

# 一种改进的贝叶斯算法在短信过滤中的研究

裴向杰,唐红昇,陈 鹏

(江苏省气象局 江苏省气象信息中心,江苏 南京 210008)

**摘要:**随着信息交流的频繁性,各种骚扰和垃圾短信充斥手机,严重干扰了人们的正常生活。针对垃圾短信过滤技术,研究基于最小风险决策贝叶斯的文本分类器构造方法以及实现。对于朴素贝叶斯在短信过滤系统中过分依赖样本空间的分布和内在的不稳定性,造成了时间复杂度的增加,提出了一种基于改进贝叶斯的垃圾短信文本分类器构造方法。主要利用最小风险决策算法结合贝叶斯理论完成对批量短信的训练,形成对应的集合模型。对实现文本分类的关键技术做了重点叙述,并对文本分类算法进行了实现。最后对算法进行测试,结果表明:基于最小风险决策贝叶斯的文本分类器不仅训练简单,而且分类准确度高,解决了朴素贝叶斯算法的不稳定性,为短信过滤技术提供了借鉴。

**关键词:**贝叶斯;最小风险;文本分类;短信过滤

**中图分类号:**TP301.6

**文献标识码:**A

**文章编号:**1673-629X(2015)09-0089-05

**doi:**10.3969/j.issn.1673-629X.2015.09.019

## Research on Optimized Naive Bayesian Algorithm in SMS Spam Filtering

PEI Xiang-jie, TANG Hong-sheng, CHEN Peng

(Information Center of Meteorological Bureau of Jiangsu Province, Nanjing 210008, China)

**Abstract:** With frequent exchanges of information, various harassing messages with mobile phone disturb the normal life for people. For spam filtering technology, research the constructing method and its realization for text classifier based on optimized Naive Bayesian algorithm. The distribution of Naive Bayesian over-reliance on sample space in the short message filtering system and the inherent instability cause an increase in time complexity, propose a spam message structure text classifier based on the improved Bayesian method. The method uses the Bayesian theory and minimum risk decision algorithm to complete the training of bulk SMS. Describe the key technologies of text classification and implement the text classification algorithm. The test results show that the new algorithm can easily train and improve the classification accuracy, solving the instability of Naive Bayesian algorithm, which provides a reference for filtering technology.

**Key words:** Bayes; minimum risk; text classification; message filtering

## 0 引言

科学技术的快速发展,带来信息交流的快速性,造成社会的各个领域都充斥着各种信息的交流,而手机作为信息交流的一种成熟载体被广泛应用。随着手机的频繁使用,手机接收信息内容出现了一些新的情况,主要表现在会收到一些骚扰信息、不良信息甚至违法信息,而这些信息存在的原因主要是由于短信成本较低和我国相关法律监管的缺失造成。这些信息不仅严重干扰了人们的正常生活,也给其他方面带来了不利影响。从手机使用者的角度来看,垃圾短信给用户带来情绪和心情上的波动,影响和干扰了人们正常的工

作和生活,查看垃圾短信也是对时间的一种浪费;从通信资源来看,大量的垃圾短信占用了大量的通信信息空间,在有线的通信资源状况下,会造成通讯的拥堵和不畅,对于通信公司来说,影响其业务的发展和运转;从社会稳定性上来看,大部分的垃圾短信都是一些负面信息,具有对一般民众的欺骗或者诱导作用,易造成不满或者不安的个人情绪,败坏了社会环境。为了实现垃圾短信的整治,消除对部分人群造成的不良影响,即受到手机的不良信息的影响,国外一些国家已经在法律层面上对垃圾短信进行了立法,做到对垃圾短信治理过程中的有法可依,同时有的国家还成立了相关机构,专门负责对垃圾短信的监控、处理和惩治,具

收稿日期:2014-10-23

修回日期:2015-01-26

网络出版时间:2015-08-26

基金项目:国家自然科学基金资助项目(60802034)

