新文献检测报告（简明版）

报告编号：PL-20190611-1A01CCA4-JM　　检测时间：2019-06-11 23:28:43

题　　名：基于朴素贝叶斯的垃圾邮件过滤
作　　者：罗鹏娟
检测范围：
☑中国学术期刊数据库　☑中国学位论文全文数据库　☑中国学术会议论文数据库
☑中国学术网页数据库　☑中国专利文献数据库　☑中国优秀报纸数据库

检测结果

总相似比：6.08%

检测字数：6676
参考文献相似比：0.00%　/　排除参考文献相似比：6.08%
可能引用本人已发表论文相似比：0.00%　/　辅助排除本人已发表论文相似比：6.08%
可能引用本人学位论文相似比：0.00%　/　辅助排除本人学位论文相似比：6.08%
单篇论文最大相似比：1.71%

相似片段分布图

注：绿色区域为参考文献相似部分，蓝色区域为本人已发表论文相似部分，黄色区域为本人学位论文相似部分，红色区域为其他文献相似部分

相似文献列表

序号	相似比	题名	作者	文献类型	来源	发表时间	是否引用
1	1.71%	朴素贝叶斯算法在垃圾邮件过滤方面的应用	徐梦龙 等	期刊论文	《网络安全技术与应用》	2018-07-28	否
2	0.84%	基于适应概念漂移的垃圾邮件过滤系统设计与实现	党建军	学位论文	电子科技大学	2010-05-24	否
3	0.76%	贝叶斯文本分类器的研究与改进	史瑞芳	期刊论文	《计算机工程与应用》	2008-09-25	否
4	0.63%	Hadoop平台垃圾邮件过滤算法研究与实现	种飞	学位论文	沈阳理工大学	2018-03-07	否
5	0.52%	改进的贝叶斯分类法在垃圾邮件过滤中的应用研究	钱诚	学位论文	华东理工大学	2012-06-01	否
6	0.40%	基于ISOMAP算法的贝叶斯分类模型及应用	许义仿	学位论文	华北水利水电大学	2018-05-01	否
7	0.25%	DHTnil垃圾邮件过滤系统的改进方法研究	张晓明	学位论文	南开大学	2011-05-01	否
8	0.22%	基于Spark的分布式NB算法的垃圾邮件过滤研究	张亚斌	学位论文	内蒙古科技大学	2018-06-09	否
9	0.21%	基于粗糙集的最小风险贝叶斯垃圾邮件过滤算法	郝建忠 等	会议论文	第六届中国信息和通信安全学术会议(CCICS'2009)	2009-05-30	否
10	0.21%	基于内容过滤的反垃圾邮件技术研究	王平	学位论文	北京邮电大学	2006-02-28	否
11	0.16%	文本挖掘在垃圾邮件过滤中的应用研究	李春玲	学位论文	中国人民大学	2008-06-01	否
12	0.15%	进化计算在反垃圾邮件系统中的应用研究	李姝亚	学位论文	电子科技大学	2008-05-01	否

原文

1 绪论

1.1垃圾邮件的危害

信息技术的发展使电子邮件成为互联网信息传播和通信的主要方式，具有快速，方便，低成本的优点。虽然电子邮件有利于生活，但垃圾邮件的泛滥也给电子邮件的用户带来了许许多多问题和麻烦。大量的非正常邮件给人们的生活带来的很大的问题，这可能导致邮箱的用户变少，因此让很多研究人员的对这个问题进行了深入的研究。根据中国的反垃圾邮件的第一次调查报告显示，在2007年下旬至2008年上旬内，垃圾邮件的比例由60.97%上升到了63.79%在中国邮箱用户收到邮件中的，上升了2.82 %，大大高于之前所调查的结果。非正常邮件的问题，对邮箱用户的工作，生活，有不可挽回的损失。

严重的垃圾邮件问题主要有以下两个原因：

垃圾邮件中广告居多，发送者只用获得邮箱用户的列表，就可以把垃圾邮件发送到用户的邮箱中。

目前没有很有效的限制垃圾邮件的技术使，这使得电子邮箱的用户经常受到垃圾邮件的烦扰。所以，有效的垃圾邮件过滤技术来解决垃圾邮件给用户带来的困扰是非常必要的。

1.2 过滤垃圾邮件的技术现状

垃圾邮件问题越来越严重，人们开始从不同的角度寻找解决方案，目前已经在世界多个地方建立了各种组织来缓解垃圾邮件所带来的困扰。在与垃圾邮件的持续斗争中，一些技术不断改进，新的垃圾邮件的特征不断被开发出来。在垃圾邮件的过滤方面，对此目前主要有以下几种解决办法：使用IP地址对垃圾邮件进行过滤，读取邮件的内容对垃圾邮件进行过滤及判断发送邮件者的行为对垃圾邮件进行过滤。

