基于朴素贝叶斯算法的垃圾邮件过滤系统的研究与实现

王 斌
（商洛学院 陕西 商洛 726000）

摘要：针对垃圾邮件泛滥的问题，本文基于朴素贝叶斯算法构建了邮件过滤系统，并采取平滑、归一化等方法进行数据预处理，提取结构与统计特征，通过邮件地址、邮件内容等多个方式进行过滤。计算机测试后表明本算法提高了垃圾邮件识别精度与准确率。

关键词：模式识别；邮件过滤；朴素贝叶斯；数据预处理
中图分类号：TP311 **文献标识码：**A **文章编号：**1674-6236(2018)17-0171-04

Research and implementation of spam filtering system based on Naive Bayes algorithm

WANG Bin
(*Shangluo University*, *Shangluo* 726000, *China*)

Abstract: In order to satisfy the requirement of the spamrecognition, a professional recognition system based on Naive Bayesis realized in this paper. The learning samples are smoothed and normalized. The spam is filtered by the e-mail address, e-mail content and other waysto improve the accuracy .The results show that the algorithm can improve the precision and accuracy of spam detection.

Key words: pattern recognition; spam filtering; NB; data preprocessing

随着信息技术的发展，电子邮件凭借其快捷、便利、低成本的优点，成为互联网中的主要信息传播与交流方式。但是，在电子邮件便利了生活的同时，也带来了很多的问题与烦恼，其中最严重的莫过于垃圾邮件的泛滥。垃圾邮件中蕴含了大量的商业传销、色情信息甚至邮件病毒，为电子邮件用户带来了诸多的负面影响，一些不良信息会欺骗诱导用户做出错误的判断与决定，随着互电子邮件传播的病毒会对用户的计算机造成重大损伤。同时，垃圾邮件还恶化了互联网环境，滥用网络带宽与计算机存储资源。毫不夸张的说，垃圾邮件对于计算机网络、社会和国家安全造成了严重的威胁。不法分子利益的驱使与客观存在的技术漏洞，加上相关法律的不健全导致了垃圾网络的肆意传播。在本文中，为了遏制垃圾邮件的传播，深入研究了垃圾邮件的过滤技术。基于朴素贝叶斯算法，提取邮件的地址、主题与内容信息，进行特征工程与邮件类别判定，帮助用户远离垃圾邮件干扰。

收稿日期：2018-01-23 **稿件编号：**201801116
作者简介：王 斌(1975—)，男，陕西商洛人，工程师。研究方向：计算机科学与技术。

1 系统简介

1.1 贝叶斯技术

贝叶斯算法是一种基于概率分析事件发生可能性的方法，其核心思想在于选择发生概率高的作为分类的结果。贝叶斯算法名字来源于十八世纪著名神学家“托马斯.贝叶斯”。在了解贝叶斯原理前，先介绍几个基本概念。

条件概率：假设以 $P(A)$ 代表事件A的发生概率；以 $P(B)$ 代表B事件所发生的概率；那么就可将事件A发生下的事件B发生的概率看作为 $P(B|A)$ ；将 $P(AB)$ 看作A事件与B事件同时发生概率。则 $P(B|A)=\frac{P(AB)}{P(B)}$ 。

全概率公式：假设将试验E的样本空间表示为 Ω ；B代表着样本空间 Ω 的事件；$A_1,A_2,\cdots,A_n$ 代表 Ω 的一个划分，$P(A_i)>0$ 。如此以来，将会得出全概率公

式为：

$$P(B)=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(B|A_n)P(A_n)=\sum_{i=1}^{n}P(A_i)P(B|A_i) \quad (1)$$

全概率公式常被用于复杂概率的计算中，采用该公式的思想进行概率问题分析时，主要通过将较为复杂的概率问题细分为n个简单事件集合，并在此基础上对概率的可加性进行利用，以此计算出目标事件发生的相关概率。

根据条件概率与全概率公式，可以得出贝叶斯公式。与全概率公式的前提一样，将Ω看作是试验E的样本空间；以B代表样本空间Ω的事件；$A_1,A_2,\cdots,A_n$代表Ω的一个划分，$P(A_i)>0$。以此便能得出贝叶斯公式：

$$P(A_i|B)=\frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n}P(B|A_j)P(A_j)},\quad i=1,2,\cdots,n \quad (2)$$