作者简介:裴向杰(1977-),男,硕士,工程师,研究方向为网络管理、软件开发。

网络出版地址:<http://www.cnki.net/kcms/detail/61.1450.TP.20150826.1535.016.html>

有很强的针对性。在我国,为了进化信息传递过程中的健康性,颁布了相关的条例来约束垃圾短信的发送,并成立了受理垃圾短信的信息中心,有效抑制了垃圾短信的传播。此外,为了更加有效地拦截垃圾短信,从自动过滤系统上面去考虑对垃圾短信的阻拦。据中国互联网组织调查,2012 年我国手机用户收到的垃圾短信数量高达 13.6 条/周<sup>[1]</sup>,因此建立有效的垃圾短信过滤系统显得尤为重要,具有重要的经济价值和社会效益。

近年来,我国手机用户通过各种方式来拒绝垃圾短信,目前常用的垃圾短信过滤方法有两种方式:黑白名单过滤方法和基于统计文本分类的智能短信过滤<sup>[2]</sup>。黑白名单过滤方法原理较为简单,即将拒绝短信发送方的号码设置为黑名单,这种方式使用起来简单,但是如果短信发送方采用动态号码或者号码欺骗的方式发送短信,则该方式将无法过滤。智能短信过滤系统是根据短信文本的内容进行分类,将短信分成正常短信和垃圾短信两大类。目前基于统计文本分类的过滤方法有决策树和贝叶斯推理过滤法<sup>[3-4]</sup>。其中,贝叶斯过滤算法简单且高效,广泛运用于垃圾短信过滤技术中。

朴素贝叶斯过滤算法通过对类别和词的概率的计算,来对给定文档进行判别,并依据概率设定阈值来实现对文档的归类<sup>[5-6]</sup>,但该方法具有一定的局限性。文中通过对目前垃圾短信过滤系统的分析与研究,针对贝叶斯在垃圾短信过滤系统中应用的局限性,提出了一种新型的基于最小风险决策贝叶斯的分类算法,该算法能够有效提高垃圾短信过滤的准确度。

## 1 关键技术

### 1.1 朴素贝叶斯

朴素贝叶斯作为一种重要的分类技术,是统计学内容中的一种非常重要的数学处理技术,可以根据预测的概率判别预测成员的种类。而朴素贝叶斯判别模型或者分类模型来自于各成员在各类别出现规律的统计。朴素贝叶斯分类器是一种稳定且分类效率较高的分类器,具有算法简单、待估参数较少和对数据不敏感的特点,与其他分类方法相比,具有最小的误差特点。

作为一种数学统计方法,贝叶斯可以预测成员关系的可能性。文中对于垃圾短信的分类过程描述如下:

(1) 假设短信样本表达方式是:使用一个  $n$  维特征向量  $C = C_n (n = 1, 2, \dots)$  表示对  $n$  个样本的  $n$  个度量。

(2) 假设  $V = (v_1, v_2, \dots, v_n)$  表示短信类别,一般分

为正常短信类别、垃圾短信类别。对于一个不知道类型的短信  $C$ , 可以通过式(1) 计算  $C$  属于  $V_i$  的后验概率。

$$P(V_i | C) = \frac{P(C | V_i)P(V_i)}{P(V)} \quad (1)$$

其中,  $P(V_i)$  是  $V_i$  的先验概率;  $P(V)$  是训练集  $C$  的先验概率;  $P(C | V_i)$  是分类  $V_i$  样本数据  $C$  的概率;  $P(V_i | C)$  是  $V_i$  的后验概率。

