[综述]

文章编号: 1003－0077(2005)05－0001－10

基于内容的垃圾邮件过滤技术综述①

王 斌, 潘文锋

(中国科学院 计算技术研究所, 北京 100080)

摘要: 垃圾邮件问题日益严重, 受到研究人员的广泛关注。基于内容的过滤是当前解决垃圾邮件问题的主流技术之一。目前基于内容的垃圾邮件过滤主要包括基于规则的方法和基于概率统计的方法。本文综述了目前用于垃圾邮件过滤研究的各种语料和评价方法, 并总结了目前使用的垃圾邮件过滤技术以及它们之间的对比实验, 包括Ripper、决策树、Rough Set、Rocchio、Boosting、Bayes、kNN、SVM、Winnow等等。实验结果表明, Boosting、Flexible Bayes、SVM、Winnow方法是目前较好的垃圾邮件过滤方法, 它们在评测语料上的结果已经达到很高水平, 但是, 要走向真正实用化, 还有很多的工作要做。

关键词: 计算机应用; 中文信息处理; 综述; 垃圾邮件; 反垃圾邮件; 信息过滤; 文本分类

中图分类号: TP391 文献标识码: A

A Survey of Content-based Anti-spam Email Filtering

WANG Bin, PAN Wen-feng

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: The volume of junk emails on the Internet has grown tremendously in the past few years and is causing serious problems. Content-based filtering is one of the mainstream technologies used so far. This paper aims to provide an overview on the state of art in this research field, including benchmark corpora, evaluation methods and filtering approaches. Many filtering approaches, including Ripper, Decision Trees, Rough Sets, Rocchio, Boosting, Bayes, kNN, SVM and Winnow, are discussed and compared in this paper. The experimental results show that some approaches, such as Boosting, Flexible Bayes, SVM, Winnow, can achieve very good results on research corpora. However, much more work should be done for practical use.

Key words: computer application; Chinese information processing; overview; junk email; anti-spam; information filtering; text classification

1 引言

作为互联网的第一大应用, 电子邮件一直受到广大网民的青睐。但是近些年来, 垃圾邮件问题日益严重。2004年1月, 中国互联网络信息中心(CNNIC)发布的《第十三次中国互联网发展状况统计报告》显示, 中国网民平均每周收到13.7封电子邮件, 其中垃圾邮件占了7.9封, 垃圾邮件数量已经超过了正常邮件数量。垃圾邮件不仅耗费网络带宽和计算机时空开销, 而

① 收稿日期: 2004－09－02 定稿日期: 2005－03－10
基金项目: 国家973项目资助(2004CB318109)
作者简介: 王斌(1972—), 男, 博士, 副研究员, 主要研究方向: 信息检索、自然语言处理、内容安全.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

且会对企业的正常运作和用户的正常工作造成严重的干扰。

迄今为止,垃圾邮件在国际上并没有一个标准的定义。垃圾邮件的基本特征是“不请自来”,而且大部分垃圾邮件都带有商业或者其他宣传目的。同时,垃圾邮件的判定和邮件的接收者有很大关系,不同用户对同一邮件的判断结果可能会存在差异。

要解决垃圾邮件问题,必须综合法律、技术等手段。反垃圾邮件技术上可以分成两类:“根源阻断”和“存在发现”。“根源阻断”是指通过防止垃圾邮件的产生来减少垃圾邮件。据报道,比尔盖茨曾提出通过对发送邮件收费来减少垃圾邮件。另一种称为 Zomail 的技术试图通过给每个用户分配密码锁的方法来保护邮件接收者的安全。这些方法目前还没得走向实用,即使走向实用也需要对全球的邮件系统进行全面改造。所以,对垃圾邮件进行阻断还有很长的路要走。目前,主流的反垃圾邮件技术是“存在发现”,即对已经产生的垃圾邮件进行过滤。反垃圾邮件的发现可以通过邮件的内容特征或者其他特征(如群发特征)来实现,其中基于内容的反垃圾邮件过滤技术是研究的重点。

从内容上看,垃圾邮件过滤可以看成一个“二类”问题:垃圾邮件类和合法邮件类。因此,各种分类方法可以用于垃圾邮件的过滤。然而,垃圾邮件过滤是一个特定领域的分类问题,它至少在以下几个方面与一般的分类存在不同:

一、通常认为,用户宁愿接收更多的垃圾邮件,也不能接受将合法邮件错判成垃圾邮件。因此,与通常的分类方法相比,垃圾邮件过滤更重视正确率;

二、垃圾邮件过滤实现的环境通常都有较高的性能要求,因此,要求垃圾邮件过滤的方法不仅要重视实现的效果,也要重视实现的效率;