贝叶斯定理，又称贝叶斯法则。对于事件A与B，事件A发生下的事件B的条件概率$P(B|A)$与事件B发生下事件A的条件概率往往是不一样的，但二者间存在着固定的关系，文中用贝叶斯准则来描述二者间的关系。在机器学习中，具备训练数据集D以及假设空间H。通常情况下，均是采用给定训练数据集D的方式来确定出最佳假设空间H，基于概率的思想，寻找最大可能性的假设。对于某一个假设h，在没有使用训练数据D前存在先验概率$P(h)$，类似的$P(D)$可以表示没有某一种假设成立时D的概率，则根据条件概率的定义$P(D|h)$表示当假设h成立时可以观察训练数据集D的概率，而应该关注的是$P(h|D)$即在给定训练数据D后，假设为h的概率，称此概率为后验概率。根据贝叶斯法则，有如下计算方法：

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)} \quad (3)$$

1.2 朴素贝叶斯算法

在机器学习领域，基于贝叶斯准则衍生了很多的分类方法，其中最基本的形式是朴素贝叶斯算法，它是一种应用广泛，简单有效的基于概率分类的方法，是本邮件过滤系统中的主要算法。

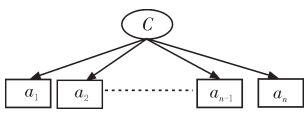

图1 算法思想

在朴素贝叶斯算法中，如图1所示，需要将训练实例表示成属性（特征）向量A和决策类别变量C。同时，在该算法中，假定每个特征之间相互独立对于决策变量它们具有独立作用。

此时，根据该假设，如果样本A表示的属性向量集合中的每一个属性ak都是相互独立的，那么$P(A|C_i)$可以分解为：

$$P(A|C_i)=\prod_{k=1}^{n}P(a_k|C_i) \quad (4)$$

则对于后验概率$P(C_i|A)$，表示特征A属于类别C_i的概率，根据贝叶斯准则，可以由下式计算：

$$P(C_i|A)=\frac{P(C_i)}{P(A)}\prod_{j=1}^{m}P(a_k|C_i) \quad (5)$$

其中，C_i的先验概率$P(C_i)$很容易求得：训练集中属于C_i的数量/训练集的总量。朴素贝叶斯算法在计算条件概率中常常采用多变量的贝努力模型与多项式模型。

2 系统实现

2.1 算法实现

在算法的实现上，对于垃圾邮件的过滤，可以将邮件分为垃圾邮件与非垃圾邮件（合法邮件），因此垃圾邮件的过滤系统可以简化为二值分类问题。具体垃圾邮件过滤算法训练流程如图2所示。

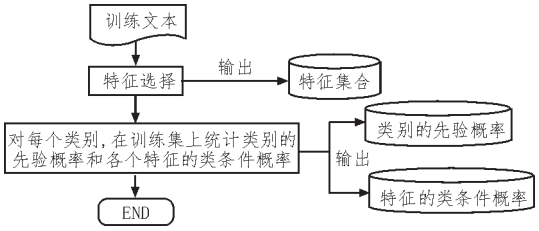

图2 垃圾邮件过滤算法训练流程

如图2所示，在进行垃圾邮件过滤算法训练时，首先，本文将采用RSS源进行垃圾邮件以及合法邮件的收集工作，以此建立其相应的邮件集合；其次，从邮件集合中对字符串进行提取，如sale，cash作为Token串，并对字顿进行统计；然后，建立连个哈希表hash_good，hash_bad分别映射Token中统计的垃圾邮件与合理邮件的字频，即在每个训练集上统计类别的先验概率和每个特征的类条件概率；最后，计算每个hashtable中的Token串的频率。

文中将上节中的事件C定义为接收到垃圾事件，在本节中记作A。将第i个Token串记作t_i，则

$P(A|t_i)$ 可以表示，当某一封邮件中，出现了 t_i 时，该邮件被判别为垃圾邮件的概率。t_i 出现在 hash_good 中的概率记作 $P_1(t_i)$，在 hash_bad 中记作 $P_2(t_i)$。则

$$P(A|t_i)=\frac{P_2(t_i)}{P_1(t_i)+P_2(t_i)} \quad (6)$$

对 $P(A|t_i)$ 进行平滑处理后建立第三个 hash 映射 t_i 到 $P(A|t_i)$，得到该 hash 后训练完成，根据该 hash 表进行邮件类别的判别。当收到邮件时，判别步骤如图 3 所示。