(3) 当系统特征较多的情况下,  $V_i$  的后验概率  $P(V_i | C)$  计算开销比较大,此时假定样本之间不存在依赖关系,即样本变量  $C_i$  之间相互独立,则可由式(2) 计算出联合样本特征向量  $C$  的概率。

$$P(C | V_i) = \prod P(C_i | V_i) \quad (2)$$

使用贝叶斯方法对短信进行分类过滤时具有一定的局限性,即对垃圾短信与正常短信误判<sup>[5]</sup>。

### 1.2 最小风险贝叶斯决策

贝叶斯决策是对未知部分的判断估计过程,同时也是对事件发生概率不断修正的过程,其特点主要在于估计过程存在风险,能够掌握除了客观因素外的其他因素变化带来的概率分布变化,得出未来的期望状况的决策准则。其基本思想主要包括概率密度参数的表达式,后验概率以及决策分类。

通过文本特定特征将文本归类,即通过机器学习来获取分类的规则,根据文本特点分配到设定的类别中<sup>[7-9]</sup>。在分类过程中,分析短信分类的随机性和不确定性,故在短信文本分类时,不能对某条短信进行肯定的分类,只能说其属于某个类别的概率有多大<sup>[10]</sup>。采用朴素贝叶斯算法构造的分类器容易造成过低估计概率的情况,针对这一问题,文中分类器采用  $m$ -估值算法进行改进。

贝叶斯方法处理文本分类使用两种机制<sup>[11-12]</sup>, 分别为:选择后验概率最大的类别为目标类别和选择效用最大的类别为目标类别。

分类的目的就是把  $C$  归入到某个类别  $v_i$  中。第一种方法选择后验概率最大的类别,即  $P(v_i | c) \geq P(v_j | c), j \in (1, 2, \dots, n)$ , 此时取判别函数:  $r_i(x) = p(c_i | x)$ 。第二种方法是根据文本不确定的因素大小判断,利用平均效益的值去评估决策风险的过程。假定特征向量  $C$  划分到类别  $c_j$  (错误判断)的损失为  $L_{ij}(X)$ , 其正确所属类别为  $v_i$ , 即:

$$\min \left\{ \sum_{j=1}^n L_{ij}(x) \times P(c_j | x) \right\}$$

此时取判别函数:

$$R_i(x) = \sum_{j=1}^n L_{ij}(x) \times P(C_j | x) \quad (3)$$

由式(3)可知,正确分类损失为零时,对角线

$L_{ij}(x)$  为 0,反之,非对角线值影响着错误分类的损失。

文中将短信分为诈骗短信、反动短信、广告短信、非健康短信和正常短信五类。因此,在接收到一条短信时,该未知短信的分配类别就有 5 种,因此不同类别之间的决策就有 25 种情况。

表 1 为基于最小风险的贝叶斯决策理论定义的决策表。

表 1 贝叶斯决策表					
	正常短信	广告短信	反动短信	黄色短信	诈骗短信
正常短信	0	$K_1$	$K_1$	$K_1$	$K_1$
广告短信	$K_2$	0	$K_3$	$K_3$	$K_3$
反动短信	$K_2$	$K_3$	0	$K_3$	$K_3$
黄色短信	$K_2$	$K_3$	$K_3$	0	$K_3$
诈骗短信	$K_2$	$K_3$	$K_3$	$K_3$	0

从表 1 可知,正常短信的损失定义为 0,将垃圾短信误判为正常短信过程,其中的损失定义为  $K_2$ 。正常短信判为垃圾短信过程,用  $K_1$  代表其中的损失。垃圾短信之间的相互误判损失定义为  $K_3$ 。对比三种损失,一般情况是正常短信误判为垃圾短信的损失要远大于另外两种,垃圾短信误判为正常短信的损失次之,垃圾短信之间的相互误判给用户带来的损失最小,因此  $K_1 > K_2 > K_3$ 。

一般情况下,垃圾短信之间的误判给用户带来的损失可以忽略不计,因此为了简化计算,降低算法的时间复杂度,定义  $K_3 = 0$ ;将垃圾短信误判为正常短信的损失定义为  $K_2 = 1$ ;将正常短信误判为垃圾短信的损失定义为  $K_1 > 1$ 。