使用IP地址对垃圾邮件进行过滤的方法：①黑白名单 ②安全认证法③监测邮箱服务器端

读取邮件的内容对垃圾邮件进行过滤包含有使用规则的内容进行的过滤和使用概率统计的内容进行的过滤：

①使用规则的内容：

a.对关键字进行匹配：此技术对邮件的主题进行监测，匹配邮件的主题与垃圾邮件的数据库的常用词汇。如果得到高于阈值的结果，则认定此邮件为垃圾邮件，并将这封邮件过滤掉，以便不能将其发送给电子邮件的用户。如果得到低于阈值的结果，则确定为普通邮件并且可以正常发送。利用关键字进行判断的优点是该方法相对容易判断，判断的速度也较快，但缺点是误报率较高，给电子邮件用户带来损失。

b.采用RoughSet:RoughSet是研究不确定、不完整的知识的重要的方法。RoughSet的核心是对多个值的属性的向量的集合进行考虑。它在集合的顶部构建等价的关系，然后确定分类的最小的对象集和最大的对象集。在粗糙集理论中，决策表也叫做为信息系统。属性分为要求属性和决定属性。利用粗糙集对垃圾邮件进行过滤，据实验测试可以达到80%的正确率。

c.采用决策树：数据进行分类的一种重要方法就是决策树对。该过程采用树的形式。 形成一个“问题 - 判断 - 问题”的树。决策树的节点具有分支节点和叶节点。将树转到另一个分支，最后转到没有分支的叶节点，叶节点指示相应的类别。过滤垃圾邮件时采用决策树可以达到非常好的过滤结果，但是不能直接采用决策树方法去过滤垃圾邮件，他需要使用其他一些辅助方法。

②使用概率统计的内容进行的过滤的方法 ：

根据概率统计过滤内容是指将文本计算为垃圾邮件以过滤垃圾邮件的概率。朴素贝叶斯算法用作过滤垃圾邮件是目前最好也是最常用的。贝叶斯分类器的原理是以令牌为单位算出文本内容是属于垃圾邮件还是属于普通邮件的概率，对邮件进行判别并分类。

（3）判断发送邮件者的行为对垃圾邮件进行过滤的方法：

①对发送邮件的工具进行识别:基本原则是根据邮件头中显示的发送工具或发件人的字段信息确定批量发送工具是否发送邮件。如果发现发送的邮件是用批量发送的工具发送的，就将其判断为垃圾邮件并被阻止，使其不能发送到邮箱用户的邮箱中，如果发现无异常，则将其放行。此方法的优势是它可以直接匹配出邮件头中的某些词语是否属于常见的垃圾邮件中的词语但它的缺点也是非常大的，它需要不断地去更新批量发送软件工具和工具的功能;同时，一些普通的电子邮件如果也通过群发发送，如公司给员工群发送信息或公告等,也会拦截，如此就会给邮箱用户带来不便。

②利用流量去控制垃圾邮件的发送:基本方法是根据电子邮箱的管理者去监控电子邮件。根据邮件的内容，电子邮箱的管理者给发送垃圾邮件者进行警告或拒绝它使用电子邮箱。该技术需要对现有邮件网络进行大量更改，效果显而易见,但也是存在误判和难以忍受的危害。

1.3 研究的目的和意义

目前，垃圾邮件泛滥，给人们的生活带来了诸多麻烦，基于目前过滤垃圾邮件的方法，本文采用了朴素贝叶斯垃圾邮件过

滤算法，它的正确率是最高的。朴素贝叶斯算法过滤垃圾邮件可以在样本数据较少的情况下仍然有效，也可以有效的处理多类别问题。

2系统简介

2.1贝叶斯原理

垃圾邮件的过滤算法中贝叶斯理论被大量的应用，具体是把它看作是一个分类问题，第一是我们要收集大量的邮件作为邮件样本，第二我们要对收集到的邮件样本进行训练，第三利用训练好的贝叶斯分类器对其他的邮件进行分类。通过训练样本邮件，贝叶斯分类器可以获取所有垃圾邮件的所有特征，并根据垃圾邮件的特征来计算邮件文本属于垃圾邮件或正常邮件的概率，将想要分类的邮件利用邮件特征计算它的概率，基于它的概率把它分到最大的类别中去,来比较精确地概率对实现垃圾邮件的过滤。