三、垃圾邮件过滤中的类别有别于通常分类中的类别,一方面,垃圾邮件、合法邮件在语义上并不象通常分类中的类别(如体育、军事等等)能够被人理解;另一方面垃圾邮件的类别定义可能会因人而异,也可能会随着时间而改变。

本文后续内容组织如下:第二节和第三节分别介绍目前垃圾邮件研究中用到的公共语料库和评价方法;第四节总结了目前用到的基于内容的垃圾邮件判别方法;最后,第五节给出了总结和展望。

2 垃圾邮件语料库

为了能对不同的垃圾邮件过滤算法进行比较,必须有一个可供比较的平台,包括公共的评测语料和统一的评价方法。本节主要介绍目前该项研究工作中用到的公共语料。这些语料都可以免费下载。目前,所见到的大都是英文语料,还没见到公开的中文垃圾邮件语料库。

2.1 PU 系列语料

PU 系列语料[①] 由希腊学者 Androutsopoulos 提供,其来源于提供者收到的某个时段的真实邮件。语料只保留了这些邮件的标题和正文中的纯文本内容。为了保护提供者的隐私,语料中将邮件中的不同词汇用不同整数代替。PU 系列语料目前包括 PU1、PU2、PU3 和 PUA 四个语料。每个 PU 语料平均分成 10 份,可以每次取其中的 9 份作为训练集,另外 1 份作为测试集进行交叉验证(cross validation)。PU 系列语料的详细情况可参见表 1。PU 系列语料的不足在于内容加密后不能试验语言学相关的方法。

① 可从http://iit.demokritos.gr/skel/i-config/downloads/下载。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

2.2 Ling-Spam 语料

Ling-Spam[①] 也由 Androutsopoulos 等人提供, 由提供者收到的垃圾邮件和来自于语言学家邮件列表(Linguist list)的合法邮件组成。Ling-Spam 中公用的合法邮件没有加密。Ling-Spam 和 PU1 使用相同的垃圾邮件样本。

Ling-Spam 语料的不足在于它的合法邮件都来自于一个特定的邮件列表, 因此这些合法邮件都偏向于一个主题, 而对某个特定用户来说, 真正收到的合法邮件内容的主题一般都很发散。从实验上看来, 这种内容专指性可以导致比较好的过滤结果。如果我们关注某个邮件组或者邮件列表上的垃圾邮件过滤, 可以采用在 Ling-Spam 语料上表现较好的算法。

2.3 其他语料

除了上述两种类型的语料外, 目前见到的公开垃圾邮件语料还包括 Spam Assassin 语料[②] 和 Spambase 语料[③] 等等, 前者是 Network Associates 的 Justin Mason 提供的。与 Ling-Spam 有些类似, 其合法邮件来自公众论坛。Spambase 语料由 Hopkins、Reeber 等人提供。与前面提到的那些语料不同, Spambase 将每一封邮件都表示为向量的形式, 每个向量都是 57 维(预先选择出来的 57 个词特征), 权重一般是词频。Spambase 来自于提供者的私人邮件。

表 1 垃圾邮件语料情况(数量单位: 封)

语料名称	非垃圾邮件数量	垃圾邮件数量	总数量	备 注
PU1	618 Pr	481 Pr	1099	加密, 多种形式
PU2	579 Pr	142 Pr	721	加密
PU3	2313 Pr	1826 Pr	4139	加密
PUA	571 Pr	571 Pr	1142	加密
LingSpam	2412 Pu	481 Pr	2893	未加密
Spam Assassin	4150 Pu	1897 Pr	6047	未加密
Spambase	2788 Pr	1813 Pr	4601	向量

Pr —— 来自私人邮件, Pu —— 来自公众论坛

3 评价体系

垃圾邮件过滤的性能评价通常借用文本分类的相关指标。具体地, 假设待测试的邮件集合中共有 N 封邮件, 一个垃圾邮件过滤系统的判定结果如表 2 所示:

表 2 垃圾邮件系统判定情况分布(单位: 封)

	实际为垃圾邮件	实际为合法邮件
系统判定为垃圾邮件	A	B
系统判定为合法邮件	C	D

其中, $N=A+B+C+D=N_s+N_l$。$N_s=A+C$ 为实际的垃圾邮件数目, $N_l=B+D$ 为实际的合法邮件数目。则可定义如下几个评价指标来衡量不同垃圾邮件过滤系统的性能:

(1) 召回率(Recall): $R=\frac{A}{A+C}=\frac{A}{N_s}$, 即垃圾邮件检出率。这个指标反映了过滤系统发现垃圾邮件的能力, 召回率越高, "漏网"的垃圾邮件就越少。

(2) 正确率(Precision): $R=\frac{A}{A+B}$, 即垃圾邮件检对率。正确率反应了过滤系统"找对"垃

① 可从 http://iit.demokritos.gr/skel/i-config/downloads/ 下载。

② 可从 http://www.spamassassin.org 获得。

③ 可从 http://www.ics.uci.edu/~mlearn/MLRepository.html 获得。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