根据式(3),将短信  $x$  判定为正常短信的条件风险为:

$$\begin{aligned} R(\text{normal}) &= 0 \times P(c = 0 | x) + K_2 \times P(c = 1 | x) \\ &\quad + K_2 \times P(c = 2 | x) + K_2 \times \\ &\quad P(c = 3 | x) + K_2 \times P(c = 4 | x) = \\ &\quad K_2 \times [1 - P(c = 0 | x)] \end{aligned}$$

将短信  $x$  判定为垃圾短信的条件风险为:

$$\begin{aligned} R(\text{spam}) &= K_1 \times P(c = 0 | x) + 0 \times P(c = 1 | x) + \\ &\quad K_3 \times P(c = 2 | x) + K_3 \times \\ &\quad P(c = 3 | x) + K_3 \times P(c = 4 | x) = \\ &\quad K_1 \times P(c = 0 | x) + K_3 \times [1 - P(c = \\ &\quad 0, 1 | x)] \end{aligned}$$

根据前面定义的  $K_3 = 0, K_2 = 1$  可以得出公式(4)、(5)。式(4)为将短信  $x$  判定为正常短信的条件风险;式(5)为将短信  $x$  判定为垃圾短信的条件风险。

$$R(\text{normal}) = \sum_{j=1}^4 P(C_j | x) \tag{4}$$

$$R(\text{spam}) = K_1 \times [1 - \sum_{j=1}^4 P(C_j | x)] \tag{5}$$

假设决策为垃圾短信,则此时判定为垃圾短信的条件风险要小于判定为正常短信的条件风险,即  $R(\text{normal}) > R(\text{spam})$ ,带入公式(4)、(5)可以得出

$$\sum_{j=1}^4 P(C_j | x) > K_1 \times [1 - \sum_{j=1}^4 P(C_j | x)]$$

,通过计算得出:

$$\sum_{j=1}^4 P(C_j | x) > \frac{K_1}{K_1 + 1} \tag{6}$$

当决策为四类垃圾短信并且满足式(6)时,可以保证决策为垃圾短信的风险比决策为正常短信的风险小。此时,根据  $K_1$  的值可以定义短信的阈值,从而保证决策风险最小。在这里,假设  $K_1 = 1$ ,那么短信的阈值便为 0.5;假设  $K_1 = 4$ ,那么短信的阈值便为 0.8;假设  $K_1 = 9$ ,那么短信的阈值便为 0.9。

## 2 贝叶斯分类器的实现

### 2.1 算法分析

应用贝叶斯分类器时需要解决两大问题:一是对贝叶斯分类器所需的概率的估算方法;二是对文档的属性值的建立过程。

例如某条短信其内容如下:

这/是/关于/贝叶斯/分类器/的/一篇/论文/贝叶斯/属于/统计学/分类方法/可/预测/成员关系/的/可能性

以上短信文本中,对应的第一个属性为“这”,第二个属性为“是”,第三个属性为“关于”,第四个属性为“贝叶斯”,第五个属性为“分类器”,等等。总共 17 个属性分类。

按照 Naive 贝叶斯的假设<sup>[13-14]</sup>,文本的各个属性是相互独立的,但是这一假设具有一定的局限性。例如“论文”出现的概率会因为它前面的属性“一篇”而增大,因此文中采用的是一种新型的 m-估值法来完成贝叶斯机器学习算法,估计其概率。因为短信文本是巨大的,每个短信文本的先验概率是无法知道的,因此采用均匀分布法计算先验概率,对应的概率值公式为:

$$P(A_i | C_i) = \frac{n_k + l}{n + |\text{Vocabulary}|} \tag{7}$$

其中,  $n$  为训练集中词条的位置总数;  $n_k$  为词条  $k$  在  $n$  个单词中出现的次数;  $l$  为调节项; Vocabulary 为词条总数。式中  $n_k + l$  的值是固定的,  $n_k$  与  $l$  之间的变化是相反的,  $l$  增加,则  $n_k$  减少,反之亦然。

