·先进计算与数据处理· 文章编号：1000—3428(2013)11—0057—04 文献标识码：A 中图分类号：TP338.6

基于 Hadoop 的贝叶斯过滤 MapReduce 模型

曾青华，袁家斌，张云洲

(南京航空航天大学计算机科学与技术学院，南京 210016)

摘 要：传统分布式大型邮件系统对海量邮件的过滤存在编程难、效率低、前期训练耗用资源大等缺点，为此，对传统贝叶斯过滤算法进行并行化改进，利用云计算 MapReduce 模型在海量数据处理方面的优势，设计一种基于 Hadoop 开源云架构的贝叶斯邮件过滤 MapReduce 模型，优化邮件的训练和过滤过程。实验结果表明，与传统分布式计算模型相比，该模型在召回率、查准率和精确率方面性能较好，同时可降低邮件过滤成本，提高系统执行效率。

关键词：云计算；MapReduce 模型；Hadoop 架构；贝叶斯算法；垃圾邮件；反垃圾邮件过滤

Hadoop-based MapReduce Model of Bayesian Filtering

ZENG Qing-hua, YUAN Jia-bin, ZHANG Yun-zhou

(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

【Abstract】There are some disadvantages of mass mail filtering for large mail systems on the traditional distributed system including programming difficulties, low efficiency, mass system and network resources consumed. Taking advantage of the high performance of the cloud computing in processing data processing effectively, a MapReduce model of Bayesian mail filtering based on Hadoop is proposed. It improves the traditional Bayesian filtering algorithms and optimizes the mail training and filtering processes. Experimental results show that, compared with traditional distributed computing model, the Hadoop-based MapReduce model of Bayesian anti-spam mail filtering performs better in recall, precision and accuracy, reduces the cost of mail learning and classifying and improves the system efficiency.

【Key words】cloud computing; MapReduce model; Hadoop framework; Bayesian algorithm; spam mail; anti-spam mail filtering

DOI: 10.3969/j.issn.1000-3428.2013.11.012

1 概述

随着 Inernet 数据规模的增加和应用类型的丰富，海量数据的存储和分析处理给传统的系统框架带来巨大的挑战。云计算的出现和发展，打破了传统分布式垃圾邮件过滤系统的固有模式，新型分布式并行编程模型的提出，为海量数据计算处理提供了新的思路。

本文以垃圾邮件过滤问题为背景对云计算的 MapReduce 模型进行研究。现有的邮件过滤产品，主要采用贝叶斯算法、黑白名单、基于关键词和规则等[1-2]技术在传统分布式计算系统中进行实现，普遍存在集中管理难、成本高、维护困难、重复建设等问题[3]。其中，贝叶斯邮件过滤技术是一种基于内容统计的过滤技术，具有较强的文本分类能力和较高的准确性。但在传统分布式实现中，前期由大量垃圾邮件和合法邮件组成的样本集的训练过程，占用较多的系统资源和网络资源[4-5]。文献[6]利用粗糙集(Rough Set, RS)在处理不精确、不一致及不完备信息问题的有效性，提出了基于 Rough Set 的加权朴素贝叶斯分类算法，克服了朴素贝叶斯分类中的条件独立性假设问题。文献[7]提出一种最小风险的贝叶斯决策，根据误判与漏判之间的代价比值，设定阈值，进行分类决策，即根据计算得到邮件的后验概率，采用人为设定概率阈值的方法进行分类决策。文献[8]提出一种新型的最小风险的贝叶斯决策，从直线几何分割的角度改进了贝叶斯邮件分类决策模型，并定义了新的风险因子，但仍然是一种基于概率阈值的分类决策。

本文设计并实现一种基于 Hadoop 开源云架构的分布式贝叶斯邮件过滤 MapReduce 编程模型，一方面对传统贝叶斯过滤算法进行并行化改进，另一方面利用 MapReduce 模型在海量数据处理方面的优势优化邮件样本集的训练过程与待过滤邮件的过滤过程。

