢迎. 但是朴素贝叶斯分类技术在垃圾邮件分类的前期的训练阶段会耗费大量的时间和系统资源进行训练和学习因此引入了大数据分布式处理平台.

在众多应用大数据平台的软件中，Hadoop是一款重量级的基础平台，非常适用于大规模数据的处理[3]. 许多研究人员也曾通过Hadoop平台用于垃圾邮件的分类. 如陶永才等人提出的基于MapReduce的贝叶斯垃圾邮件过滤机制[4]，而Spark[5]作为大数据计算平台的后起之秀，提供了内存计算. 在2014年成功打破了Hadoop保持的基准排序(Sort Benchmark)纪录，仅仅使用了Hadoop的计算资源的十分之一，却获得了比之快3倍的速度. 也因此表明了spark可以作为一个更加快速高效的大数据计算平台[6].

收稿日期：2017—11—20；**修回日期**：2018—01—05

针对传统的朴素贝叶斯分类方法无法胜任大规模邮件的过滤任务，本文将朴素贝叶斯过滤算法与大数据平台 spark 结合进行研究. 提出在分布式平台 spark 上运行朴素贝叶斯算法对垃圾邮件进行分类. 加快分类的速率.

2　基于 spark 的朴素贝叶斯邮件分类模型

朴素贝叶斯分类算法[7-11]的“独立性假设”虽然与具体算法没有直接的联系，增进邮件文本中的各个特征词的独立性对于分类的效果没有显著的功效，因此，为了朴素贝叶斯的算法复杂度不在增加，本文选择仍然基于算法的“独立性假设”.

2.1　基于 spark 的朴素贝叶斯分类并行化

Spark 分布式平台是基于内存的特点大大减少了在运算过程中的因读取和写出造成的资源和时间的消耗，通过 RDD 机制合理的构建有向无环图 DAG 以并行的方式加速实验的进程，很大程度的提高了计算的效率.

图 1 所示为邮件分类流程. 训练邮件集先经过分词与去停用词处理形成特征词集合，在经向量化形成特征向量空间. 联系原本的输入集共同进行样本的学习与训练，最后得到各个类别的先验概率和各特证词的条件概率. 新邮件的判断即为邮件的后验概率值大小的比较，选出最大的一方作为最终的类别.

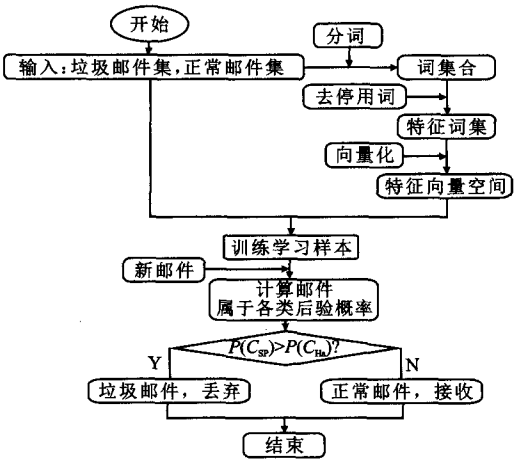

图 1　邮件判别流程图

垃圾邮件的分类在大数据计算平台 spark 上以朴素贝叶斯分类算法进行分类也主要是以＜key，value＞的形式并行处理. 通过 Map 和 Reduce 两个阶段进而完成分类，而不同的 Map 和 Reduce 处理各自之间独立运行从而实现并行化. spark 从最开始读入数据构造原始 RDD，通过不断的“转换”构造新的 RDD. 这之间描述出了各个 RDD 间浓厚的“血缘”关系，通过构造有向无环图 DAG 合理的处理各个操作[11].

垃圾邮件的分类主要分为预处理阶段、训练阶段和分类阶段，预处理阶段主要包括邮件集的分词、去停用词等一系列操作. 训练阶段为得出邮件的类别所需的各个特征词的条件概率和类别的先验概率等，形成下一个阶段分类所需的“知识库”，分类阶段主要通过前两个阶段形成的知识库对测试集内的邮件进行分类. 图 2，3 为垃圾邮件在 spark 上的分类主要过程.