圾邮件的能力, 正确率越大, 将合法邮件误判为垃圾邮件的可能性越小。

(3) 精确率(Accuracy): $Accur=\frac{A+D}{N}$, 即对所有邮件(包括垃圾邮件和合法邮件)的判对率。

(4) 错误率(Error rate): $Err=\frac{B+C}{N}=1-Accur$, 即对所有邮件(包括垃圾邮件和合法邮件)的判错率。

(5)F 值: $F=\frac{2PR}{R+P}$, F 实际上是召回率和正确率的调和平均, 它将召回率和正确率综合成一个指标。

除此之外, 垃圾邮件过滤中还常常采用虚报率(Fallout)、漏报率(Miss rate)等指标。

另外, 我们在前面提到, 在实际的垃圾邮件过滤中, 人们往往不希望将合法邮件误判成垃圾邮件。为了表示不同情况下垃圾邮件系统的代价。Androutsopoulos[8] 等人提出了代价因子的概念。假设将合法邮件误判为垃圾邮件的损失为是垃圾邮件判为合法邮件的 λ 倍(如: λ=9 表示一封合法邮件误判的损失是一封垃圾邮件误判的 9 倍)。则可以定义:

$$WErr=\frac{\lambda B+C}{\lambda N_l+N_s},\ WErr_b=\frac{N_s}{\lambda N_l+N_s}$$

代价因子 $TCR=\frac{WErr_b}{WErr}=\frac{Ns/(\lambda N_l+N_s)}{(\lambda B+C)/(\lambda N_l+N_s)}=\frac{N_s}{\lambda B+C}$

TCR 越高, 表明当前垃圾邮件过滤系统的损失越低。

另外, 在评价一个真实垃圾邮件过滤系统时, 还要考虑算法实现的时空效率。

4 基于内容的垃圾邮件过滤方法

基于内容对垃圾邮件进行判别, 目前一些垃圾邮件过滤工具中常采用黑名单—白名单或者手工制订规则的方法。黑名单—白名单可以看成手工制订规则方法的特例。黑名单给出了发送垃圾邮件的邮件地址(或者是 IP 地址范围、域名等属性)列表, 凡是属于黑名单的邮件被判定为垃圾邮件。白名单收录了邮件接收者确信的邮件地址信息, 凡是属于白名单的邮件都被判定为合法邮件。手工建立规则的方法通过用户建立一系列规则来判定垃圾邮件。显然, 这些方法的主观性会造成大量合法邮件的误判和垃圾邮件的漏判。因此, 目前的垃圾邮件工具逐渐倾向于引入基于内容的机器学习判别方法。

目前基于内容垃圾邮件判别的机器学习方法可以大体分成基于规则的方法和基于概率统计的方法。前者常常得出人们可以理解的显式规则; 后者往往通过某种计算表达式推出结果。本质上, 概率统计方法可以看成规则方法的一种特例, 只不过概率统计方法中得到的规则是一种不被人轻易理解的"隐式规则"。

不管是基于规则的方法还是基于概率统计的方法, 在使用时都经历从训练到过滤的过程。通过已有的训练集合(正例+反例)训练出相应的垃圾邮件规则(包括显式规则或隐式规则), 然后将规则应用到新的邮件判定中去。在实际系统中可能还会加入人机交互过程, 通过用户对判定结果的认可与否对已有的垃圾邮件规则进行更新。

4.1 基于规则的方法

基于规则的方法通过训练得到显式规则(通常用产生式表示, 如: if 邮件包含 sare money Then 该邮件为垃圾邮件)。规则方法学习的过程实际上是归纳总结的过程, 通过考查一个个的训练样本, 归纳总结出其中规律性的东西来形成规则。规则方法的主要优点是可以生成人

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

类理解的规则。缺点是在规律性不明显的应用领域效果较差。

4.1.1 Ripper

Ripper 是 William W. Cohen[17] 提出的一种基于规则的方法。它比传统的规则方法速度更快、性能更高。Cohen[18] 的普通文本分类的实验表明, Ripper 方法的正确率和决策树方法 C4.5 相差不大, 但是速度却提高了两个数量级。Drucker[5] 将 Ripper 方法用于垃圾邮件过滤, 在 1000 个文本特征的情况下, 通过从正例中学习规则并对规则进行修剪来获取垃圾邮件的覆盖规则, 在某单位员工真实邮件语料库(含 850 篇垃圾邮件和 2150 篇非垃圾邮件)上取得了 80%以上的精确率。