2 研究背景

2.1 Hadoop 云计算

云计算是网格计算、分布式计算、并行计算等传统计

基金项目：国家“863”计划基金资助项目(2009AA044601)；国家自然科学基金资助重点项目(61139002)；南京航空航天大学基本科研业务费专项基金资助项目(NS2010230)

作者简介：曾青华(1987－)，女，硕士，主研方向：云计算，并行计算；袁家斌，教授、博士、博士生导师；张云洲，硕士

收稿日期：2012-10-15 **修回日期：**2013-01-01 **E-mail：**zeng_qh@126.com

算机和网络技术发展融合的产物[1]，也是效用计算、虚拟化、硬件即服务(HaaS)、软件即服务(SaaS)、平台即服务(PaaS)等概念结合创新的结果[9]。

云计算为解决当前网络化制造存在的问题提供了新的思路和契机。所谓“算法再好，通常也难敌更多的数据”，而云计算中“移动数据的代价总是高于移动计算的代价”。这种新兴的商业计算模型整合大量计算机或服务器，组成的集群上，将大规模计算任务分布使各种应用系统能够根据需要获取计算力、存储空间和各种软件服务。其中，Hadoop[10]是 Apache 组织的一个开源的分布式系统架构，作为云计算的实现，解决了 TB 级以上数据集的存储、分析和学习问题，具有低成本、可扩展性、可伸缩性、高效性、高容错性等优点，被广泛应用于多种平台。Hadoop 的三大核心设计 MapReduce、HDFS 和 HBase，分别是 Google 云计算核心技术 MapReduce、GFS 和 BigTable 的开源实现。

2.2 MapReduce 模型

MapReduce 模型[10]是 Google 公司首先提出的一种能在大型计算机集群上并行处理海量数据的框架模型。

作为一种简化的并行计算模型，编程模型借鉴了函数式程序设计语言的设计思想，把处理并发、容错、数据分布等的细节抽象到一个库里面，将数据处理过程归结为 Map(映射)阶段和 Reduce(规约)阶段 2 个主要步骤：

(1)在 Map 阶段，MapReduce 模型以一组<*key*, *value*>作为 Map 函数的数据输入，经过映射，聚合所有具有相同的 *key* 值的中间结果的 *value* 值，产生一组中间结果<*key*1, *value*1>。

(2)在 Reduce 阶段，所有 Map 中输出的数据都传入到 Reduce 函数，函数把具有相同 *Key* 值的中间结果进行合并产生最终结果<*key*2, *value*2>。

MapReduce 模型为编程人员提供功能强大但实现简单的 API。使用这些接口，编程人员无需具备大量的分布式并行计算程序设计的经验，甚至无需关心后台复杂的任务调度和负载均衡等问题，而只需负责设计简单的业务处理逻辑流程，就可实现大规模计算任务的分布式并行执行。另外，使用这些接口，可以将应用部署到由普通 PC 构成的集群上，从而获得较高的性能。

2.3 贝叶斯邮件过滤算法

在传统的贝叶斯算法[11]中，贝叶斯决策就是根据先验概率，利用贝叶斯公式转换成后验概率，再根据后验概率大小进行决策分类。该算法有如下假定条件：

(1)m 个事件类 $D_1, D_2, \cdots, D_m$ 为事件样本空间 S 的一种划分，某一事件类 D_i 发生的概率为 $P(D_i)$，则有：$P(D_i) \geqslant 0$ $(i=1,2,\cdots,m)$，且满足 $P(D_1)+P(D_2)+\cdots+P(D_m)=1$。

(2)在假设(1)的基础上，各事件发生的概率之间不存在任何联系，即相互独立，且事件类 D_i 发生的概率对于给定类空间的影响独立于其他事件类。

(3)在假设(1)、假设(2)的基础上，对任一事件 d，$P(d) \geqslant 0$，则事件 d 属于事件类 D_i 的贝叶斯概率公式，有：

$$P(D_j \mid d)=\frac{P(d \mid D_j)P(D_j)}{P(d)} \tag{1}$$

$$P(d)=\sum_{k=1}^{m}P(d \mid D_k)P(D_k), k=1,2,\cdots,m \tag{2}$$

通过计算概率 $P(D_k \mid d)$，判断一个事件 d 的类别。假定 w_j 表示事件 d 的第 j 个特征项，根据事件 d 的特征项 $w_j(1 \leqslant j \leqslant n)$ 与事件类 D_i 的向量空间的匹配情况，决定该事件属于事件类 D_i 的概率。