2.2贝叶斯技术

随机试一个验 E的可能出现的情况组成的结果是 E 的样本空间将其标记为 S。样本空间中E 的每个结果，即在样本空间中的所有元素，称为一个样本点。随机试验 E 的样本空间S的子集都是 E 的一个随机事件。表示在事件E发生情况下，某一个的结果A 可能的发生的概率的值。假定 A，B 都是随机的事件，且 事件A发生的概率大于零，即有：

公式（2-1）

在上式中在事件 A 发生的条件下事件 B 发生的条件概率为P（B|A）。

假定事件A发生的概率大于零，则有

公式（2-2）

上式是非常著名乘法公式，意为两个随机事件的乘积可能出现的概率。

假定随机事件 E 的样本空间是 S，并且A是事件E的子事件，Bl，B2 ，…,Bn 为 样本空间S 的一个划分，且 P(Bi)>0(i=1,2,…,n),则有

公式（2-3）

将公式（2-3）称为全概率公式。

定理 2.3：试验 E 的样本空间为 S，A 为 试验E的事件,且 P(Ai)>0 ， P(Bi)>0(i=1,2,…,n)， 则有

公式（2-4）

上述公式是著名的贝叶斯（Bayes）公式。在贝叶斯理论中，依据已有资料来确定的事件发生的概率我们称之为先验概率，其中，概率值的大小取决于已有资料，它是指计算前的概率。

基于贝叶斯公式及先验概率对将要判定的事件做出的估计我们称之为后验概率。要进行后验概率估计的前提是必须要计算先验概率并且其是有效的且充分的。

通过已有资料和计算方法得到的先验概率是贝叶斯概率，并运用事件发生的可能性的方法对后验概率进行的猜测。

2.3朴素贝叶斯算法

目前，在关于垃圾邮件内容过滤的算法中，我们使用最多，最为广泛的算法是基于朴素贝叶斯的文本分类算法。其中，朴素贝叶斯文本分类算法是通过计算目标属于某个类别的概率，最后概率最大的一类即是文本的类别。在“贝叶斯假设”的基础上引入了朴素贝叶斯的分类的算法。朴素贝叶斯的分类的算法是假设实验的目标所有的特征都互不影响，朴素贝叶斯的分类的算法的基础上用来简化概率的计算。

贝叶斯的文本的分类的算法的基本原理如公式2-5所示：

公式（2-5）

公式2-6为全概率公式：

公式（2-6）

由训练集进行计算 类的先验概率如公式2-7所示：

公式（2-7）

由文本中出现的特征的条件概率可计算公式2-7条件概率。

基于朴素贝叶斯的文本的分类的算法有两种基本的模型：A.多项式事件模型。B.多变量贝努里事件模型。

（1）多项式事件模型和多变量贝努里事件模型的区别主要在于：在公式（2-5）中以不一样的计算方法来计算P(dx| cj)。在多变量贝努里事件模型中，样本内容中的分词的每一个分词都有权重。假设有n个分词，把样本内容的分词看做是一个事件，文本是通过分词结果产生的，即目标词是否出现。

如特征wi 在文本dx 中的出现情况为Bxt = 1 / 0 为，Bxt表示是否出现则有公式2-8:

公式（2-8）

在公式2-8中文本属于 Cj 类的情况下 Wt特征的可能性为P(Wt|Cj)。由公式2-8可以得到，在多变量的贝努里事件的模型中，所有单个分词的概率之积就是样本邮件的判断依据。如果文本之中出现目标分词，则最后结果为：P(Wt| Cj) ，如果不出现，乘的项是1P(Wt| Cj)。 P(Wt| Cj) 的计算方式为公式2-9所示：

公式（2-9）

其中平滑的处理过程如公式2-10所示：

公式（2-10）

由以上可得多变量贝努里事件模型有以下两个的特征：

① 不需分词在样本中的封数，在计算 P(dx| cj) 和 P(wt| cj)时；

②没有出现在目标文本中的词，计算 P(dx| cj) 时乘以项1P(wt| cj)。 在进行判别邮件内容时，只用虑垃圾SPAM和非垃圾邮件HAM这两种情况，设c=0 表示正差邮件，c=1 表示非正常邮件，则可得到公式2-11结果：