4.1.2 决策树(Decision Tree)方法

决策树是著名的规则方法之一。通过按照某种属性的顺序自顶向下地生成一棵树, 树的每个节点是属性名, 而每条边是属性值。从树根到树叶的一条路径便对应一条规则。基于信息增益进行属性顺序选择是决策树中常用的方法之一。著名的决策树算法有 ID3、C4.5 等。Carreras[19] 使用决策树来过滤垃圾邮件, 他采用 RIM 距离方法而非信息增益来选择特征, 采用 TFIDF 来描述特征, 在 PU1 语料上得到的垃圾邮件过滤的正确率和召回率都在 88%左右。目前, 由于决策树方法效果一般, 它本身并不常常直接用于垃圾邮件过滤, 而是作为 Boosting 方法的弱学习器来使用[13, 19, 20]。

4.1.3 Boosting 方法

严格地说, Boosting 方法不是一种特定的学习方法, 而是一种在已有学习方法基础上的进行"投票"的技术。它通过对已有的分类器(称为弱规则或弱假设)进行加权求和得到最终的分类器(称为强规则或强假设)。虽然从理论上来说, 任何机器学习方法都可以作为 Boosting 方法的弱学习器, 但在实际中, Boosting 的弱规则常常采用基于规则的方法, 因此我们将它归于规则方法类。Boosting 通过关注弱规则的错误而逐渐组合成强规则, 它是一种错误驱动的方法。AdaBoost 是 Boosting 方法中最常用的一种。Carreras[19] 和 Nicholas[20] 将 AdaBoost 引入到垃圾邮件过滤, 获得了很高的性能。Carreras 的方法中还考察了不同深度决策树作为弱学习器下的过滤性能, 他们发现, 随着深度的增加, 过滤的正确率将更高。与此类似的是, Desouza[13] 使用了决策树方法作为弱学习器在 LingSpam 语料上进行垃圾邮件过滤的两组实验, 一组是采用一层的决策树进行多遍 Boosting 循环, 另一组是采用完全的决策树进行少量 Boosting 循环。他的实验证明, 两组情况下都能取得很高的精确率(98%以上)。Androutsopoulos[9] 在实验中引入了另外一种 Boosting 方法——LogitBoost, 其弱学习器采用了对数回归方法, 学习到的是实数值而不是分类结果, 最后通过一个阈值来实现分类。从发表的结果看, LogitBoost 在 PU1 上的结果略逊于 AdaBoost。Boosting 方法的最主要缺点是训练速度较慢。

4.1.4 粗糙集(Rough Sets)方法

Rough Sets 理论是由 Pawlak 于上世纪 80 年代提出的一种研究不完整、不确定知识和数据的表达、学习、归纳的理论方法。Rough Sets 的研究对象是一个多值属性集合描述的向量集合。它通过集合的等价关系操作来确定属于给定类的最大对象集合和可能属于给定类的最小对象集合, 从而指导分类决策。Rough Sets 通常经过属性约简(消除对决策属性没有影响的属性)和属性值约简(消除对决策属性没有影响的属性值)来简化分类规则。刘洋等[22] 将 Rough Sets 引入到垃圾邮件过滤, 采用了 11 种非文本属性(包括收信人个数、中继个数等等)来进行邮件分类(正常、广告和反动)。在一个小规模的垃圾邮件样本上实验, 可以达到 80%左右的正确率。需要指出的是, 上述实验中使用的特征是所谓邮件"元信息", 并非其他实验中所使用的邮件中

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

的文字内容。实际上,同样可以使用文字内容特征来应用 Rough Sets 方法。

4.2 基于统计的方法

4.2.1 kNN 方法

kNN 是最常用的基于实例的方法。kNN 没有训练过程,分类时直接将待分类文本与训练集合中的每个文本进行比较,然后根据最相似的 k 篇文本得到新文本的类别。kNN 的原理非常直观。在文本分类中,kNN 常常能够取得较好的结果。但是由于其分类速度的局限性,不太适用于对分类速度要求较高的垃圾邮件过滤场合。尽管如此,出于研究的目的,一些文献仍然将它应用于垃圾邮件过滤领域。Androutsopoulos[7] 使用了一种类 kNN 方法,该方法使用 k 组最近的距离而不是 k 个最近的样本来计算,如果多个样本同待过滤邮件距离相差不大的话,则这些样本都将用于确定最后的结果,此时,过滤中真正使用的样本数目大于 k。实验表明,kNN 在 k 取较小值的情况性能较好,和 Naïve Bayes 的结果性能几乎相当。

