doi: 10.3969/j.issn.1673-4785.2010.03.001

反垃圾电子邮件方法研究进展

谭 营[1,2],朱元春[1,2]

(1. 北京大学 机器感知与智能教育部重点实验室,北京 100871; 2. 北京大学 信息科学技术学院,北京 100871)

摘 要: 随着垃圾电子邮件对互联网技术的威胁日益严峻,反垃圾电子邮件研究已成为当今的研究热点. 综述了反垃圾电子邮件研究的历史、现状和最新进展. 首先介绍并分析了3种类型的邮件特征提取方法——基于文本、图片和行为的特征提取方法. 然后,在此基础上,详细论述了当前的反垃圾邮件技术——法律手段、简单方法和智能型处理技术. 接着,介绍了反垃圾邮件系统性能评估准则和标准数据集. 最后,对反垃圾电子邮件现状进行了分析和总结,并指明了未来反垃圾电子邮件研究的发展方向,包括改进邮件特征提取技术、完善相关法案和引入新的智能反垃圾邮件方法.

关键词: 反垃圾电子邮件; 特征提取; 智能检测技术; 性能评估

中图分类号: TP393 **文献标识码:** A **文章编号:** 1673-4785(2010)03-0189-13

Advances in anti-spam techniques

TAN Ying[1,2], ZHU Yuan-chun[1,2]

(1. Key Laboratory of Machine Perception (MOE), Peking University, Beijing 100871, China; 2. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China)

Abstract: As the threat of spam on the Internet grows increasingly severe, anti-spam techniques have become a hotspot for researchers. The authors reviewed the history, current situation, and latest advances in research on spam control. First, we introduced and analyzed three different types of feature extraction methods for email. These were text-based, image-based, and behavior-based approaches. Then, current anti-spam techniques were described and discussed. These included laws, simple methods, and intelligent approaches. After that, performance evaluation methods and standard data sets were discussed. Finally, we summarized the current research on anti-spam techniques and pointed out directions for future research, including improvements to e-mail feature extraction techniques, improvements to laws, and new intelligent anti-spam approaches.

Keywords: anti-spam; feature extraction; intelligent detection technique; performance evaluation

随着信息技术的持续发展和互联网的日益普及,电子邮件(E-mail)已成为人们日常通讯交流的重要方式之一. 然而,垃圾电子邮件(unsolicited bulk email—UBE, or Spam)的涌入,给电子邮件通讯带来诸多不便,引发了日益严重的问题. 垃圾电子邮件不仅会耗费通信带宽、网络资源,而且消耗人们大量的处理时间,造成生产力浪费,使公司蒙受巨大经济损失. 因此,垃圾邮件检测技术和方法的研究,已成为国内外研究的热点,具有必要性和重大意义.

在反垃圾电子邮件技术研究中,学者们相继提出众多的邮件特征提取方法和垃圾邮件检测过滤方法. 本文是对反垃圾邮件技术和方法研究现状的综述,重点介绍以下内容: 垃圾电子邮件的现状、用于垃圾邮件检测的邮件特征提取方法、现有的反垃圾邮件技术以及反垃圾邮件系统评估准则和标准数据库.

1 垃圾电子邮件现状

1.1 定 义

在反垃圾电子邮件技术研究中,一些专家学者和研究机构给出不同的垃圾电子邮件定义.

收稿日期: 2009-11-20.
基金项目: 国家"863"计划资助项目(2007AA01Z453); 国家自然科学基金资助项目(60673020,60875080).
通信作者: 谭 营. E-mail: ytan@pku.edu.cn.

Cranor 等人[1]将其定义为 “未经请求的大量电子邮件(unsolicited bulk email, UBE) ”. 垃圾电子邮件还被定义为[5]: “未经请求的商业电子邮件(unsolicited commercial email, UCE) ”. 中国互联网协会将垃圾电子邮件定义为[6]: 收件人事先没有提出要求或者同意接收的广告、电子刊物、各种形式的宣传品等宣传性的电子邮件; 收件人无法拒收的电子邮件; 隐藏发件人身份、地址、标题等信息的电子邮件; 含有虚假的信息源、发件人、路由等信息的电子邮件; 含有病毒、恶意代码、色情、反动等不良信息或有害信息的邮件.

以上 3 种定义尽管不同,却有着一个共同点: 未经请求. 这是垃圾电子邮件与正常电子邮件的本质区别. 正常电子邮件是人们正常通讯、交流的媒介,包含着交互信息的需求. 而垃圾电子邮件往往包含收件人不感兴趣的内容,且在未经许可的情况下发送给收件人. 垃圾电子邮件一般包含商业广告信息,且成批量发送,这也是定义其为 UBE、UCE 的原因. 同时,垃圾电子邮件发送者为逃避对电子邮件的反向追踪,会刻意伪造发件人、路由、信息源等信息. 故在多数情况下,这 3 种定义是一致的.

1.2 垃圾电子邮件的规模与影响

根据 Symantec 公司的统计报告,2008 年全球范围垃圾电子邮件的平均比例已经占到了总邮件数的 80% 左右[4]. 依据 Ferris Research 的研究估计[5],2009 年垃圾电子邮件将耗费全球1 300亿美元的开销,其中,劳动力浪费引起的开销占总开销的 85%. 这将比 2007 年的估计增长 30%,而比 2005 年的数据增长 100%. 根据 Sophos 公司最新调查结果显示[6],中国的垃圾电子邮件的数量继美国、巴西之后,位列第 3 位. 图 1 显示出各国家的垃圾电子邮件数量比例.

中国互联网协会 2009 年第一季度中国反垃圾电子邮件调查结果[7]指出,中国网民平均每周收到 17.68 封垃圾电子邮件,与去年同比增加 0.04 封,占邮件总数的 57.52%. 图 2 给出中国网民在 2008 年第一季度至 2009 年第一季度平均每周收到垃圾电子邮件的比例. 调查报告还指出,处理这些垃圾电子邮件将耗费中国网民平均每周 12.35 min. 仅考虑浪费时间的因素,2009 年第一季度垃圾电子邮件致使中国损失人民币 339.59 亿元,与 2007 年同比增长 151.19 亿元,涨幅为 80.25%. 鉴于垃圾电子邮件所引发的这些严重社会问题,近年来,反垃圾电子邮件策略受到了前所未有的关注.