### 2.2 算法实现

文中分类算法流程如下:给定一组短信样本,对样本内容进行分类,其中  $C$  为短信样本的特征向量。贝叶斯分类器首先学习概率项,该概率项描述从特征向量  $C$  文档  $C_i$  中某个位置随机抽取一个词条的概率。

算法的流程如图 1 所示。

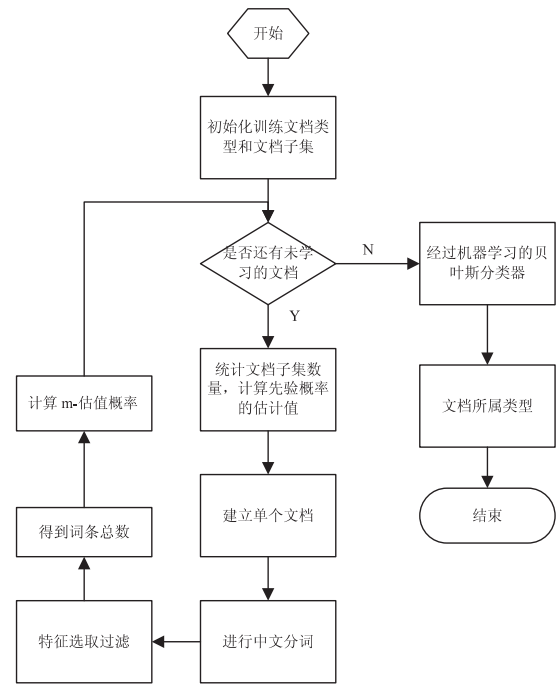


图 1 贝叶斯机器学习与分类流程

根据以上的算法流程,首先对文档的类型和子集进行训练。然后对文档初始化训练进行检测,如果发现存在未训练的文档,对文档的子集数进行统计,计算先验概率的估计值,并建立单个文档,对文档进行识别并基于特征过滤,最后得到  $m$  估值概率;如果未发现未学习的文档,直接对文档进行贝叶斯分类,建立文档属性。至此,整个算法完成。

根据上述流程实现该贝叶斯分类器的主要接口如下:

```
Public interface Learner{
Public void learn( HashMap results,int category)
Public void GetLastResults();
Public void SaveResults();
Public HashMap GetTerms();
}

Public interface IntelligentClassifier{
//LinkedList
Publicdouble[] getClassierInfo( LinkedList words);
Public static void machineLeaning( LinkedList words,int category);
}
```

3 算法测试

根据文中提出的贝叶斯分类器实现方法,本次测试选择 1 600 条短信样本进行实验,采用预留法作为测试方法,将 1 600 条短信分为 5 份,每份大约 320 条,每次选择其中 1 份为训练集。首先对朴素贝叶斯分类器进行测试,得到的短信样本分类结果见表 2。

表 2 朴素贝叶斯分类器短信样本分类结果

类别	样本数	分类数量	正确分类	查准率	查全率
正常	320	297	276	0.925	0.857
广告	320	328	292	0.879	0.904
反动	320	330	268	0.827	0.845
色情	320	321	299	0.901	0.910
诈骗	320	324	284	0.871	0.889

从表 2 中可以看出,对于正常、广告、反动、色情、诈骗五种类别的短信进行识别分类,在查全率方面基本能够达到 85% 以上,只有对反动短信查全率较低;在查准率方面,同样除了在反动方面的准确率不高,对其他垃圾短信的查准率在 90% 左右。为对比文中算法的有效性和优越性,对实验样本进行  $m$ -估值最小风险贝叶斯分类器测试,测试的结果如表 3 所示。

表 3 短信样本分类结果(最小风险决策贝叶斯)

类别	样本数	分类数量	正确分类	查准率	查全率
正常	320	338	309	0.892	0.958
广告	320	326	299	0.895	0.914
反动	320	314	286	0.827	0.877
色情	320	302	293	0.901	0.925
诈骗	320	320	301	0.937	0.898