4.2.2 SVM

支持向量机(Support Vector Machine,简称 SVM,也叫做支撑向量机)是在二十世纪 90 年代以来发展起来的一种统计学习方法,它通过构造最优线性分类面来指导分类。SVM 可以直接用于线性可分问题,而对于线性不可分的情形,可以构造一个变换,将问题转换到一个新的空间,在这个新空间中线性可分。在文本分类中,SVM 是公认的较好的方法之一。Drucker[5] 将线性 SVM 用于垃圾邮件过滤,得到的结果再次印证了这一点。Drucker 还指出,采用二值表示的 SVM 的性能稍高于采用多值表示的 SVM。另一个使用 SVM 的好处就是不需要进行特征选择就可以直接利用 SVM 进行过滤,其结果和经过特征选择差别不大。Androutsopoulos[9] 也在实验中引入了 SVM 方法,与 Drucker 不同的是,他使用了实数值作为特征权重。Kolcz[1] 则采用了多种 SVM 方法的变形进行垃圾邮件过滤。

4.2.3 Rocchio 方法:

Rocchio[11] 方法由是信息检索领域常常用于相关反馈的方法。它用于分类的基本思路很简单:将所有训练文本向量化,类别向量等于所有正例向量和反例向量的加权差。形式地:

$$\vec{C} = \frac{a}{|D^+|}\sum_{\vec{x}_i \in D^+} \vec{x}_i - \frac{\beta}{|D^-|}\sum_{\vec{x}_i \in D^-} \vec{x}_i$$

其中 D^+、D^- 分别表示正例和反例集合,$|D^+|$、$|D^-|$ 分别表示正例集合和反例集合的大小。α、β 为加权系数。计算得到的结果 $\vec{C}$ 表示该类的类别向量。用于垃圾邮件过滤时,通过上式可以得到垃圾邮件类的类别向量。新的邮件与类别向量计算距离,距离小于某个阈值 θ,则判定该邮件属于垃圾邮件类,否则为合法邮件。实际应用中,α 可以设定为 1,β 和 θ 可以通过训练得到(使得训练集合的分类错误率最低)。Drucker[5] 将该方法用于垃圾邮件过滤。该方法十分简洁,分类时间短,但是过滤效果稍差。Rocchio 其实是一种基于向量表示的方法。邮件和类别都表示成特征组成的向量。为了减少向量的维数,选择更"语义化"的特征。陈华辉[30] 在垃圾邮件过滤中引入了潜在语义索引方法。但文章没有关于实验结果的叙述。

4.2.4 Winnow 方法

Winnow① 是一种线性分类器。它训练的目的是为了找到某个类所有特征的权重向量 $w = < w_1, w_2, \cdots, w_N >$ (N 是特征数) 和阈值 θ,对于新文本 $x = < x_1, x_2, \cdots, x_N >$,若两者的内

① 严格地说,Winnow 实际是一种基于神经网络的方法。本文为了叙述方便,将之结为统计方法。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

积 $w^T \cdot x > \theta$, 则判定属于该类。否则, 不属于该类。Winnow 在学习 w 时采用的是一种错误驱动的方法。在训练时, 一旦发生错误, 将根据需要降低或者升高 w 里相应特征的权重值。Balanced Winnow[14] 算法是 Winnow 算法的一种, 它和普通 Winnow 算法的不同在于引入了两个权重向量 w^+ 和 w^-, 训练时通过同时变化 w^+ 和 w^- 来达到更新权值的目的。潘文锋[23] 将 Balanced Winnow 算法引入到垃圾邮件过滤, 在应用时使用了固定的阈值 θ, 它的大小是每篇训练文本所含的平均特征数, 而初始的 w^+ 和 w^- 分别取全 2 向量和全 1 向量。在不同语料库上的实验结果表明, 该方法效果接近目前所发表的最好结果, 而 Winnow 在训练速度和分类速度上具有较大的优势, 所以具有更高的实用价值。另外, 作为一种在线(On-line) 学习方法的 Winnow, 在训练集合不断扩大的情况下能够快速对分类器进行更新。

4.2.5 Bayes 方法

Bayes 方法是通过计算文本 d 属于每个类别 C_i ($i=1, 2, \cdots, M$, M 为类别个数) 的概率 $P(C_i \mid d)$, 并将它们排序取其最大值来得到 d 所属的类别。根据 Bayes 公式, 最后归结于求每个类别的概率 $P(C_i)$ 和从类别 C_i 生成文本 d 的概率 $P(d \mid C_i)$。这两个概率都可以通过训练语料得到。Naïve Bayes 是 Bayes 方法中使用最广泛的一种。在这种方法中, 假设 d 由互相独立的多个特征 w_j ($j=1, 2, \cdots, N$, N 是 d 中不同特征数) 生成, 于是 $P(d \mid C_i)$ 由可以归结成求 $P(w_j \mid C_i)$。Naïve Bayes 方法被广泛用于文本分类中, 取得了不错的效果。