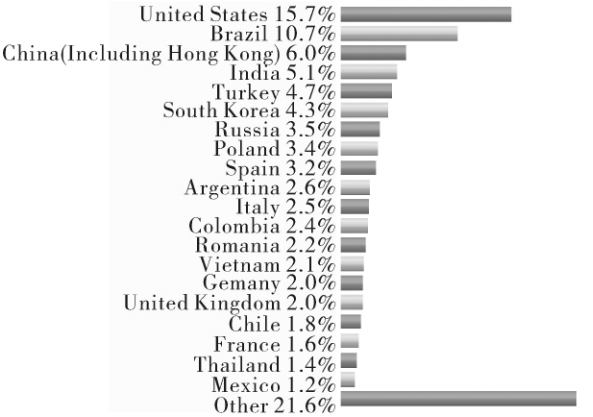


图 1 各国家垃圾电子邮件数量比例

Fig. 1 The proportion of spam-relaying of different countries

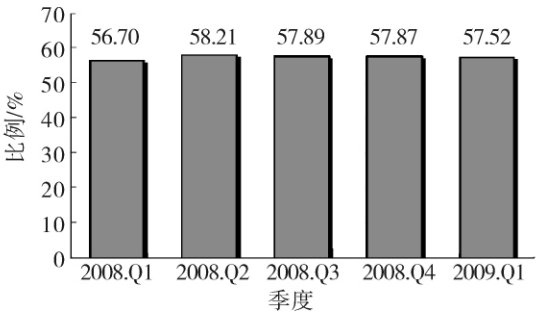


图 2 中国网民平均每周收到垃圾电子邮件的比例

Fig. 2 Weekly average ratio of spam received by cybercitizens in China

2 邮件特征提取方法

对于垃圾邮件检测系统来说,邮件特征提取是极其关键的环节,甚至比模式识别方法的选择、分类器的设计与使用更为重要. 邮件特征提取方法的准确性、可区分性、稳定性和自适应性将会直接影响到系统整体的分类效果与性能. 据中国互联网协会 2008 年第四季度中国反垃圾邮件调查统计[8],用户收到垃圾邮件的正文格式主要是 3 种: 图片 + 文本格式、纯文本格式和纯图片格式. 本节将综述经典的基于文本的邮件特征提取方法、基于图片的邮件特征提取方法和基于行为的邮件特征提取方法.

2.1 基于文本内容的邮件特征提取方法

基于文本内容的邮件特征提取方法一般包含 2 个阶段: 1) 词筛选(terms selection) : 依据词的重要性(可区分度) 对特征词进行排序,选择可区分度好的特征词进入下一阶段; 2) 特征提取与表示: 提取出邮件特征并表示成统一的形式.

2.1.1　文本词筛选方法

当邮件库中的邮件经历切词阶段后,大量的单词被获取,如果不经过词筛选过程,会导致特征维度过高,引发维度灾难. 词筛选一方面可以降低特征维度和计算复杂度,另一方面还可以减小噪声(区分度差的单词)的不良影响. 下面介绍几种常用的词筛选方法:

1) 信息熵.

在信息论中,信息熵(IG)又被称为 Kullback-Leibler 距离[9]. 它能够度量 2 个概率分布 $P(x)$ 和 $Q(x)$ 的距离. 在垃圾邮件检测技术研究中,它被用于度量单词的优良度(区分度). 根据该方法,可以计算出,当知道给定单词 t_i 是否在邮件中出现时,所能获得的邮件类型信息的量. 单词 t_i 的信息熵被定义如下:

$$I(t_i) = \sum_{C\in\{c_s,c_l\}}\{\sum_{T\in(t_i,\bar{t}_i)} P(T,C)\log\frac{P(T,C)}{P(T)P(C)}\}.$$

式中: C 表示邮件类型,c_s 和 c_l 分别表示邮件类型是垃圾邮件(spam)和正常邮件(legitimate email),t_i 表示单词 t_i 在邮件中出现,而 $\bar{t}_i$ 表示单词 t_i 未在邮件中出现. 式中的概率可以根据训练集数据进行估计. 根据该式,每个单词的信息熵值将被计算出来,信息熵值大的单词将被选择进入下一阶段.

2) 词频方差.

Koprinska 等人[10]研究出词频方差法(term frequency variance, TFV),来选取具有高词频方差的词. 他们认为词频方差大的词包含更多的信息量. 依据该方法,那些倾向于出现在某一种类型邮件(垃圾邮件或正常邮件)的词将被选择,而那些在 2 种类型邮件中出现频率相当的词将被移除. 在反垃圾邮件技术研究领域中,词频方差被定义如下:

$$T(t_i) = \sum_{C\in\{c_s,c_l\}}[T_f(t_i,C) - T_f^{\mu}(t_i)]^2.$$

式中: $T_f(t_i,C)$ 表示单词 t_i 在类型为 C 的邮件中的出现频率,$T_f^{\mu}(t_i)$ 表示单词 t_i 在 2 种类型邮件中出现的平均频率.

文献[10]指出在多数情况下,词频方差方法性能优于信息熵方法. 具有最大信息熵值和最大词频方差的前 100 个词的对比显示,这些词具有以下特征: a) 在内容为语言学相关的正常邮件中频繁出现; b) 在垃圾邮件中频繁出现,却在正常邮件中极少出现.

3) 文档频率.

文档频率(document frequency, DF)指的是某一特定的单词 t_i 所出现过的邮件的数量. 依据该方法,文档频率值大于预设阈值的词将被选择,而文档频率值小于该阈值的词将被舍弃. 单词 t_i 的文档频率被定义如下:

$$D(t_i) = |\{m_j \mid m_j \in M, t_i \in m_j\}|.$$

式中: M 表示整个训练集,m_j 表示 M 中的一封邮件.

文档频率法认为低频单词所含的类别信息量较少,移除它们不会影响整体分类性能. 文献[11]指出,当移除 90% 的低信息量单词时,文档频率方法与信息熵和 χ^2 统计量方法的性能相当. 文档频率方法的主要优点是,计算复杂度低,与训练样本的数量成线性比例增长.

4) 其他词筛选方法.

词筛选方法在垃圾邮件检测系统中起着重要的作用. 为了更好地理解词筛选方法,下面列出 3 种其他的常用方法的计算式[11-13].

a) χ^2 统计量(CHI):

$$\chi^2(t_i,c) = \frac{|M|(P(t_i,c)P(\bar{t}_i,\bar{c}) - P(\bar{t}_i,c)P(t_i,\bar{c}))^2}{P(t_i)P(\bar{t}_i)P(c)P(\bar{c})};$$

b) 比值比(odds ratio):

$$\tau(t_i,c) = \frac{P(t_i\mid c)}{1-P(t_i\mid c)}\cdot\frac{1-P(t_i\mid \bar{c})}{P(t_i\mid \bar{c})};$$

c) 术语强度(terms strength):

$$S(t_i) = P(t_i\in y \mid t_i\in x).$$

式中: $c\in\{c_s,c_l\}$ 表示给定的邮件类型,相应的 $\bar{c}\in\{c_s,c_l\}/c$,x 和 y 表示训练集中类型相同的任意 2 封不同邮件.

2.1.2　文本特征提取方法

1) 词汇袋法

词汇袋法(bag-of-words, BoW)也被称为向量空间模型,是垃圾邮件检测技术研究领域应用最广泛的方法之一[2]. 通过观察特征词是否在邮件中出现,将每封邮件转换成一个 d 维的特征向量 $<x_1,x_2,\cdots,x_d>$,其中每维特征值 x_i 可以看作是特征词 t_i 的函数. 对于 x_i,有 2 种常用的类型表示方法: 布尔型和频率型[4]. 在布尔型表示下,x_i 按下列方式赋值: 若 t_i 在邮件中出现,那么给 x_i 赋值 1,否则给其赋值 0. 如果采用频率类型表示,那么 x_i 则表示为该邮件中特征词 t_i 的词频. Schneider 的实验显示,这 2 种类型的表示法性能相当[5].

2) 稀疏二元多项式哈希.

稀疏二元多项式哈希(sparse binary polynomial

hashing, SBPH)运用滑动窗口方法,能够从邮件中提取出大量的不同特征[6-17]. 它使用一个长度为 N 个单词的滑动窗口依次滑过邮件中的单词,窗口移动步长为1个单词. 在每次窗口的滑动中,都将按以下方式提取 2^{N-1} 个特征:最新进入窗口的单词被保留,而窗口中的其他单词被选择保留或删除,选择之后,整个窗口被整体映射为一个特征. 对于窗口中的 $N-1$ 个单词,存留选择有 2^{N-1} 种,故可映射成 2^{N-1} 个不同的特征. 然后,每个特征将被计算为一个特定的哈希值,特征提取之后可以根据前面介绍的词筛选方法进行特征筛选,以降低特征维度. 该方法的分类准确度较高,但因为特征数量的庞大计算复杂度很高.