由表 3 可以看出,最小风险决策贝叶斯分类器对短信的过滤作用,无论是在查准率方面还是在查全率方面,准确率都达到 90% 左右,特别是在对正常短信的查全率方面达到 95% 以上,远远大于朴素贝叶斯分类器对正常短信分类的准确率 85%,同时对诈骗信息上也表现出较高的查准率。为了能够比较直观地对比两种算法的性能,根据表 2 和表 3 得到如图 2 所示的两种过滤器的对短信的分类效果。

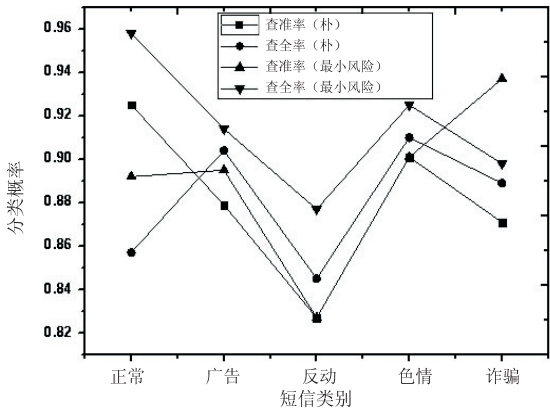


图 2 两种算法对短信分类的效果

图中,能够比较清晰地看出两种算法对不同类型的短信过滤作用,对比显示文中设计的最小风险决策贝叶斯分类器,将样本判定为正常短信的概率增加,使得正常短信的查全率得到较大的提高;将短信样本判



定为垃圾短信的风险加大,使得垃圾短信的查准率得到提高。

4 结束语

文中针对传统垃圾短信过滤系统中存在的主要问题,设计并实现了最小风险决策的贝叶斯分类算法,使得对垃圾短信判定为正常短信的概率降低,正常短信误判过程减少,即认为正常短信是垃圾短信的概率降低。通过实验表明,该算法不仅训练简单,而且分类准确度高,具有很强的自我学习能力,解决了朴素贝叶斯算法的局限性,大大提高了垃圾短信的拦截率以及降低了短信的误判率。

参考文献:

[1] 张东亮,董礼. 基于改进的朴素贝叶斯算法在垃圾短信过滤中的研究[J]. 计算机测量与控制,2012,20(2):526-528.

[2] Zhang Harry, Su Jiang. Naïve Bayes for optial ranking[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2012,20(2):79-83.

[3] 范晶,刘菊新,陈峰,等. 基于 Hash 函数和贝叶斯方法的垃圾短信在线过滤系统[J]. 计算机应用,2008,28(4):1084-1086.

[4] Ren J T, Lee S D, Chen X L, et al. Native Bayes classification of uncertain data[C]//Proc of 9th IEEE international conf on

data mining. Miami, FL: IEEE, 2009: 944-949.

[5] Hagmayer Y, Osman M. From colling billiard balls to colluding desperate housewives: causal Bays nets as rational modes of everyday causal resoning [J]. Synthese, 2012, 189(1): 89-92.

[6] 蔡嵩,张建明,陈继明,等. 云计算环境中基于朴素贝叶斯算法的负载均衡技术[J]. 计算机应用,2014,34(2):360-364.

[7] 张依杨,向阳,蒋锐权,等. 朴素贝叶斯算法的 MapReduce 并行化分析与实现[J]. 计算机技术与发展,2013,23(3):23-26.

[8] Cantwell J. On an alleged counter-example to causal decision theory[J]. Synthese, 2010, 173(2): 127-152.

[9] 段震,王倩倩,张燕平,等. 覆盖算法下文本分类特征选择的研究[J]. 计算机技术与发展,2008,18(11):29-31.

[10] 樊建聪,张问银,梁永全. 基于贝叶斯方法的决策树分类算法[J]. 计算机应用,2005,25(12):2882-2884.