已有多位学者将 Bayes 方法应用于垃圾邮件的判别。Stanford 大学的 Sahami[16] 将 Naïve Bayes 方法引入到垃圾邮件过滤进行实验。Sahami 采用了自己收集的邮件作为实验数据。值得一提的是, Sahami 除了使用词汇作为特征外, 还使用了词组特征和其他属性特征(如标题中非字母和数字字符所占的百分比), 实验结果表明, 其他属性特征能够较大幅度地提高过滤结果(精确率在 95% 左右)。Sahami 的另外一项工作是将垃圾邮件细分为色情和非色情邮件, 再加上合法邮件, 变成一个三类问题进行实验(当然实验的最终目标还是区分垃圾和合法邮件两类)。Sahami 的实验结果却表明, 将垃圾邮件判别看成三类问题反而降低了效果, 文章对该出乎意料的结果进行了分析。

Androutsopoulos[8] 也利用 Naïve Bayes 来判别垃圾邮件。他采用了公开语料 Ling-spam 进行实验。实验中考查了不同文本预处理形式对过滤结果的影响, 得出的结论如果对原始文本除去停用词和进行词汇还原, 能得出最佳的实验结果。该论文的另一个工作是提出垃圾邮件的代价因子指标, 并分析了不同过滤阈值条件下代价因子的变化情况, 文章指出, 一味地追求高的邮件正确率在系统实现时可能去造成很大的代价。本文所做实验的正确率和召回率略低于上一篇文章。

Schneider[12]、潘文峰[23] 也利用 NaïveBayes 模型来判别垃圾邮件, 他们使用了两种不同的概率估计方法: 贝努利分布模型和多项式分布模型。比较发现, 前者不仅计算上更简便, 效果上也略优于后者。

除了 Naïve Bayes 外, 不少学者还使用了其他的 Bayes 模型。IBM 的 Mertz[2] 不是采用独立性假设而是考虑使用 N 元语言模型来估计相关的概率。文章发现三元语言模型是一个很好的选择。Androutsopoulos[9] 使用了一种 Flexible Bayes 模型, 虽然该模型仍然采用独立性假设, 但是对概率的估计使用了高斯分布模型。该方法获得了很好的效果。

4.3 对比实验

目前一些学者对上述的各种方法进行了对比实验。Drucker[5] 比较了 Ripper、SVM、Boosting

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

和 Rocciho 方法,得出的结论是 SVM 和 Boosting 方法相当, Ripper 的结果最差。而 SVM 和 Boosting 相比,训练时间更短。Androutsopoulos[6,7] 比较了 Naïve Bayes、kNN 以及基于关键词过滤(Outlook 所用)的方法,得出的结论是基于关键词过滤的效果最差, kNN 和 NäiveBayes 效果相当,但是 kNN 的过滤时间较长。潘文锋[23] 比较了 Bayes 方法和 Winnow 方法,得出了结论是 Winnow 方法效果好于 Bayes, Winnow 速度上虽略微差一点,但是增量更新非常容易,是一种很有前途的垃圾邮件过滤方法。Carreras[19] 比较了决策树、Naïve Bayes 和 Boosting 方法。结论是 Boosting 方法最好, Naïve Bayes 和决策树方法性能相当,但 Naïve Bayes 的正确率高于决策树方法。

以上这些实验由于采用不同的语料,特征选择也不尽相同,比较的方面也不完善,因此,实验之间缺乏可比性。有鉴于此, Androutsopoulos[9] 使用 4 种不同语料,对 4 种垃圾邮件过滤方法(Naïve Bayes、Flexible Bayes、SVM 以及一种 Boosting 方法 LogitBoost)使用相同的特征选择方法,在每种语料上进行了时间、性能、代价因子等的全面比较。4 种算法的时间分析如表 3 所示。

表 3　算法时间比较

算　法	训练时间	过滤时间
Naïve Bayes	$O(mN)$	$O(m)$
Flexible Bayes	$O(mN)$	$O(mN)$
SVM	$O(m^2N^2)$	$O(m^2N)$
LogitBoost	$O(m'mN^2)$	$O(m')$

其中, N 是训练邮件个数, m 是特征个数, m' 是 LogitBoost 算法中的循环次数。从表中可以得到, Naïve Bayes 方法训练时间和过滤时间都相对较短。而 SVM 方法的训练时间和过滤时间都较长。虽然如此,文章同时指出,以上分析基于最坏情况,在算法实际实现时,通常的时间代价会远小于理论分析。

文章的对比实验发现,在相同特征的情况下, Flexible Bayes 方法在性能和代价因子等方面占据上风。其他方法则大抵相当(考虑到实现的可行性,作者的 Boosting 方法中采用了较少的循环次数)。作者同时比较了采用单个词或者 N-gram 做为特征的过滤效果,结果发现 N-gram 在加大计算复杂度的同时并不能显著提高结果性能。文章在此基础上介绍了一个实际的垃圾邮件过滤原型系统的流程和实现。