3) 正交稀疏双词.

为了降低SBPH方法的冗余度和复杂度,Siefkes等人[7]提出正交稀疏双词法(orthogonal sparse bigrams, OSB)来提取一个较小的特征集合. 该方法同样使用长度为 N 个单词的滑动窗口提取特征,与SPBH方法不同的是,只有具有共同单词的单词对被提取作为特征. 对于每个窗口来说,最新进入窗口的单词被保留,并作为共用单词. 然后,从剩下的 $N-1$ 个单词中选择1个与其组成单词对,如此每个窗口可以构造出 $N-1$ 个单词对,映射出 $N-1$ 个特征. 与SPBH方法相比,这样做大大减少了特征的数量. 文献[7]中的实验表明OSB性能略优于SBPH方法.

4) 基于人工免疫系统.

Oda等人[8]设计出一种反垃圾邮件免疫模型,运用正则表达式构造抗体(检测器). 正则表达式的运用,使得每个抗体都能够匹配大量的抗原(垃圾邮件),这样能有效降低抗体(特征)集合. 模仿生物免疫系统(biological immune system, BIS)的功能,他们给每个抗体赋予不同的权重. 算法初期,所有的抗体权重被初始化为一个缺省值,经过一段时间的运行,那些匹配垃圾邮件较多的抗体的权重将被增加,而那些与正常邮件匹配的抗体的权重将被降低. 当抗体的权重低于预设阈值时,该抗体将从系统模型中被移除.

Ruan等人[9]提出一种基于免疫浓度的特征构造方法. 该方法根据单词的倾向性构建出2个基因库. 若一个单词在垃圾邮件中出现频率高(倾向在垃圾邮件中出现),那么将该单词添加到垃圾邮件基因库,否则将其添加到正常邮件基因库. 然后,根据邮件中单词在2个基因库中的出现情况计算出每封邮件的"自己浓度"和"异己浓度". 这2个浓度值共同构成邮件的二维特征向量.

2.2 基于图片的邮件特征提取

为了避开垃圾邮件检测系统的过滤,垃圾邮件发送者有时会采用图片型邮件来发送广告信息. 检测这类垃圾邮件的关键在于提取有效的图片特征. 目前,基于图片的特征提取研究仍处于初步,常用的图片特征包括以下方面:

1) 图像属性特征.

这些特征包括图片类型、大小、颜色、饱和度等. 垃圾邮件发送者往往倾向选择高压缩率的图像格式,从而能够在较短时间内发送出大量的垃圾邮件. 故可以选取图片的类型作为其中一个特征,来检测图片型垃圾邮件[10]. 图像的这些属性均包含了一定的类别信息,广告图片的这些属性值往往与正常邮件有一定的差异.

2) 边缘特征.

相对正常邮件来说,垃圾邮件图像中往往包含更多的文字信息. 而包含大量文字的图片会具有不同的边缘特性. 因此可以利用边缘特性,如:方向性、边缘强度、边缘轮廓形状,来有效地检测垃圾邮件[11].

3) 文字特征.

可以利用文字识别工具将图片中的文字提取出来,然后对文字进行语言分析、关键词匹配,也可以采用基于文本的特征提取方法,从而有效检测垃圾邮件.

4) 其他特征.

除了上述特征外,可以利用图片的纹理特征、异质特征、噪声特征等有效地对邮件类型进行区分,对垃圾邮件进行过滤.

2.3 基于行为的邮件特征提取方法

基于行为的垃圾邮件检测技术是一种新型过滤垃圾邮件的手段,通过提取垃圾邮件与正常邮件有区分的行为特征,来过滤垃圾邮件. 本节对常用的基于行为的反垃圾邮件技术进行综述,从4个方面阐述常用的邮件行为特征:基于邮件头部信息及系统日志的行为特征、基于附件的行为特征、基于网络的行为特征以及基于用户行为的特征.

2.3.1 基于邮件头部信息及系统日志的行为分析

正常情况下,邮件的头部信息能反映邮件传送

信息及发信人的基本意图: 发件人、收件人、抄送、发送时间等. 一般情况下,正常的邮件在这些条目中将用正确的格式填入完整的信息. 但为了避开一些常用的反垃圾邮件机制,垃圾邮件发送者往往在这些条目中填入伪造的数据和错误的格式.

文献[22]针对这种行为模式提出一种基于行为的反垃圾邮件机制: 首先,该文献在头部信息中选取最能区分出垃圾邮件的 7 个条目,如 From field、To field、Reply-To field 等; 然后,基于这些基本的特征,从他们的交叉比对组合中选出 10 个特征,如 From-To、From-Reply-To 等; 接着针对各条目的数据正确、错误、伪造类型分别定义出相应的类别,并进行编码,得到 113 维的特征向量; 最后,作者使用支持向量机、贝叶斯和决策树 3 种分类方法对特征化后的邮件数据库进行分类. 实验中,支持向量机在各数据集上的性能优于其他 2 种方法,但决策树有较高的准确度. 相对于基于内容的机制来说,该机制拥有较高的准确度、较低的特征维度和较低的时间复杂度.

文献[23-24]在此基础上加入系统日志中的一些条目信息作为特征,并利用一种增强型的 BP 神经网络对特征化后的邮件数据进行分类,根据各特征的重要程度赋予各个特征不同的权重. 文献[25]指出,有 190 多个头部信息条目和 23 个系统日志条目可以被邮件用户代理/邮件传送代理(mail user Agent/mail transfer Agent, MUA/MTA) 使用. 文献[23]研究探讨了多达 13 种形态 24 种类型的垃圾邮件行为形态,选取 32 个基本条目及 38 个交叉比对条目提取特征. 并且还进行实验验证交叉比对条目的重要性. 文献[24]观察得出,MUA/MTA 并没有使用所有的头部信息和系统日志条目,文中选出 6 个最有意义的头部信息条目和 4 个最有意义、最高出现频率的系统日志条目,以及基于此选择出 16 个交叉比对条目进行研究实验.

文献[26]提出基于行为的分阶段过滤垃圾邮件技术. 在过滤的过程中,该机制不仅分析处理到目前阶段为止的所有行为信息,而且还特定分析处理新增的行为信息. 根据 SMTP 协议,它将处理分为 4 个阶段: HELO、FROM、RCPT TO 和 DATA, 利用各个阶段中的属性信息进行分类处理. 如果邮件在前一个阶段中被确定分类为垃圾邮件,那么邮件就会被直接拒绝掉,而不会进入下一个阶段,这样做能够节省资源. 文章采用贝叶斯分类方法,实验效果在时间性能和资源耗用上优于其他的一些算法.

文献[27]对发送人 IP 地址、SMTP ID 序列、URL 连接和回复邮件地址进行分析,对其按照设定的公式计算相应邮件的评分,然后用人工免疫系统对处理过的数据进行分类. 该机制具有可靠性、有效性和可扩充性.

文献[28]针对 IP 和域名,发送者、接收者的对应关系,发送者、接收者邮件地址的长度,以及发送频率等信息为特征,用决策树进行分类.