公式（2-11）

由以上的公式（2-9）和公式（2-8）可可有如下结果：

公式（2-12）

公式（2-13）

公式（2-14）

对样本文件进行计算 P(c =1)、P(wt|c=1)和 P(wt| c =0) ，判别时由以上公式计算即可知道目标文本分类正确与否。

（2）在多项式事件模型中，对评判结果造成影响的一个重要原因是样本内容的分词在邮件中出现的封数。如果样本中的分词是彼此不重复的。设为样本中包含的分词数，由分词在样本中的出现封数的和，文本中分词的数量为n，可得 适合该模型，可得公式2-15结果：

公式（2-15）

计算训练集时如公式2-16所示：

公式（2-16）

平滑处理如公式2-17所示：

公式（2-17）

在对邮件进行分类时，只需判断邮件是正常的还是非正常的，假定c=0 表示正常邮件， c=1 表示非正常邮件，由公式(2-15)和公式(2-16)可以得到公式2-18的结果：

公式（2-18）

对同一封邮件，由公式（2-18）得：

公式（2-19）

朴素贝叶斯文本分类器是对文本的内容进行判断并计算它出现的可能性，并计算它出现的可能性，将所得概率与设定的阈值相比较，判别目标内容是有用的还是无用的。阈值是概率计算中非常重要的一个参数。垃圾邮件与样本的相像程度阈值就大一些。将一些与垃圾邮件相似程度低的邮件判定为正常邮件，这种情况下能减少判断的情况，但是，垃圾邮件与样本邮件的相似度低，阈值就会小一些，表示将与垃圾邮件相像程度小的垃圾邮件正常发送，但是这种情况下可能把正常邮件误判为垃圾邮件的概率会变大，一旦判断错误那么就会对邮箱用户带来很多的损失。因此可以通过不断地训练文本来调整阈值的大小。

2.4训练算法

2.4.1 计算先验概率

使用贝叶斯算法的本质是由概率来判断一件事出现的概率的一种方式，使用出现可能性大的当作分类的依据，这是贝叶斯算法的中心的内容，其中最能代表它的是对事件可能发生的可信度来进行分析。置信度是事件的概率的一种：先验概率是人们由自己了解的知识对事件可能发生概率一种评估。

2.4.2 计算每个词语各类别邮件中的概率

在垃圾邮件的条件下第i个特征的概率，首先先将所有的类别为1的词向量相加，可以得到每个特征的个数，所以再除以在类别为1的单词总数得到的就是在垃圾邮件中每个单词的概率了。

2.5算法修正

在进行分类时，多个概率乘积得到类别，但是如果有一个概率为0,则最后的结果为0，因此未来避免未出现的属性值，在估计概率时同必须要进行“平滑”处理， 常用的是平滑处理是使用“拉普拉斯修正”（Laplacian correction）,具体来说，在计算的时候，将分子加1，分母加上类别数N.同样在计算)的时候在分子加1，分母加2，这样就避免了整个算式为 0 的特殊情况。

当值过小时可能会出现相乘溢出，在实际中对概率取对数的形式，再进行相乘，可以防止相乘结果溢出的现象。

3系统设计与实现

3.1 流程图

图3-1 垃圾邮件过滤算法基本思想

3.2样本数据准备

使用公开数据，公开数据总共包含垃圾邮件为250封和正常邮件250封，将其作为训练的样本。

3.3分词

分词是把邮件内容划分成一个个单词的形式，可以想到用正则表达式，但正则表达式不能很好的分割中文，所以这里选用的是python的第三方库jieba来实现邮件内容分割成单个词语。代码示例如图4-1

图4-1 分词示例

分词结果如图4-2 分词结果所示

图4-2 分词结果

3.4生成词汇表

将所有的邮件进行分词后生成一个dataSet，然后生成一个词汇表，这个词汇表是一个集合，即每个单词只出现一次，词汇表是一个列表形式如：[“cute”,” love”,” help”,garbage”,” quit”…]

3.5生成词向量

每一封邮件的词汇都存在了词汇表中，因此可以将每一封邮件生成一个词向量，存在几个则为几，不存在为0，例如：[“love”,” garbage”],则他的词向量为[0,1,0,1,0,…],其位置是与词汇表所对应的，因此词向量的维度与词汇表相同。