目前,已经实现的垃圾邮件过滤方法的大致比较可以参见下表①。从表中可以看出, Naïve Bayes 和 Rocchio 在训练和分类上速度占优但结果一般,而 Flexible Bayes、SVM、Boosting、Winnow 方法分类速度很快结果性能很好。Boosting 在训练速度上处于下风。如果考虑更新的方便程度,则 Winnow 算法占据优势。

表 4　目前各种垃圾邮件过滤方法的大致对比

Classifier	Training	Classification	Update	Storage	Performance
Ripper	* *	*	* * *	* * *	* * * * *
Rocchio	*	*	*	*	* * * * *
Naïve Bayes	*	*	* * *	*	* * *
Flexible Bayes	* *	*	* * * *	*	*
SVM	* * *	*	* * * *	*	*
KNN	None	* * * * *	None	* * * * *	* * *
Boosting	* * *	*	* * * *	* *	*
Winnow	* *	*	*	*	*

*越多表示这种方法越耗时间、空间或者性能更差,相同的 * 个数表示两种方法大概处于同一量级。

需要指出的是,垃圾邮件过滤使用了文本分类方法,但其不完全等同于文本分类。首先,垃圾邮件具有自己的特殊特征。一些方法在实验时同时考虑这些特殊特征。比如,刘洋[22] 的

① 该表只是通常情况下的大致比较,实际上,有些方法受参数影响较大。比如:规则方法的速度和所使用规则的数目有很大的关系。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

实验中只使用邮件的元信息(发信人、收信人等)。Sahami[16]通过加入垃圾邮件本身的一些特征(如非正常字符的比率)来提高过滤效果。其次,很多学者都考虑了垃圾邮件过滤中不同错误的代价,从而通过改造基本的学习器来适应这种代价计算。改造中主要的思想就是提高样本中合法邮件的权重。再次,垃圾邮件过滤不仅要考虑过滤性能,也关注实际应用的时空效率。最后,由于邮件到达的时间序列性,垃圾邮件过滤还有一个随时间不断更新以适应用户兴趣变化或到达邮件特征发生变化的情况。更新问题在文献[23]中略有提及。但是相关的详细研究进展还未见到。

5 总结和展望

垃圾邮件问题已经受到了包括学者、商业界、法律界等各界人士在内的广泛关注,目前已经有专门针对该领域的多项国际会议。基于内容的垃圾邮件过滤是解决垃圾邮件的重要方法。本文总结了目前该方法的研究现状。目前的研究主要借鉴机器学习的方法,从已有的训练语料中学习垃圾邮件的规律来指导后续邮件的判断。包括 Bayes、决策树、SVM、Boosting 在内的很多机器学习方法都被应用到垃圾邮件过滤中。从目前的研究结果看,Flexible Bayes、SVM、Boosting、Winnow 等这些机器学习方法在一些小规模语料上似乎可以达到实用的程度。但是,在实际应用时仍然有许多工作要做。这是因为,垃圾邮件过滤是一项长期的斗争。在我们对付垃圾邮件的同时,垃圾邮件制造者也在挖空心思制造更"合理"或者严重干扰过滤器的垃圾邮件。因此,一方面,对垃圾邮件的预处理显得越来越重要(即还原到邮件的真实内容);另一方面,垃圾邮件过滤器也要不断地更新来适应垃圾邮件的一些新特征。另外,在真正的实用垃圾邮件系统中,综合各种方法(包括各种机器学习方法、黑白名单人工规则方法甚至图片分析方法等)和各种特征(除正文内容外,还包括群发特征、元信息特征等)是垃圾邮件工具研制的趋势。

参 考 文 献:

[1] A. Kolcz and J. Alspector, SVM-based Filtering of E-mail Spam with Content-specific Misclassification Costs[A]. In: Proc. ICDM-2001 Workshop on Text Mining (TextDM 2001)[C]. Nov. 2001.

[2] D. Mertz, Six appQroaches to eliminating unwanted e-mail[EB]. from http://www-900.ibm.com/developerWorks/cn/linux/other/l-spamf/index_eng.shtml, September, 1999.

[3] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists, Information Retrieval[J]. Vol. 6, No. 1, pp. 49—73, Kluwer Academic Publishers, 2003.

[4] H. Katirai, Filtering Junk E-Mail: A Performance Comparison between Genetic Programming & Naive Bayes[D]. available online at: http://members.rogers.com/hoomank/katirai99filtering.pdf, Sep. 1999.

[5] H. Drucker, D. Wu, and V. N. Vapnik, Support Vector Machines for Spam Categorization[J]. IEEE Transactions on Neural Networks, Vol. 20, No. 5, pp. 1048—1054, Sep. 1999.

[6] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos and C.D. Spyropoulos, An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal E-mail Messages[A]. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)[C], Athens, Greece, pp. 160—167, 2000.