2.3.2　基于附件的行为分析

文献[29-30]分析邮件的附件行为用于发现带病毒的可疑垃圾邮件. 文中 MET 客户端(malicious email tracking) 采用 MD5 哈希技术给每个附件赋予一个特定标识,并保存一个相关记录(标识、时间戳、附件有无病毒、发件人地址、收件人地址) . MET 服务器端接收 MET 客户端的信息,并根据附件的特征进行分析处理——病毒事件、附件产生率、病毒生命周期、病毒事件频率、病毒死亡率、病毒流行程度、病毒威胁、病毒传播等. 当 MET 客户端发现某一附件的产生率或流行率大于给定的阈值时,将会对其他的特征进行进一步分析,来确定是否为病毒. 如果是病毒,就将此报告给中心服务器. 中心服务器将会基于其他客户端关于此附件的报告来作出最终决定,判明其是否为病毒. 若为病毒,则将相关标识、病毒死亡率、该种病毒发生频率等信息发给客户端,来避免将来的感染. 如果客户端提供了邮件地址和 IP 地址,那么就可以根据信息追踪出病毒的制造者.

文献[31]提到将邮件携带附件的类型(图片、二进制文件、文本文件等) ,以及附件的数量作为区分垃圾与非垃圾邮件的行为特征.

2.3.3　基于网络的行为分析

1) 基于社会网络的特征提取.

文献[10,32]分析邮件传送过程中的簇行为特征,即用邮件经常交流的一些人形成特定的簇,邮件发送行为一般发生在簇内部. 比如说,一般情况下,一个用户不会将同一个邮件信息同时发送给他的配偶、上司、朋友等,这种概率非常小. 然而一个对用户地址簿的攻击者显然不知道这些社会关系模式,当他试图给地址簿中的所有人发送邮件时就会违反正常邮件的簇行为特征. 从概念上来说,有 2 种簇模式: 用户簇模式和群落簇模式.

用户簇模式通过对单个用户帐户的邮件历史分析计算得到. 对于某一邮件来说,收件人列表(收件人、抄送、密送)中的所有帐户看作一个簇关系. 为了避免簇的数量过大,以及冗余现象,只选定那些最大化的簇,即所选定的每个簇都不是其他簇的子集. 例如,有3个收件人列表: [A,B,C], [A,B]和[A,B,D],则会选择2个作为簇——[A,B,C]和[A,B,D]. 若某一邮件的收件人列表不是任何用户簇的子集,那么称其为不一致簇行为. 这种方法往往要与其他模型结合使用,以处理特殊的收件人列表情况. 如果用户曾发过一个全体收件人列表的广播邮件,那么该机制就会失效. 然而,这种情况较少发生,一般情况下,用户只会给地址簿中少于10%的帐号同时发送邮件.

群落簇模式通过2个用户间的邮件交流数量建立相应的联系. 若两帐户间交换的邮件数量超过给定阈值,那么就认为这两帐户间存在联系. 然后,利用层次算法,逐步建立大小为 n 的簇. 例如,当前层次为2,存在AB、AC、AD、BC、BD、CE 6个簇. 只有当只是最后一个成员不同时,2个簇才能进行融合,以避免重复. 例如,AB、AC形成候选簇ABC,但是AB、BC不再融合. 当所有候选簇形成完毕后,要对其合法性进行检查. 只有当前层次中同时存在AB、AC、BC时,候选簇ABC才是合法的. 最后,要将那些是其他簇的子集的簇去掉,如AB、AC、BC将会被去除. 如此进行下去,形成大小为 n 的群落簇.

文献[13]定义3种类型的图,来描述邮件的发送行为:有向图、无向图和差分图. 在有向图中,节点代表至少进行了一次发送或接收行为的电子邮件用户,有向图的边表示一用户从另一用户那里接收或向其发送了一封邮件. 无向图中,节点代表至少进行了一次与另一用户发送和接收行为的那些邮件用户,边代表两用户间交换了信息. 差分图是基于2个有向图建立的,用于表示那些存在某一图中,而不存在于另一图中的那些边. 基于此,算法共分为3个阶段: a) 基于服务器的系统日志,建立3种类型的图; b) 利用有向图和无向图,对邮件发送者进行初步分类,列入黑名单、白名单或灰名单; c) 利用差分图,对b)阶段的分类结果进行调整,得到最终分类结果.

2) 邮件的网络分布特征.

文献[14]分析垃圾邮件发送的网络层次行为,是首次分析垃圾邮件、僵尸网络和网络路由的相互关系. 该文献通过研究IP地址空间分布特征,来分析垃圾邮件发送者、垃圾邮件僵尸网络和正常邮件发送者的网络分布. 大多情况下,正常邮件与垃圾邮件分布大致相同,大多数的邮件都来自一小部分IP地址空间. 但有一小部分例外的情况,在地址段80.* ~90.* 中,绝大多数邮件都是垃圾邮件,在地址段60.* ~70.* 中,绝大多数邮件都是正常邮件. 这表明可以将IP地址作为一个区分特征. 该文献还分析了僵尸网络的行为特征,分析得出:绝大部分的垃圾邮件是从Windows操作系统中发出的,并且有很大比例(25%)的垃圾邮件来自僵尸网络. 65%的已感染的IP地址仅发送了一次垃圾邮件,且其中75%发送时间短于2 min. 由于这些IP地址生命周期短,这种情况使得黑名单方法失效. 研究还表明,每个僵尸网络节点在整个周期发送的垃圾邮件数量少于100封. 垃圾邮价发送者利用大量的僵尸网络节点发送垃圾邮件,且对每个节点来说,只利用很短的时间,发送少量的邮件. 因此,基于黑名单和发送数量的方法对这种情况都会失效. 文献还分析了边界网关协议(border gateway protocol, BGP),用路由广播传播垃圾邮件. 该机制使用了大量的IP地址空间,并且发送者在空间中分散分布,使得不容易被察觉. 目前使用这种机制发送的垃圾邮件比例还很小,大约为1% ~10%.

2.3.4 基于用户行为的技术

文献[15]分析用户的行为特征,用户查收邮件可以归纳为以下几类行为:在远程邮件箱中将认为无用的邮件删除;打开邮件并且阅读时间超过给定阈值 N;打开邮件但在低于 N 将邮件删除;将邮件移存至邮件箱目录;回复、转发邮件;将发件人加入通讯簿. 通过收集这些用户处理邮件的行为信息,该方法将其作为垃圾邮件检测系统的反馈信息,将处理的信息反馈给反垃圾邮件网关. 网关可以将界定的垃圾邮件作为其他过滤器的训练或学习样本,提交共享黑名单等. 另外,还应清除邮件系统中某些用户收件箱中未阅读的但已被其他用户界定的垃圾邮件.

文献[10]提出使用模型来描述用户发送邮件的特征. 它统计出每个用户在每个小时段的发送行为(向外发送邮件的数量、附件数量、邮件大小、收件人数量),建立柱状图. 通过将当前阶段的行为特征柱状图与历史行为特征柱状图进行对比分析,来发现异常行为(垃圾邮件).

3　反垃圾邮件技术

3.1　法律手段

为了应对垃圾邮件带来的巨大损失,一些国家制定出相应的法律来规范邮件发送行为,力图减少垃圾邮件的数量. 美国在 2003 年制定出反垃圾邮件法案——非请求色情及广告信息攻击控制法案(controlling the assault of non-solicited pornography and marketing act, CAN-SPAM Act)[6]. 该法案明确禁止邮件头信息伪造、邮件地址骗取和邮件地址攻击等行为. 该法案同时还要求商业性邮件必须有退订链接. 然而,文献[2,37]指出该法案对垃圾邮件数量的控制不具有明显的效果,退订链接的存在反而有助于垃圾邮件制造者确认有效邮件地址.