假设事件中特征项 $w_j(1 \leqslant j \leqslant n)$ 出现的概率相对独立。设定 $P(w_i \mid D_k)$ 表示特征项 w_j 在事件类 D_i 中的条件概率，称 $P(w_i \mid D_k)$ 是特征项 w_j 对事件 d 属于事件类 D_i 的贡献值，$P(D_i)$ 表示一个事件属于事件类 D_i 的先验概率，有：

$$P(d \mid D_k)=P(w_1, w_2, \cdots, w_n \mid D_k)=\prod_{i=1,k=1}^{n}P(w_i \mid D_k) \tag{3}$$

对于垃圾邮件过滤问题，运用贝叶斯公式对邮件内容进行分析，过滤机制根据概率进行邮件分类：垃圾邮件类 *Spam* 或正常邮件类 *Ham*。

首先在训练过程中收集大量邮件，建立 *Spam* 类和 *Ham* 类。然后对邮件训练集合中的所有邮件数据，分别提取独立字符串 $w_1, w_2, \cdots, w_n$ 作为特征向量词汇，并统计各个词汇所出现的次数 $f_1, f_2, \cdots, f_n$。

假设所有邮件的正文部分文本的各特征词汇之间相互独立，使用贝叶斯公式，便可求出垃圾邮件在训练邮件集中所占比例，此数值即为某一封待过滤邮件 d 属于垃圾邮件 *Spam* 类的概率(先验概率)，如下：

$$P(Spam \mid d)=\frac{P(Spam)}{P(d)}\prod_{i=1}^{n}P(w_i \mid Spam) \tag{4}$$

待分类邮件 d 属于 *Ham* 类的概率(先验概率)，其值等于训练邮件集中正常邮件所占比例，如下：

$$P(Ham \mid d)=\frac{P(Ham)}{P(d)}\prod_{i=1}^{n}P(w_i \mid Ham) \tag{5}$$

其中，$P(w_i \mid Spam)$ 是指特征词汇 w_i 在垃圾邮件样本集合中出现的条件概率，等于特征词 w_i 在垃圾邮件样本集合中出现的次数 f_i 与所有特征词总数之比；$P(w_i \mid Ham)$ 是指特征词 w_i 在合法邮件样本集合中出现的条件概率，等于特征词 w_i 在合法邮件类中出现的次数 f_i 与合法邮件集合分词之后所含特征词总数之比；n 是指某一封待过滤邮件 d 在分词之后所含特征词的总个数；$P(d)$ 的计算式如下：

$$\begin{aligned}P(d)&=\sum_{k=1}^{2}P(C_k)\times P(d \mid C_k)=\\&P(Ham)\times P(d \mid Ham)+P(Spam)\times P(d \mid Spam)=\\&P(Ham)\prod_{i=1}^{n}P(w_i \mid Ham)+P(Spam)\prod_{i=1}^{n}P(w_i \mid Spam)\end{aligned} \tag{6}$$

3　贝叶斯邮件过滤算法的 MapReduce 模型

3.1　改进的贝叶斯算法

按照式(6)，传统贝叶斯邮件过滤算法在计算 $P(d)$ 过程中，耗用大量系统资源并降低了邮件过滤的效率。针对这一问题，利用 MapReduce 模型在分布式计算中的优势，改进如下：

从式(4)和式(5)可知，计算 $P(Ham\,|\,d)$ 和 $P(Spam\,|\,d)$ 的计算都需要 $P(d)$ 的数值。但对于任一封合法邮件 d，只需计算出 *Spam* 和 *Ham* 2 个类别概率的相对大小。因此，将式(4)的两端分别除以式(5)的两端，可得到 $P(Spam\,|\,d)/P(Ham\,|\,d)$，其含义可解释为某一待过滤邮件 d 属于 *Spam* 垃圾邮件类和 *Ham* 正常邮件类的概率之比，如式(7)所示。

$$\frac{P(Spam\,|\,d)}{P(Ham\,|\,d)}=\frac{P(Spam)\prod_{i=1}^{n}P(w_i\,|\,Spam)}{P(Ham)\prod_{i=1}^{n}P(w_i\,|\,Ham)} \tag{7}$$