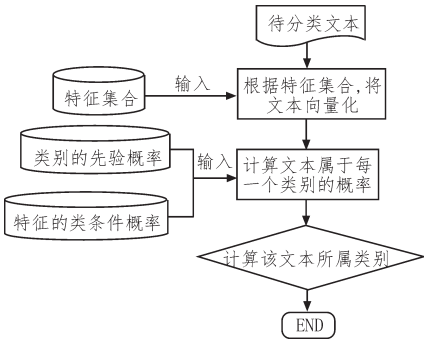

图3　垃圾邮件判别过程

对于新的邮件，首先需要提取词条生成该邮件的 token，然后查询训练模型过程中生成的第三个 hash 表得到 token 串对应的概率值。假设 N 个 token 在 hash 中对应的概率分别为 $P_1,P_2,...P_n$；以 $P(A|t_1,t_2,...,t_n)$ 代表需要判别的邮件是垃圾邮件的概率；则当该邮件中出现 token 串 $t_1,t_2,...,t_n$ 时，可用以下公式对该条件概率进行计算：

$$P(A|t_1,t_2,...,t_n)=\frac{P_1P_2...P_n}{P_1P_2P_n+(1-P_1)(1-P_n)...(1-P_n)} \quad (7)$$

若是当该概率超过给定阈值时，就可将该邮件判定为垃圾邮件。

2.2　系统设计与实现

文中在对该系统设计时，将在 windows 平台下采用 QL Server 2008 的后台数据库，使用 Python 语言设计实现。系统在设计时针对互联网上的垃圾邮件。结合了多种过滤方式分级过滤。系统的架构如图 4 所示。由图 4 可以看出，本系统分为管理与用户操作两个界面。在过滤的时候，可以选择 3 种不同的过滤方式，分别为按照邮件地址进行黑白名单判别进行过滤，根据邮件主题进行过滤和通过贝叶斯算法生成的垃圾邮件模板进行过滤。通过这种不同方式的分层过滤方法，可以大大提高系统对于垃圾邮件过滤的准确性。

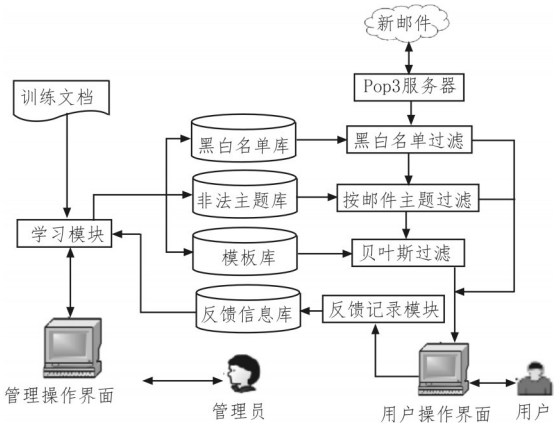

图4　系统架构

系统的核心模块是基于贝叶斯算法的过滤模块，该模块基于邮件内容进行过滤，其架构如图 5 所示。

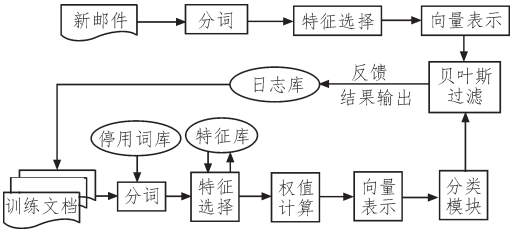

图5　过滤模块架构

该模块的构建基于 2.1 的算法实现过程，由图 5 可以看出，本模块包括分词、特征选择、向量表示等过程。对于特征选择过程，本模块在系统的训练与后续的使用分类过滤阶段都有使用，在训练的过程中，可采用对分词模块产生的关键词进行统计的方式计算出每个关键词的权重，经过排序后选择合适的作为特征。在过滤的过程中，还要对分词模块产生的无用关键词进行剔除。本模块的贝叶斯过滤过程按照 2.1 阐述的算法实行。过滤模块是一种增量式的学习思想，可以对上次的特征选择与过滤进行记录，以指导系统重新选择特征指标。

2.3　系统测试

在进行系统测试时，先定义一个查全率（Recall）用来表述系统发现垃圾邮件的能力

$$Recall=\frac{A}{N} \quad (8)$$

在公式（8）中，A 代表正确判定为垃圾邮件的数目，N 代表垃圾邮件的数目。在测试中，采用英文文件样本，样本来自公开的垃圾邮件资料库 PU1，样本库中共有 1 099 封邮件，其中 481 封是垃圾邮件，在系统的训练与测试中，将样本分为 10 份，每份 110 封邮件，选取一定的作为训练组，其余的作为测试组。为了更好地认证本系统的性能，本文引入分级