澳大利亚的电信法案第 107 条,针对个人、公司分别制定了不同的规定[2,38]. 只有得到了收件人的允许,才能向个人发送垃圾邮件(包括商业邮件,以及收件人数超过 50 人的邮件). 而它对发送给公司的邮件的限制要宽松一些,允许向公司发送包含退订链接的垃圾邮件.

欧洲议会在 2002 年 6 月通过了隐私和电子通讯法律规章[3],禁止在未征得收件人同意的情况下,向其发送垃圾邮件.

这些法律条文的制定与实施,能够在一定程度上缓解垃圾邮件问题,然而,这些法律不能彻底杜绝垃圾邮件的产生. 因此,必须将其与其他技术手段相结合,才能更好地过滤垃圾邮件,保障电子邮件通讯的便捷通畅.

3.2　简单方法

在反垃圾邮件研究初期,人们通过对垃圾邮件基本特征和垃圾邮件制造者基本手段的观察,人工制定出一些简单的对策. 这些方法在早期的反垃圾邮件工作中起到了重要的作用.

1) 地址保护.

文献[9]提出一种比较简单的反垃圾邮件技术,通过改变公开的邮件地址形式来防范垃圾邮件. 例如,将邮件地址 username@domain.com 改变为 username#domain.com 或 username AT domain.com 等形式,有时进一步地将“.”改写为 DOT. 这样做可以在一定程度上防止垃圾邮件制造者通过爬虫技术获取网页上的邮件地址.

但是,这种技术的防护能力很弱. 垃圾邮件发送者只要在收录邮件地址时加上一些简单的识别代码,依旧可以提取出真实的邮件地址. 目前通过字典攻击,邮件地址收集程序可以推算出邮件服务器中的账号,还可以提取网上非页面文档(如 DOC、JPEG、PDF、XLS、RTF、PPT 等)中的邮件地址.

2) 关键词过滤.

关键词过滤技术通过检测每封邮件中是否存在预先定义的关键词,例如发票、促销、Viagra 等,来判断邮件的类型[1]. 最初只采用完全匹配的方法,“Viagra”只能与“Viagra”匹配,而不能匹配“Viiaagra”. 这样很容易被垃圾邮件制造者通过小改动,规避这些关键词.

之后,基于正则表达式的模式匹配方法逐渐被采纳. 特定模式“V* i* a* g* r* a”可以与“V-i-agra”、“Viiaagra”、“Viagra”等关键词进行匹配. 这种模式匹配方法能够有效地减小关键词库的大小,并能在一定范围内适应垃圾邮件的小改动.

3) 黑名单和白名单.

这 2 种方法均基于对发件人身份的简单识别,当身份信息被伪造时,这 2 种方法将会失去效用[3].

黑名单方法指的是通过拒绝来自特定 IP 地址、TCP 连接,或域名的邮件,从而过滤掉垃圾邮件发送者发送的垃圾邮件. 但是这些包含在邮件头部中的信息有时会被垃圾邮件发送者伪造成其他人的地址发送,这样会使得无辜的人的电子邮件被过滤掉.

白名单方法指的是只接收来自特定 IP 地址、TCP 连接或域名的邮件,而拒绝其他所有来源的邮件. 白名单方法使用起来不是很方便,2 个人刚开始联系时需要发送请求确认邮件.

4) 灰名单和激励-响应.

灰名单方法会对服务器中未记录的邮件给出暂时失败的响应[10]. 对正常邮件来说,正确配置的 MTA 收到该响应后会再次发送该邮件. 当服务器在一定时间内再次收到该邮件时,会将其成功传送. 而对于垃圾邮件来说,邮件往往是通过开放转发(open-relay)的方式发送,不会因为错误响应而再次被发送,故无法成功到达收件人. 该方式的缺点是会给正常邮件的发送带来少量的延迟.

激励-响应(challenge-response)在白名单的基础上增加了激励响应策略[11]. 该方法同样维护一个白名单列表,来自白名单列表中地址的邮件会被成功发送. 而列表之外的邮件地址进行发信时,服务器

会返回给发件人一个“图灵测试”,如果发件人通过了测试,邮件将会被成功传送,而相应的发件人地址将被添加到白名单列表中. 垃圾邮件制造者一般会采用伪造的发件人地址,来逃避反向追踪,也就收不到返回的测试.

这2种方法的设计基于正常邮件和垃圾邮件发送时所能作出的不同反应,利用垃圾邮件无法正确作出响应的不足,对邮件类型进行判别. 这2种方法的不足是,响应会给正常邮件的发送带来延迟,也会占用网络带宽.

3.3 智能型垃圾邮件检测技术

1) 质朴贝叶斯.

该方法简便、有效,是商业软件中一种最常用的方法. 大量的工作表明这种方法是处理垃圾邮件最有效的方法之一,并且它能够取得较高的精确率(precision)和召回率(recall)[42-43]. 一些研究表明使用多项式模型能够比使用多元伯努利(Bernoulli)模型获得更高的正确率(accuracy)[5]. 在传统的质朴贝叶斯(naïve Bayes)方法之上,衍生出了很多变体. R. Shrestha 等人[44]利用不同位置出现的同一关键字的内部关联特性进行分类,计算关键字的协同权重(co-weighting),并取得了性能上的提高. Li 等人[45]提出了基于用户反馈的改进的 naïve Bayes 方法,获得了相对较低的丢失率(false positive)和较好的性能.

2) k-近邻方法.

Sakkis 等人[46]将 k-近邻方法(一种经典的惰性学习方法)应用于垃圾邮件检测领域. 他们通过实验方法研究了领域大小(k 的大小)、特征维数,以及训练集大小对检测器性能的影响. 文中实验表明,k-近邻方法的平均性能优于贝叶斯方法.

3) Boosting Trees.

Schapire 和 Singer[47]首先将该方法应用于文本分类领域,通过组合多个基本假设(base hypotheses)来处理多类别(multi-class)以及多标签(multi-label)的分类问题. Carreras 和 Marquez[48]实现了 AdaBoost 算法用于反垃圾邮件的邮件过滤,在基于2个公共数据集(PU1 corpus 和 Ling-Spam corpus)实验的基础上,他们得出 Boosting Trees 的方法在性能上要优于 Naïve Bayes、Decision Trees 和 k-NN 算法. 然而,Nicholas[49]认为使用 decision stumps 的 Boosting Tress 以及 AdaBoost 在正确率和速度方面都要差于 Naïve Bayse.

4) 支持向量机.

文献[50-52]中对该方法进行了深入的讨论. Drucker 等人[53]实现了一个基于 SVM 的过滤器,他们的研究表明 SVM 过滤器和 Boosting Trees 过滤器均能够达到最低的错误率(error rates),但是 Boosting Trees 花费了更多的训练时间.

5) Ripper.

和其他分类方法不同,Ripper[54]并不需要特征向量,它从训练样本集中归纳出分类的规则,通过一系列相与或者相或关系的 if-then 规则组成.

6) Rocchio.

这种类型的分类器[55-56]使用规范化的 TF-IDF 来表示训练样本的向量. 这种方法的优点是在训练和测试中具有较快的速度,缺点是在训练集上搜索最优阈值(optimum threshold)以及最优 β 时会消耗掉额外的训练时间,并且这些参数在测试集上的泛化特性也较弱.