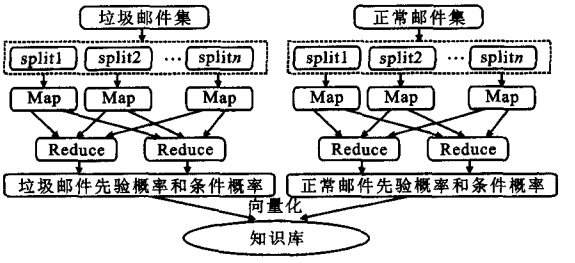

图 2　邮件训练阶段流程图

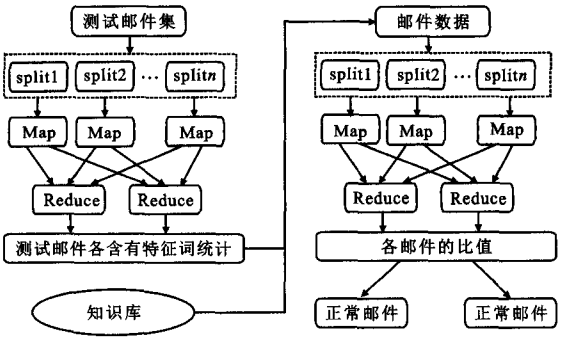

图 3　邮件分类阶段流程图

在图 2、3 中已经包括邮件的分词，去停用词等预处理阶段的内容，朴素贝叶斯在 spark 的框架上主要使用 Map 和 Reduce 的形式并行处理邮件.

3　实验仿真

3.1　实验数据和环境

本文主要采用两个邮件集，其中之一是公共平台上给出的邮件集，包含合法邮件 16 556 封，垃圾邮件 27 360 封. 第二个邮件集为中国教育和科研计算机网紧急响应组（DataSetsofChineseEmails，CCERT2005-jun），含有正常邮件 9 272 封，垃圾邮件数目为 25 088 封.

本次实验的主要采取 spark 集群来运行垃圾邮件的朴素贝叶斯分类模型，实验的具体数据参数如

下表 1 所示.

表 1 实验数据

实验项目	值
节点个数	5
JDK 版本	jdk-8u11-linux-x64
Spark 版本	spark-2.1.0
Spark 模式	Stand Alone
Hadoop 版本	hadoop-2.6.4
节点内存	8G
节点磁盘	60G
操作系统	Ubuntu 14.04

3.2 实验结果与分析

实验 1 spark 集群环境和单机环境邮件分类速率对比采用数据集部分邮件,对比不同环境下的速率.本文因采取了两个邮件集,考虑到不同邮件集的整合会影响分类的效果.分别在正常邮件和垃圾邮件集中各自去掉 3 封邮件.实验将两个邮件集打破顺序,将邮件整合在一起.并且再将邮件集平均分为 5 组(a,b,c,d,e).每组包含 5 165 封垃圾邮件和 10 489封合法邮件,分别测试两种环境下分类所用的时间.其中单机环境下的作为实验 A,在 spark 集群环境下进行的实验作为实验 B.随机抽取两组邮件集合作为实验所用数据集.结果如表 2 所示.

表 2 邮件分类时间比较

实验组	机器数	实验时间/s	
		邮件组 a	邮件组 b
实验 A	—	2 326	2 465
实验 B	1	1 835	1 887
	2	935	1 063
	3	506	596
	4	423	453
	5	392	413

对比实验结果,两组实验在随机取得数目相同邮件集的情况下运行时间相对而言并没有太大的波动,这说明了实验相对稳定,增加了可信度.在对比单机与集群的情况,在邮件集相同的情况下 spark 计算平台在开启一台机器运行比单机运行的时间要短,这充分显示了 spark 计算平台基于内存运行的优越性,如图 4 所示.

随着集群启动机器数目的增多,邮件分类运行的时间先是大幅度减小,当机器台式达到 3 台时,邮件分类的运行时间开始趋于平缓,当机器个数增多其运行的时间起伏不大,这表明了对于固定的数据集 spark 集群会随着机器台数而随着启动机器数的

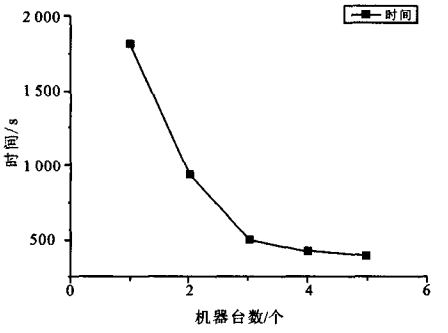

图 4 spark 机器台数与运行时间的关系

增多运行时间也越加缩短.但当数目达到一定值时对运行速度增幅就变得有限.并目数目越多,运行的时间也越加趋于平缓.

实验 2 两种环境下分辨邮件的能力对比

此次实验主要验证 NB 算法在集群环境下分辨垃圾邮件的能力是否会有所不同,从邮件语料集中取出上述实验的 a 和 b 两组邮件集作为实验的测试集,分别比较两组邮件集合在单机环境和 spark 集群环境对邮件的各个指标的影响.其中在单机环境下的 NB 分类作为实验 A,在集群环境下 NB 算法分类实验作为实验 B.分别测试 R(召回率),P(正确率)和 F 值.其中实验结果如表所示.