设 $K=P(Spam\,|\,d)/P(Ham\,|\,d)$，取某一阈值 Q，若 $K\geqslant Q$，则将该邮件判为垃圾邮件，否则判为正常邮件。

例如，取 $Q=0.8$，计算某一封邮件的 $P(Spam\,|\,d)/P(Ham\,|\,d)$，得 K=1.32，由于 K 大于 Q 值，因此，将该邮件判为垃圾邮件。

3.2　贝叶斯算法的 MapReduce 模型

根据上述的贝叶斯邮件过滤算法，设计邮件过滤流程，由基于内容的邮件训练模块和贝叶斯过滤模块组成，如图 1 所示。针对这一流程，分别设计训练部分的 MapReduce 模型和过滤部分的 MapReduce 模型。

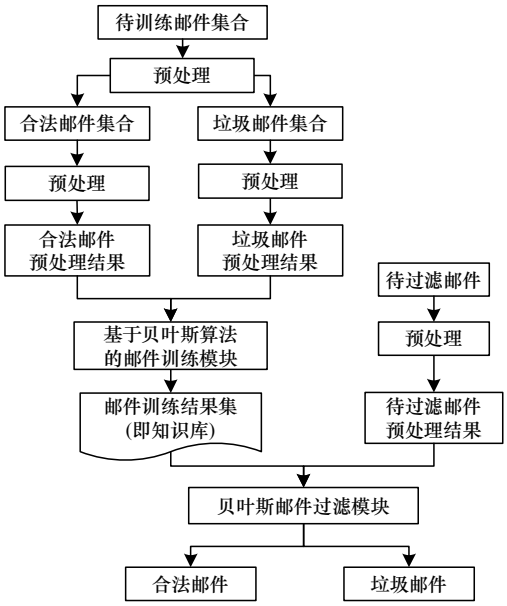

图 1　贝叶斯邮件过滤算法

在贝叶斯邮件过滤算法的传统实现中，对合法邮件集合、垃圾邮件集合、带过滤邮件集合的预处理过程同样占用一定的时间与计算资源。本文利用 MapReduce 模型对预处理过程进行改进。如图 2 所示。

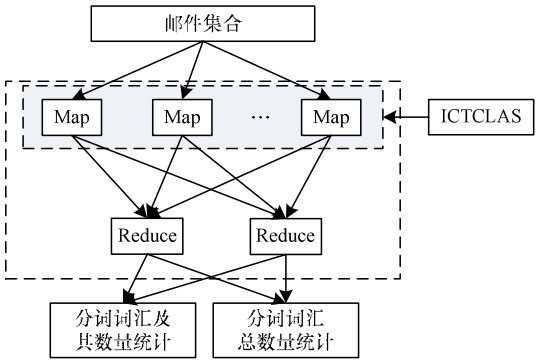

图 2　邮件预处理 MapReduce 过程

在该过程中，每个 Map 函数接收一个邮件数据块，以<数据块序号，邮件数据块>作为输入键值对，对每个数据块进行分词，映射到<分词词汇,1>键值对，并作为中间结果输出。每个 Reduce 函数接收具有相同 *Key* 值的中间结果，合并 *Value* 值，即统计每个 *Key* 值的个数，得到并输出<分词词汇,数量统计>键值对。其中，分词过程使用 ICTCLAS 中文分词器对邮件文本数据进行分词处理。中国科学院计算技术研究所研制的汉语词法分析系统 ICTCLAS 的主要功能包括中文词汇和语句的分词、词性标注、命名实体识别、新词识别等。测试实验结果表明，使用该系统的分词结果有 98.45%的正确率。因此，本文使用 ICTCLAS 提供的 API 进行邮件正文文本分词，为后续反垃圾邮件的样本训练和过滤操作的实现提供良好基础。