7) 文本聚类.

M. Sasaki 等人[57]提出基于特征空间模型的文本聚类方法,使用 spherical k-means 算法[58]来自动计算出不同的 clusters,并对抽取出的质心向量(centroid vector)分配类别标记,通过计算新邮件向量和质心向量的距离来完成分类. 该方法在 Ling-Spam corpus 数据库获得了较好的测试性能.

8) 元启发(Meta-heuristics).

C. Y. Yeh 等人[12]针对关键字变化对基于关键字的机器学习方法所造成的性能上的影响,提出了使用 spammers 的行为作为区分特征,来进行邮件的分类. 这些行为特征通过 Meta-heuristics 来描述,在给定的 Meta-heuristics 下,共抽取出了113个新的特征. 实验结果显示这种方法要优于基于关键字的过滤方式,并且训练时间也有了显著的降低.

9) 人工神经网络.

J. Clark 等人[59]利用人工神经网络自动分类邮件,他们开发的系统 Linger 在 Ling-Spam corpus 数据库获得了较高的正确率、召回率以及精确率. 在 PU1 corpus 上系统所获得的性能略有下降. I. Stuart 等人[60]基于词和消息的描述性特征,使用人工神经网络的方法对邮件进行分类,实验结果表明该方法还需要对特征集作适当地扩充或者修改以获得性能上的提高.

10) 人工免疫系统.

A. Secker 等人[61]提出基于免疫的邮件分类算法 AISEC(artifical immune system for e-mail classification) . 该算法旨在区分出用户感兴趣的邮件和不感兴趣的邮件. 在不需要进行重新训练的前提下,算法能够连续地对 e-mail 进行分类处理,并能够及时地追踪用户兴趣的变化.

T. Oda 等人[62]将人工免疫模型应用于垃圾邮件处理,主要利用免疫中自己/异己(self/non-self) 的检测原理和检测器(dector) 的概念. 在实现的邮件过滤系统中,首先从多样的来源中构建基因库,这些来源包括语言中的词汇、所收集的邮件中的词汇和词组、垃圾邮件中的联系信息和邮件头信息等. 在系统初始化的过程中,使用随机的方法从基因库中生成抗体(antibody) 及其关联的淋巴细胞(lymphocyte) . 在构建的过程中,不允许相似抗体的重复产生,每个淋巴细胞除了具有抗体属性外,还有 msg_matched 和 spam_matched 2 个属性与其关联,分别用于表示淋巴细胞所匹配的邮件的数目和垃圾邮件的数目. 在对淋巴细胞的训练过程中,对发生匹配的淋巴细胞修改其 msg_matched 和 spam_matched 这 2 个属性的值. 在系统的运行过程中,使用了带权平均值的评价方法对邮件的类别进行判断,在这种评价方法下,匹配次数多的淋巴细胞在评分中具有较大的权重.

4　性能评估方法及标准数据集

垃圾邮件检测技术仍是现今国内外研究热点之一,大量的相关工作不断涌现出来. 为了便于人们比较和选择合适的垃圾邮件过滤方法,研究人员提出一些评估标准来对比不同过滤方法、系统的性能[2-43]. 本节主要介绍并分析几种常见的性能评估方法,并给出一些标准数据集.

4.1　性能评估方法

1) 垃圾邮件召回率.

该标准能够度量出被算法模型正确检测、分类的垃圾邮件的比例. 垃圾邮件召回率(spam recall) 高的系统模型能够更好地将垃圾邮件过滤掉,更有效减少垃圾邮件对人们生活的妨碍. 下式给出垃圾邮件召回率的计算方法.

$$R_s = \frac{n_{s\to s}}{n_{s\to s} + n_{s\to l}}.$$

式中: $n_{s\to s}$ 表示被正确分类的垃圾邮件的数量,而 $n_{s\to l}$ 表示垃圾邮件被错误分类为正常邮件的数量.

2) 垃圾邮件精确率 .

该标准评估出系统检测垃圾邮件的精确性: 度量被系统分类为垃圾邮件的邮件中,分类正确的比例. 这个标准另一方面也能够反映出被系统错误分类的正常邮件所占的比例. 系统垃圾邮件精确率(spam precision) 越高,被系统错误分类的正常邮件的数量也就越少. 垃圾邮件精准率计算方法如下所示:

$$P_s = \frac{n_{s\to s}}{n_{s\to s} + n_{l\to s}}.$$

式中: $n_{l\to s}$ 表示正常邮件被错误分类为垃圾邮件的数量.

3) 正常邮件召回率和正常邮件精确率.

由于垃圾邮件检测是关于两类邮件的(正常邮件和垃圾邮件) ,这 2 种标准与垃圾邮件召回率和精准率是对称的,计算式也可以对称地推导出来.

4) 准确率.

该标准能够反映邮件过滤系统的整体性能. 它能够表示被正确分类的邮件(包括正常邮件和垃圾邮件) 的比例,被定义如下:

$$A = \frac{n_{s\to s} + n_{l\to l}}{n_l + n_s}.$$

式中: $n_{l\to l}$ 表示被正确分类的正常邮件的数量,n_l 和 n_s 分别表示正常邮件和垃圾邮件的总体数量.

5) 加权准确率.

研究人员观察得出,正常邮件的丢失(被系统错误过滤掉) 意味着人们会错过生活中的重要信息,比垃圾邮件的错误分类要严重得多. 为了反映出正常邮件的重要性,研究人员在准确率的基础上,定义出如下加权准确率:

$$A_w = \frac{n_{s\to s} + \lambda n_{l\to l}}{\lambda n_l + n_s}.$$

式中: λ 是反映正常邮件重要性的参数,它的值越大,说明正常邮件在该情景下的重要性越强,一般可以取值 9、99 或 999. 若将 λ 赋值为 999,则表明正常邮件在该情景下极为重要. 当 λ 取 1 时,加权准确率与准确率标准等价.

6) F_β 度量.

垃圾邮件召回率与精确率只能分别反映系统的单一方面,不能够反映系统整体的性能. 为了解决这一问题,F_β 度量被定义为这 2 种标准的融合,如下式所示:

$$F_\beta = (1+\beta^2)\frac{R_s P_s}{\beta^2 P_s + R_s}.$$

式中:β 表示精确度的权重,反映精确度相对召回率的重要性. 在大多数研究中β取值1,该情况下,称该标准为 F_1 度量.

4.2 标准数据集

2000 年,Androutsopoulos 等人[3]整理发布了 LingSpam 数据集[4],该数据集是早期的经典邮件分类数据集之一: 该数据集共包含 2 893封邮件,其中正常邮件 2 412封,垃圾邮件比例为 16.63%. 该数据集中的邮件都经过了预处理,所有头信息(标题除外)、HTML 标记均已被去除. 该数据集的不足是,正常邮件的内容大多与语言学话题有关. 用该数据集评估邮件检测系统会带来过于乐观的估计.

2004 年,Androutsopoulos 等人[4]经过收集、整理又发布了 PU 系列经典数据集,该数据集被广泛应用于现今各种垃圾邮件过滤系统的性能评估. PU 系列数据集中包含着 4 个独立的数据集:

1) PU1: 该数据包含1 099封邮件,其中垃圾邮件 481 封. 该数据集中的正常邮件和垃圾邮件均为英语书写的邮件. 正常邮件是文中的第 1 位作者[4]在 36 个月的时间里收集到的,而垃圾邮件是他在 22 个月的时间内收集的.