[11] 邸鹏,段利国. 一种新型朴素贝叶斯文本分类算法[J]. 数据采集与处理,2014,29(1):71-75.

[12] 杜选. 基于加权补集的朴素贝叶斯文本分类算法研究[J]. 计算机应用与软件,2014,31(9):253-255.

[13] 徐英慧,刘梅彦. 基于内容的手机端垃圾短信过滤策略研究[J]. 北京信息科技大学学报:自然科学版,2013,28(1):51-55.

[14] 钟延辉,傅彦,陈安龙,等. 基于抽样的垃圾短信过滤方法[J]. 计算机应用研究,2009,26(3):933-935.

(上接第88页)

究与改进[J]. 计算机应用与软件,2009,26(1):146-149.

[2] 万宇文,黄林颖,甘登文. 基于权值的关联规则挖掘改进算法[J]. 计算机与现代化,2014(4):73-76.

[3] 钱光超,贾瑞玉,张然,等. Apriori 算法的一种优化方法[J]. 计算机工程,2008,34(23):196-198.

[4] 李云峰,陈建文,程代杰. 关联规则挖掘的研究及对 Apriori 算法的改进[J]. 计算机工程与科学,2002,24(6):65-68.

[5] 余文礼. 基于 Apriori 算法和关联度指标的购物篮分析[J]. 科技视界,2014(4):56-57.

[6] 黄立勤,柳燕煌. 基于 MapReduce 并行的 Apriori 算法改进研究[J]. 福州大学学报:自然科学版,2011,39(5):680-685.

[7] 丁卫平. 关联规则挖掘 Apriori 算法的改进及其应用研究[J]. 南通大学学报:自然科学版,2008,7(1):50-53.

[8] Woo Jongwook, Xu Yuhang. Market basket analysis algorithm with Map/Reduce of cloud computing[C]//Proc of the 2011 international conference on parallel and distributed processing techniques and applications. Las Vegas: IEEE, 2011: 18-21.

[9] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [C]//Proceedings of

the 1993 ACM SIGMOD conference. Washington DC, USA: ACM, 1993: 207-216.

[10] Lanm Qihua, Zhang Defu, Wu Bo. A new algorithm for frequent itemsets mining based on Apriori and FP-Tree [C]//Proc of global congress on intelligent systems. Xiamen, China: Computer Science, 2009: 360-364.

[11] He Chen, Weitzel D, Swanson D, et al. HOG: distributed Hadoop MapReduce on the grid[C]//Proc of 2012 SC companion: high performance computing, networking storage and analysis. Salt Lake City: IEEE, 2012: 1276-1283.

[12] Li Ning, Zeng Li, He Qing, et al. Parallel implementation of Apriori algorithm based on MapReduce [C]//Proc of 2012 13th ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing. Kyoto: IEEE, 2012: 236-241.

[13] Wasi-ur-Rahman M, Islam N S, Lu Xiaoyi, et al. High-performance RDMA-based design of Hadoop MapReduce over InfiniBand [C]//Proc of 2013 IEEE 27th international symposium on parallel & distributed processing workshops and PhD forum. Cambridge, MA: IEEE, 2013: 1908-1917.

# 一种改进的贝叶斯算法在短信过滤中的研究

作者：[裴向杰](#)，[唐红昇](#)，[陈鹏](#)，[PEI Xiang-jie](#)，[TANG Hong-sheng](#)，[CHEN Peng](#)

作者单位：[江苏省气象局 江苏省气象信息中心, 江苏 南京, 210008](#)

刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2015 (9)

引用本文格式：[裴向杰](#).[唐红昇](#).[陈鹏](#).[PEI Xiang-jie](#).[TANG Hong-sheng](#).[CHEN Peng](#) [一种改进的贝叶斯算法在短信过滤中的研究](#)[期刊论文]-[计算机技术与发展](#) 2015 (9)