[7] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C.D. Spyropoulos and P. Stamatopoulos, Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach[A]. In: Proc. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000)[C]. pp. 1—13,

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www

Sep. 2000.
[8] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras and C. D. Spyropoulos, An Evaluation of Naive Bayesian Anti-Spam Filtering[A]. In: Proc. of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)[C]. pp. 9—17, May 2000.
[9] I. Androutsopoulos, G. Paliouras and E. Michelakis, Learning to Filter Unsolicited Commercial E-Mail[EB]. Technical report 2004/2, NCSR "Demokritos", 2004.
[10] J. M. G. Hidalgo, Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization[A]. In: Proceedings of ACM Symposium on Applied Computing (SAC 2002)[C]. pp. 615—620, Mar. 2002.
[11] J. Rocchio, Relevance feedback in information retrieval[M]. In: the SMART Retrival System: Experments in Automatic Document Processing, pp. 313—323, PrenticeHall Inc., 1971.
[12] K. Schneider, A Comparison of Event Models for Naïve Bayes Anti-Spam E-Mail Filtering[A]. In: Proc. 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)[C]. Budapest, Hungary, pp. 307—314, Apr. 2003.
[13] M. DeSouza, J. Fitzgerald, C. Kemp and G. Truong, A Decision Tree based Spam Filtering Agent[EB]. from http://www.cs.mu.oz.au/481/2001_projects/gntr/index.html, 2001.
[14] N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm[J]. Machine Learning, 2(4): 285—318, 1988[J].
[15] R. Krishnamurthy and C. Orasan, A corpus-based investigation of junk emails[A]. In: Proceedings of Language Resources and Evaluation Conference (LREC 2002)[C]. Las Palmas de Gran Canaria, Spain, pp. 1773—1780, May 2002.
[16] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, A Bayesian approach to filtering junk e-mail[A]. In: Proc. of AAAI Workshop on Learning for Text Categorization[C]. pp. 55—62, 1998.
[17] W. Cohen, Fast effective rule induction[A]. In: Machine Learning Proceedings of the Twelfth International Conference[C]. Lake Taho, California, Mongan Kanfmann, pp. 115—123, 1995.
[18] W. Cohen, Learning rules that classify email[A]. In: Proceedings of the AAAI spring symposium of Machine Learning in Information Access, Palo Alto[C]. California, pp. 18—25. 1996.
[19] X. Carreras and L. Marquez, Boosting Trees for Anti-Spam Email Filtering[A]. In: Proceedings of Euro Conference Recent Advances in NLP (RANLP-2001)[C]. pp. 58—64, Sep. 2001.
[20] T. Nicholas, Using AdaBoost and Decision Stumps to Identify Spam E-mail[EB]. Stanford University Course Project (Spring 2002/2003) Report, from http://nlp.stanford.edu/courses/cs224n/2003/fp/.
[21] Y. Diao, H. Lu and D. Wu, A Comparative Study of Classification Based Personal E-mail Filtering[A]. In: Proceedings of PAKDD—2000[C], pp. 408—419, Apr. 2000.
[22] 刘洋, 杜孝平, 罗平, 侯志辉, 郭晨, 骆焕林, 等. 垃圾邮件的智能分析、过滤及 Rough 集讨论[A]. 第十二届中国计算机学会网络与数据通信学术会议[C]. 武汉, 2002 年 12 月
[23] 潘文峰. 基于内容的垃圾邮件过滤研究[D]. 北京: 中国科学院计算技术研究所, 2004. 7.
[24] 赵晓明, 郑少仁. 电子邮件过滤器的分析与设计[J]. 东南大学学报: 自然科学版, 2001, 31(5): 19—23.
[25] 郭泓. 电子邮件过滤技术浅析[J]. 信息网络安全, 2002(10): 42—44.
[26] 落红卫, 刘建毅, 等. 智能邮件过滤系统的研究与实现[J]. 机电产品开发与创新, 2003(1): 51—52.
[27] 谭立球, 谷士文, 等. 个人化电子邮件自动过滤系统的设计[J]. 计算机应用, 2002, 22(6): 54—55.
[28] 张长君. 电子邮件的一种过滤方法[J]. 计算机安全, 2002(12): 42—43.
[29] 王庆波, 方滨兴. 电子邮件过滤检测系统的设计与实现[J]. 计算机应用研究, 2000, 17(10): 105—106.
[30] 陈华辉. 一种基于潜在语义索引的垃圾邮件过滤方法[J]. 计算机应用研究, 2000, 17(10): 17—18, 35.
[31] 刘斌, 黄铁军, 程军. 一种新的基于统计的自动文本分类方法[J]. 中文信息学报, 2002(6): 18—24.
[32] 李渝勤, 孙丽华. 基于规则的自动分类在文本分类中的应用[J]. 中文信息学报, 2004(4): 9—14.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www