2) PU2: 该数据集包含 721 封邮件,其中有 142 封垃圾邮件. 与 PU1 相似,该数据集中的邮件也都是英语邮件. 文中作者的一位同事在 22 个月的时间内收集保存了这些邮件.

3) PU3: 该数据集包含4 139封邮件,其中有 1 826封垃圾邮件. 与 PU1、PU2 不同,该数据集同时包含英语邮件和非英语邮件. 数据集中的正常邮件是文中的第 2 位作者收集的,而垃圾邮件来自其他邮件数据集.

4) PUA: 该数据集包含1 142封邮件,其中 572 封垃圾邮件. 与 PU3 相似,该数据集也包含部分非英语邮件,垃圾邮件同样来自其他数据集. 数据集中的正常邮件是文中作者的另一位同事收集提供的.

另外,Medlock[63]也整理发布了一个大规模邮件数据集 GenSpam[64]: 该数据集由 3 部分组成: 训练集(包含8 018封正常邮件,31 235封垃圾邮件)、测试集(包含 754 封正常邮件,797 封垃圾邮件)、自适应集(包含 300 封正常邮件,300 封垃圾邮件,该部分集合用于测试垃圾邮件过滤系统的动态性、自适应性).

ZH1 数据集是中文邮件数据集[65-66],其中的邮件已进行过中文分词处理,处理后的单词被映射为整数,以保护邮件所有者的隐私. 该数据集包含 1 633封邮件,其中正常邮件 428 封,垃圾邮件比例为 73.79%. 数据集中正常邮件平均长度为 819.06 个单词.

5 总结及展望

在现有的反垃圾电子邮件技术方法中,智能型反垃圾邮件技术方法仍然是最有效、最有前景的方法. 法律手段和简单方法只能对部分符合定义特征的垃圾电子邮件起一定作用,且这 2 种方法不具备自适应性,不能有效过滤垃圾邮件的变种. 在智能型反垃圾邮件技术方法中,邮件特征提取方法起着至关重要的作用,将直接影响反垃圾邮件系统的各项性能.

邮件特征提取是反垃圾邮件系统的核心部分,对系统的分类性能起着决定性作用. 目前,绝大多数邮件集中于文本、图片类型. 基于文本、图片的邮件特征提取方法有着良好的应用前景,是当今的研究热点. 基于行为的邮件特征提取方法,是一种与邮件类型无关的特征提取方法,该方法通过区分垃圾邮件、正常邮件发送过程中表现出的不同行为,过滤垃圾邮件,是一种有效、鲁棒性强的方法,非常值得进一步地研究与探讨. 新的邮件特征提取方法的研究,将极大地推进反垃圾邮件系统的发展.

现有的反垃圾邮件相关法律,对垃圾邮件发送行为进行了一定的限制. 然而,现有的相关法律,并不能从根本上解决垃圾邮件问题,需要反垃圾邮件技术的协同支持. 现有的相关法案也急需进一步完善.

智能型反垃圾邮件技术是在简单反垃圾邮件方法的基础上,发展出的新型反垃圾邮件技术. 该技术在提取邮件特征的基础上,运用现代机器学习方法等各种智能方法对邮件类型(是否为垃圾邮件)进行分类,以过滤垃圾邮件. 新的智能型方法的提出及其在反垃圾邮件领域的应用将是未来的研究方向,具有很大的发展前景.

参考文献:

[1] CRANOR L F, LAMACCHIA B A. Spam! [J]. Communications of the ACM, 1998, 41(8): 74-83.

[2] GANSTERER W, ILGER M, LECHNER P, et al. Anti-

spam methods—state-of-the-art [EB/OL]. [2009-11-05]. http://spam.ani.univie.ac.at/files/FA384018-1.pdf.

[3] 中国互联网协会反垃圾邮件中心. 2008 年第一次中国反垃圾邮件状况调查报告 [EB/OL]. [2009-11-05]. http://www.anti-spam.cn/.

[4] Symantec Inc.. The state of spam, a monthly report—February 2009 [EB/OL]. [2009-11-05]. http://eval.symantec.com/mktginfo/enterprise/other_resources/b-state_of_spam_report_02-2009.en-us.pdf.

[5] JENNINGS R. Cost of spam is flattening—our 2009 prediction [EB/OL]. [2009-11-05]. http://www.ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/.

[6] Sophos Inc.. Security threat report, July 2009 update: a look at the challenge ahead [EB/OL]. [2009-11-07]. http://www.inuit.se/pub/1214/sophos-security-threat-report-jul-2009-na-wpus.pdf.

[7] 中国互联网协会反垃圾邮件中心. 2009 年第一季度中国反垃圾邮件状况调查报告 [EB/OL]. [2009-11-07]. http://www.anti-spam.cn/pdf/2009_01_mail_survey.pdf.

[8] 中国互联网协会反垃圾邮件中心. 2008 年第四季度中国反垃圾邮件状况调查报告 [EB/OL]. [2009-11-07]. http://www.anti-spam.cn/pdf/2008_4_dc.pdf.

[9] Wikipedia. Kullback-Leibler divergence [EB/OL]. [2009-11-07]. http://en.wikipedia.org/wiki/Information_gain.

[10] KOPRINSKA I, POON J, CLARK J, et al. Learning to classify e-mail [J]. Information Sciences, 2007, 177: 2167-2187.

[11] YANG Y M, PEDERSEN J O. A comparative study on feature selection in text categorization [C]//Proceedings of International Conference on Machine Learning (ICML'97). San Francisco, USA: Morgan Kaufmann Publishers Inc., 1997: 412-420.

[12] GUZELLA T S, CAMINHAS M. A review of machine learning approaches to spam filtering [J]. Expert Systems with Applications, 2009, 36: 10206-10222.

[13] BLANZIERI E, BRYL A. A survey of learning-based techniques of email spam filtering [EB/OL]. [2009-11-07]. http://eprints.biblio.unitn.it/archive/00001070/.

[14] ANDROUTSOPOULOS I, PALIOURAS G, MICHELAKIS E. Learning to filter unsolicited commercial e-mail, technique report No. 2004/2 [R]. Agia Paraskevi, Greece: NCSR "Demokritos", 2004.

[15] SCHNEIDER K M. A comparison of event models for naive Bayes anti-spam e-mail filtering [C]//Proceedings of the 10th Conference of European Chapter of the Association for Computational Linguistics. Morristown, USA: Association for Computational Linguistics, 2003: 307-314.

[16] YERAZUNIS W S. Sparse binary polynomial hashing and the CRM114 discriminator [EB/OL]. [2009-11-07]. http://fozzolog.fozzilinymoo.org/images/CRM114_slides.pdf.

[17] SIEFKES C, ASSIS F, CHHABRA S, et al. Combining winnow and orthogonal sparse bigrams for incremental spam filtering [C]//Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases. New York, USA: Springer-Verlag, 2004: 410-421.

[18] ODA T, WHITE T. Developing an immunity to spam [J]. Lecture Notes in Computer Science, 2003, 2723: 231-242.

[19] RUAN Guangchen, TAN Ying. A three-layer back-propagation neural network for spam detection using artificial immune concentration [J]. Soft Computing, 2010, 14: 139-150.

[20] KRASSER S, TANG Y C, GOULD J, et al. Identifying image spam based on header and file properties using C4.5 decision trees and support vector machine learning [C]//Proceedings of IEEE SMC Information Assurance and Security Workshop. New York, USA, 2007: 255-261.

[21] NHUNG N P, PHUONG T M. An efficient method for filtering image based spam [J]. Lecture Notes in Computer Science, 2007, 4673: 945-953.