表 3 邮件的分辨能力对比 (%)

邮件	实验	R	P	F
a	A	92.56	92.87	92.71
	B	92.49	92.89	92.69
b	A	92.45	92.86	92.65
	B	92.48	92.81	92.64

通过表 3 可以看出在 spark 集群上运行朴素贝叶斯算法和传统朴素朴素贝叶斯模型在召回率、正确率和 F 值方面几乎没有差别,这说明了朴素贝叶斯算法对垃圾邮件分类在 spark 集群上并没有对召回率、正确率上相对稳定,并未对垃圾邮件的分类造成不良影响.

4 结束语

本文针对大数据带来的邮件数据量大,种类更加复杂传统邮件分类系统已无法应对的问题,提出了在 spark 计算平台上实现垃圾邮件的分类过滤机制.利用其基于内存的特性和 RDD 的"血缘关系"通过行动和转换组成 NB 分类运行结构,实现并行过滤.spark 计算平台分布式框架的特性提高邮件的分类速率.

本次实验虽然提高了邮件分类的效率，但存在不足之处. 应用的 spark 集群框架对邮件分类加速，在相同的数据集上分类时，数据的正确率等会有不同程度的起伏甚至下降，分析问题原因为不同的邮件分类时结果不同因此邮件分类的精度还需提高. 虽然此次实验虽然在精度上并没有下降，但在 NB 算法精度方面有待提高. 后续研究可以考虑在 spark 上运用特征提取和增加权值等方面增加分类的精度[12]；另一方面，朴素贝叶斯算法主要是研究整体邮件的合法与垃圾邮件的分类，是一种大众化的分类. 但是在邮件的具体分类上个体的看法不一定与众人相同，意味着可以考虑根据每个人兴趣情况取舍邮件，研究可以考虑根据个人情况进一步提高邮件的过滤精度.

参考文献：

[1] 方言. 邮件安全不容忽视[J]. 中国信息安全，2017，7(3)：100-100.

[2] 张铭锋，李云春，李巍. 垃圾邮件过滤的贝叶斯方法综述[J]. 计算机应用研究，2005，22(8)：14-19.

[3] Ding Q，Boykin R. A framework for distributed nearest neighbor classification using Hadoop[J]. Journal of Computational Methods in Sciences & Engineering，2016(17)：1-9.

[4] 陶永才，薛正元，石磊. 基于 MapReduce 的贝叶斯垃圾邮件过滤机制[J]. 计算机应用，2011，31(9)：2412-2416.

[5] 曾青华，袁家斌，张云洲. 基于 Hadoop 的贝叶斯过滤 MapReduce 模型[J]. 计算机工程，2013，39(11)：57-60.

[6] Davoodi M，Segal S，Peretz N K，et al. Semi-automated program for analysis of local Ca 2+，spark release with application for classification of heart cell type[J]. Cell Calcium，2017(64)：83-90.

[7] 刘月峰，苑江浩，张晓琳. 改进 NB 算法在垃圾邮件过滤技术中的研究[J]. 微电子学与计算机，2017，34(4)：115-120.

[8] 于苹苹，倪建成，姚彬修，等. 基于 Spark 框架的高效 KNN 中文文本分类算法[J]. 计算机应用，2016，36(12)：3292-3297.

[9] Ramírez-Gallego S，Krawczyk B，García S，et al. Nearest neighbor classification for high-speed big data streams using spark[J]. IEEE Transactions on Systems Man & Cybernetics Systems，2017(99)：1-13.

[10] Geng Y，Zhang J. Optimized parallelization of binary classification algorithms based on spark[J]. IEEE Robotics & Automation Magazine，2006，13(4)：11-13.

[11] 王雯，赵衍衍，李翠平，等. Spark 平台下的短文本特征扩展与分类研究[J]. 计算机科学与探索，2017，11(5)：732-741.

[12] 李涛，刘斌. Spark 平台下的高效 Web 文本分类系统的研究[J]. 计算机应用与软件，2016，33(11)：33-36.

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.

(上接第 59 页)

作者简介：

姚玉坤　女，(1964-)，硕士，教授. 研究方向为宽带无线自组织网络、网络编码.

李小勇(通讯作者)　男，(1992-)，硕士研究生. 研究方向为无线宽带自组织网络路由、网络编码. E-mail：lixy_hbjz_cqupt@163.com.

徐栋梁　男，(1990-)，硕士. 研究方向为无线传感器网络拓扑维护算法.

刘江兵　男，(1989-)，硕士研究生. 研究方向无线自组织网络路由算法.