邮件训练部分包含 3 个 MapReduce 过程，如图 3 所示。

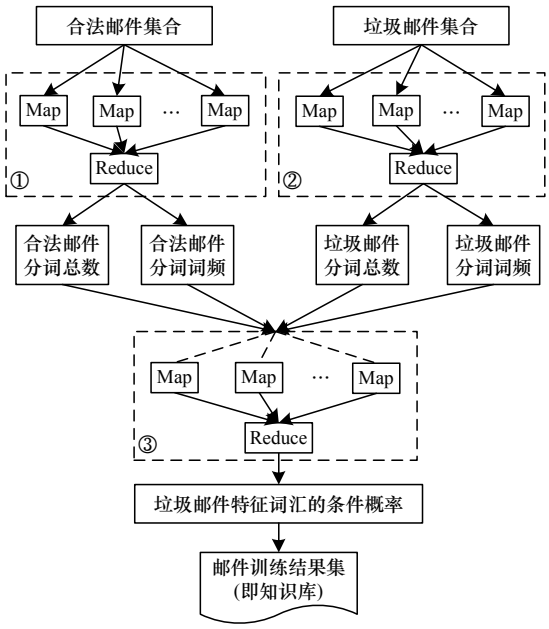

图 3　邮件训练部分 MapReduce 过程

第 1 个、第 2 个 MapReduce 过程分别接收经过预处理的合法邮件和垃圾邮件数据块，各自计算邮件数据的词汇总数及每个词汇的词频。第 3 个 MapReduce 过程接收前两个的 MapReduce 过程生成的中间数据，计算得到各类别先验概率和各词的特征权值得到贝叶斯分类模型。

邮件过滤部分包含 2 个 MapReduce 过程，如图 4 所示。

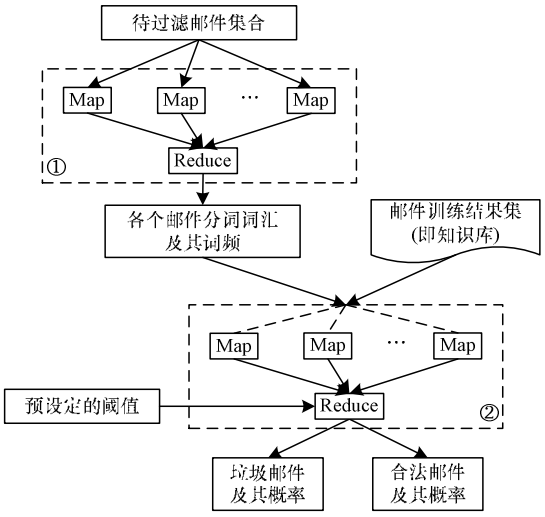

图 4 邮件过滤部分 MapReduce 过程

第 1 个 MapReduce 过程分别接收经过预处理的待过滤邮件数据，计算邮件数据的词汇总数及每个词汇的词频。第 2 个 MapReduce 过程接收第 1 个的 MapReduce 过程生成的中间数据及邮件训练结果集，计算每封邮件的 $K=P(Spam|d)/P(Ham|d)$ 值，再依据预设定的阈值 Q，对其进行判断，分别输出合法邮件和垃圾邮件的列表及其概率。