3.6词云

词云最早是来自 里奇·戈登，他曾是美国西北大学的新闻学副教授、新媒体专业主任。词云就是对文本中出现次数较多的词语给予以在视觉上的突出的显示，形成一种类似于"关键词的渲染"，从而让阅读者快速滤掉大量的无用的信息的效果，让浏览的人只需粗略扫过词云就可以领略文本的主要意思。在文本分类结束后，为了防止把正常的邮件错误分类为了垃圾邮件，在分类结果为垃圾邮件时，通过突出展现邮件关键词，以便避免此问题。

图3-2 词云图示例

4测试环境

4.1 python 3.7环境

Python是解释型的脚本语言，可以应用在很多领域：Web 开发、科学计算、统计、教育、桌面界面开发、软件开发、后端开发等。使用Python做科学计算的优点有很多：

完全免费。

许多科学计算都提供python接口，可以很方便的调用。

相较其他软件来说Python是一种非常容易学习和严谨的计算机的语言，其代码非常易读并易于维护。

非常易于操作文本文件；使用非常广泛，存在大量的开发文档。

有着非常丰富的扩展库，可以非常容易用来完成高级的任务，程序的各种功能会和容易实现。

4.2 pycharm 开发工具

PyCharm是一个非常好用的编译工具对于Python来说，有很多帮助对于使用者来说，使用者在使用Python语言进行程序编写时，可大大的增加开发效率，比如在程序中间进行调试时，运行界面如图5-1所示。

图5-1pycharm 运行界面图

5实验结果及分析

5.1处理数据验证过程

这里首先将500封邮件读进列表中，然后生成一个词汇表包含所有的单词，接下来使用交叉验证，随机的从样本中选择150个样本进行测试，350个进行训练。

第三阶段应用阶段调用第三方库对邮件内容进行分词，去重，再计算单个词语为垃圾邮件词语的概率及各个词语的联合概率，进行判别是否为垃圾邮件。

图5-2 运行结果图

经过10次重复交叉验证，最后的正确分类邮件个数1445，错误分类邮件个数5个，其中把正确邮件分类为错误邮件个数为2，错误邮件分为正确邮件个数为3个，由此可知由朴素贝叶斯算法过滤垃圾邮件的正确率在到99.3%以上。由此分析此系统对垃圾邮件过滤有效。

结语　　对垃圾邮件进行过滤是全世界一直研究的，因为垃圾邮件的大量传播会给 整个互联网造成很大的压力。大量的垃圾邮件会造成的邮箱的服务器的拥堵，并且会极大的占用了邮箱使用者的内存，会占用邮箱用户的时间和精力也可能会对邮箱用户带来金钱上的损失；甚至还有的邮件还会盗用别人的邮箱的地址做为发信的地址，这将非常大地对其他人的信誉造成损坏，这些结果一旦发生最终使得邮箱用户流失。

现代很多人在研究过滤垃圾邮件的技术，也很很多有用的方法，但是垃圾邮件也是依据更新的过滤方法不断地修改垃圾邮件的最新特征,这也使得传统的垃圾邮件过滤的技术无法被及时的发现和检测出来。所以，如果要把垃圾邮件完全阻挡在系统外部，光靠过滤垃圾邮件的手段不能完全解决的，还必须需要有关部门的重视和积极参与，可通过宣传或者立法等办法，通过使用法律的手段来对实现垃圾邮件的制造者进行相关的制裁。所以只有我们每个-个人都自觉行动，加入到抵制垃圾邮件中，并且使用先进的技术武装网络，用法律的规管理制度为从根本上消除垃圾邮件。

本系统使用了基于朴素贝叶斯的方法来过滤邮件，虽然能够在比较大的程度上将垃圾邮件过滤掉，但不能很好的应对变化的邮件特征，希望有关部门能够积极的行动起来，积极加入垃圾邮件过滤，共同维护网络安全中来，从根本上杜绝垃圾邮件。

说明：

1.送检文献总字数=送检文献的总字符数，包含汉字、非中文字符、标点符号、阿拉伯数字（不计入空格）

2.总相似比=送检论文与检测范围全部数据相似部分的字数/检测总字符数

3.参考文献相似比=送检论文与其参考文献相似部分的字数/检测总字符数

4.辅助排除参考文献相似比=总相似比-参考文献相似比

5.“单篇文献最大相似比”：送检文献与某一文献的相似比高于全部其他文献

6.“是否引用”：某一相似文献是否被送检文献列为其参考文献