最小算法作为测试对比。分别采用两种算法进行训练测试，从测试结果可以看出，朴素贝叶斯算法的查全率大大高于分级最小算法。测试结果如图6所示。

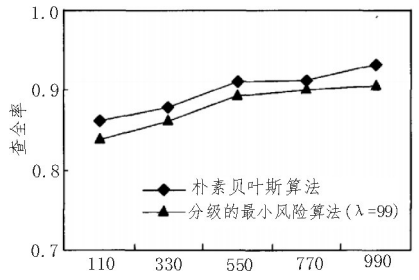

图6　算法测试对比

3　结束语

文中介绍了一种基于朴素贝叶斯算法实现的垃圾邮件的识别与过滤系统。通过朴素贝叶斯算法对用户收到的邮件的地址、主题与内容信息，进行邮件类别判定，帮助用户远离垃圾邮件干扰。经测试本识别系统有较高的识别正确率，可以广泛应用在未来的垃圾邮件过滤系统中。

参考文献：

[1] 曾纪川. 基于安卓的篇章级手写识别输入法[D]. 哈尔滨：哈尔滨工业大学，2013.

[2] 郑炜，沈文，张英鹏，等.基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究[J].西北工业大学学报，2010，28(4)：622-627.

[3] 马小龙.一种改进的贝叶斯算法在垃圾邮件过滤中的研究[J].计算机应用研究，2012，29(3)：1091-1094.

[4] 邹磊，卢炎生，崔得暄，等.一种基于最小损失的垃圾邮件屏蔽算法[J].华中科技大学学报：自然科学版，2005，33(z1)：352-355.

[5] 王忠建，张树舰，李颖，等.一种改进的基于贝叶斯的垃圾邮件过滤方法[J].黑龙江科技信息，2014，(21)：175.

[6] 张泽明，罗文坚，王煦法，等.一种基于人工免疫的多层垃圾邮件过滤算法[J].电子学报，2006，34(9)：1616-1620.

[7] 梁志文，杨金民，李元旗，等.基于多项式模型和低风险的贝叶斯垃圾邮件过滤算法[J].中南大学学报：自然科学版，2013，44(7)：2787-2792.

[8] 王潇杨，陈南飞，张登科，等.图片型垃圾邮件过滤分析系统设计和实现[J].大连理工大学学报，2011，51(z1)：69-72.

[9] 王晓龙，关毅. 计算机自然语言处理[M]. 北京：清华大学出版社，2005.

[10]张付志，伍朝辉，姚芳，等.基于贝叶斯算法的垃圾邮件过滤技术的研究与改进[J].燕山大学学报，2009，33(1)：47-52.

[11]曲美亭，于静潇，卜巍，等.基于语义引导贝叶斯算法的交互建筑设计模型[J].科技通报，2016，32(5)：133-136.

[12]赵文涛，孟令军，赵好好，等.分布式朴素贝叶斯算法在文本分类中的应用[J].测控技术，2016，35(6)：50-55.

[13]卢苇，彭雅.几种常用文本分类算法性能比较与分析[J].湖南大学学报：自然科学版，2007，34(6)：67-69.

[14]李国钊.智能垃圾邮件过滤系统的实现研究[J].信息与电脑，2016(11)：92-93.

[15]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.

[16]雷剑刚，孙细斌.一种智能垃圾邮件过滤模型的仿真研究[J].计算机仿真，2013，30(5)：370-373.

（上接第170页）

[10]魏涛.分布式计算机网络结构分析与优化探讨[J]. 信息系统工程，2016(10)：41.

[11]周巾裕，赵守盈，赵德轩.分布式计算在计算机自适应测验系统中的应用[J]. 黔南民族师范学院学报，2016，36(4)：107-110，115.

[12]赵婕. 基于SNMP协议的分布式计算机网络监控系统设计[J]. 自动化与仪器仪表，2016(6)：124-125.

[13]朱云生. 分布式技术与数据库应用于计算机技术领域解析[J]. 数字技术与应用，2016(5)：62.

[14]刘可可. 基于安全信息传输网的分布式计算机联锁系统设计[J]. 信息技术与信息化，2016(3)：54-56.

[15]潘培雯. 基于分布式计算机技术的数据库管理系统研究[J]. 电子技术与软件工程，2016(5)：184.

[16]潘培雯. 基于分布式计算机技术的数据库管理系统研究[J]. 电脑知识与技术，2015，11(6)：6-7.