4 实验与结果分析

实验基于 Hadoop 云计算平台进行研究，Hadoop 云计算平台的配置结构如图 5 所示。

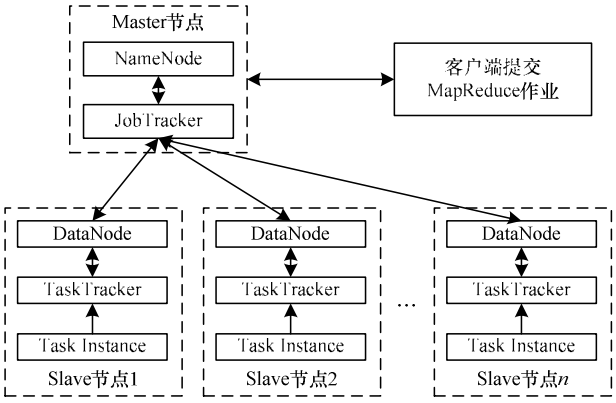

图 5 Hadoop 云计算平台的配置结构

Hadoop 集群中各节点配置如下：Intel Pentium Dual CPU、1.6 GHz、1 024 MB Memory，并采用 Ubuntu 64 位操作系统、jdk-6u19-linux-i586 版本的 jdk。根据 Hadoop 项目官方网站介绍的方法配置基于 Hadoop 版本的集群。其中，随机选取一台主机作为 Master 主节点，启动 NameNode 和 JobTracker 进程，其余主机作为 DataNode 和 TaskTracker，为 Slave 从节点。

实验数据使用中文自然语言处理开放平台(http://www.nlp.org.cn/docs/download.php?doc_id=1207)和中国教育和科研计算机网紧急响应组(http://www.ccert.edu.cn/spam/sa/datasets.htm)提供的邮件数据集，共计 37 360 封，其中垃圾邮件 26 588 封、合法邮件 10 772 封。

考虑到邮件数量问题会影响实验精确度，整合 2 个邮件数据源，在合法邮件中随机剔除 2 封邮件，在垃圾邮件中随机剔除 3 封邮件，再分别均分为 5 份，每份包含 2 154 封合法邮件和 53 176 封垃圾邮件，再编号为 a、b、c、d、e。

针对邮件过滤部分，本组实验采用 4 个评价指标：垃圾邮件查准率 P、垃圾邮件召回率 R、F 值以及全部邮件判对率 T。其中，查准率 P 反映识别垃圾邮件的准确性，其数值等于实为垃圾邮件的邮件数量在判为垃圾邮件的总数之比；召回率 R 反映识别垃圾邮件的完整性，其数值等于实为垃圾邮件的邮件在真实垃圾邮件中的比例，F 值为 $2PR/(P+R)$，兼顾了查准率和召回率 2 个指标，是这 2 个指标的综合反映；判对率 T 等于所有待分类邮件被正确归类的邮件的比例，反映正确归类邮件的能力。

本组选取 b、c、d、e 这 4 组为邮件训练样本，a 为待过滤邮件样本。对所有邮件数据进行实验并将结果进行分析，计算查准率、邮件召回率、F 值和判对率，如表 1 所示。

表 1 邮件过滤性能比较 (%)

模型	查准率	召回率	F 值	判对率
传统分布式计算模型	95.86	93.17	94.77	96.03
本文模型	95.73	94.51	95.22	97.44

可以看出，贝叶斯邮件过滤算法在传统分布式计算模型和基于 Hadoop 的 MapReduce 模型上实现的效果相当。在 Hadoop 云平台上使用 MapReduce 模型实现贝叶斯邮件过滤在查准率、召回率、F 值、判对率方面未造成不良影响。

得到上述的在计算实验时间性能后，本文针对云计算的海量数据处理方面的能力进行实验。使用上述实验中邮件样本训练所得的垃圾邮件知识库，再对 5 组待过滤邮件进行实验。待过滤邮件样本集分别由 a、d、a+d、a+b+d 这 5 组不同邮件集合组合而成。实验结果如表 2 所示。

表 2 海量数据处理性能比较

测试数据组	时间/s		加速比
	传统分布式计算模型	本文模型	
a	4.402	2.838	1.550
d	4.492	2.889	1.555
a+d	8.137	5.054	1.610
a+b+d	11.899	7.282	1.634
a+b+c+d	15.187	9.199	1.651

可以看出，a 组和 d 组的邮件数量相同，邮件过滤的运行效率也大体相同，因此，可在此基础上扩展待过滤邮件集合的规模。

由上述结果可知，在邮件过滤的贝叶斯算法实现方面上，对贝叶斯邮件过滤算法的分布式并行化改进，可减少贝叶斯过滤判定公式的计算量，基于 Hadoop 的 MapReduce 模型对比传统分布式并行模型，减少了邮件过滤的运行时间，提高了执行效率，使得总体性能有了一定程度的提高。

(下转第 64 页)