[22] YEH C Y, WU C H, DOONG S H. Effective spam classification based on meta-heuristics [C]//Proceedings of 2005 IEEE International Conference on Systems, Man, and Cybernetics. Waikoloa, HI, USA, 2005: 3872-3877.

[23] TASI C H, WU C H. Learning typed behaviors of spam e-mails using back-propagation neural networks [D]. Kaohsiung, China: Shu-Te University, 2004.

[24] WU C H, TSAI C H. A time-robust spam classifier based on back-propagation neural networks and behavior-based features [C]//Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. Hong Kong, 2007: 19-22.

[25] COSTALES B, ALLMAN E. Sendmail [M]. 3rd ed. Sebastopol, USA: O'Reilly & Associates, Inc., 2002.

[26] LIU M, LI Y C, LI W. Spam filtering by stages [C]//Proceedings of 2007 International Conference on Convergence Information Technology. Washington, DC, USA: IEEE Computer Society, 2007: 2209-2213.

[27] YUE X, ABRAHAM A, CHI Z X, et al. Artificial im-

mune system inspired behavior-based anti-spam filter [J]. Soft Computing, 2007, 11: 729-740.

[28] GUO Y H, ZHANG Y L, LIU J Y, et al. Research on the comprehensive anti-spam filter [C]//Proceedings of IEEE International Conference on Industrial Informatics. Singapore, 2006: 1069-1074.

[29] BHATTACHARYYA M, SCHULTZ M G, ESKIN E, et al. MET: an experimental system for malicious email tracking [C]//Proceedings of the 2002 New Security Paradigms Workshop. Virginia Beach, VA, USA, 2002: 3-10.

[30] HERSHKOP S. Behavior-based email analysis with application to spam detection [D]. New York, USA: Columbia University, 2006.

[31] MARTIN S, SEWANI A, NELSON B, et al. Analyzing behavioral features for email classification [C]//Proceedings of Conference on Email and Anti Spam. Stanford University, USA, 2005.

[32] STOLFO S J, HERSHKOP S, HU C W, et al. Behavior-based modeling and its application to email analysis [J]. ACM Transactions on Internet Technology, 2006, 6(2): 187-221.

[33] BRENDEL R, KRAWCZYK H. Detection methods of dynamic spammers' behavior [C]//Proceedings of 2nd International Conference on Dependability of Computer Systems. Washington, DC, USA: IEEE Computer Society, 2007: 145-152.

[34] RAMACHANDRAN A, FEAMSTER N. Understanding the network-level behavior of spammers [C]//Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. New York, USA: ACM, 2006: 291-302.

[35] **陈建发,吴顺祥. 一种基于用户行为分析的协同反垃圾邮件策略 [J]. 电脑知识与技术: 学术交流,** 2007(7): 36-37.

CHEN Jianfa, WU Shunxiang. An cooperate anti-spam strategy based on user's behavioral analysis [J]. Computer Knowledge and Technology: Academic Exchange, 2007(7): 36-37.

[36] SPAM LAWS. The CAN-SPAM Act of 2003 [EB/OL]. [2009-11-07]. http://www.spamlaws.com/federal/index.shtml.

[37] GRIMES G A. Compliance with CAN-SPAM Act of 2003 [J]. Communications of the ACM, 2007, 50: 55-62.

[38] Rundfunk and Telekom Regulierungs-GmbH. Telekommunikationsgesetz 2003 (TKG 2003) [EB/OL]. [2009-11-07]. http://www.rtr.at/de/tk/TKG2003#p107.

[39] HOANCA B. How good are our weapons in the spam wars? [J]. IEEE Technology and Society Magazine, 2006, 25(1): 22-30.

[40] HARRIS E. The next step in the spam control war: greylisting [EB/OL]. [2009-11-07]. http://projects.puremagic.com/greylisting/whitepaper.html.

[41] LODER T, ALSTYNE M V, WASH R. An economic answer to unsolicited communication [C]//Proceedings of the 5th ACM Conference on Electronic Commerce. New York, USA: ACM, 2004: 40-50.

[42] SAHAMI M, DUMAIS S, HECKERMAN D, et al. A Bayesian approach tofiltering junk e-mail [C]//Proceedings of the 1998 Workshop on Learning for Text Categorization. Madison, USA, 1998: 55-62.

[43] ANDROUTSOPOULOS I, KOUTSIAS J, CHANDRINOS K V, et al. An evaluation of naive Bayesian anti-spam filtering [C]//Proceedings of the Workshop on Machine Learning in the New Information Age. Barcelona, Spain, 2000: 9-17.

[44] SHRESTHA R, LIN Y P. Improved Bayesian spam filtering based on co-weighted multi-area information [J]. Lecture Notes in Computer Science, 2005, 3518: 650-660.

[45] LI Yang, FANG Binxing, GUO Li, et al. Research of a novel anti-spam technique based on users' feedback and improved naive Bayesian approach [C]//Proceedings of the International Conference on Networking and Services. Washington, DC, USA: IEEE Computer Society, 2006: 86.

[46] SAKKIS G, ANDROUTSOPOULOS I, PALIOURAS G, et al. A memory-based approach to anti-spam filtering for mailing lists [J]. Information Retrieval, 2003, 6(1): 49-73.

[47] SCHAPIRE R E, SINGER Y. BoosTexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39(2): 135-168.

[48] CARRERAS X, MARQUEZ L. Boosting trees for anti-spam e-mail filtering [C]//Proceedings of 4th International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria, 2001: 58-64.

[49] NICHOLAS T. Using AdaBoost and decision stumps to identify spam e-mail [EB/OL]. [2009-11-07]. http://nlp.stanford.edu/courses/cs224n/2003/fp/tyronen/report.pdf.

[50] VAPNIK V N. Estimation of dependencies based on empirical data [M]. New York: Springer-Verlag, 1982.

[51] VAPNIK V N. The nature of statistical learning theory [M]. 2nd ed. New York: Springer-Verlag, 2000.

[52] DRUCKER H, BURGES C J C, KAUFFMAN L, et al. Support vector regression machines [C]//Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 1997: 155-161.

[53] DRUCKER H, WU D, VAPNIK V N. Support vector machines for spam categorization [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048-1054.

[54] COHEN W W. Fast effective rule induction [C]//Proceedings of 12th International Conference on Machine Learning. San Mateo, USA: Morgan Kaufmann, 1995: 115-123.

[55] SCHAPIRE R E, SINGER Y, SINGHAL A. Boosting and Rocchio applied to text filtering [C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1998: 215-223.

[56] JOACHIMS T. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization [C]//Proceedings of 14th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufman Publishers Inc., 1997: 143-151.

[57] SASAKI M, SHINNOU H. Spam detection using text clustering [C]//Proceedings of International Conference on Cyberworlds. Washington, DC, USA: IEEE Computer Society, 2005: 316-319.

[58] DHILLON I S, MODHA D S. Concept decompositions for large sparse text data using clustering [J]. Machine Learning, 2001, 42(1/2): 143-175.

[59] CLARK J, KOPRINSKA I, POON J. A neural network based approach to automated e-mail classification [C]//Proceedings of IEEE/WIC International Conference on Web Intelligence. Washington, DC, USA: IEEE Computer Society, 2003: 702.

[60] STUART I, CHA S H, TAPPERT C. A neural network classifier for junk e-mail [J]. Lecture Notes in Computer Science, 2004, 3163: 442-450.

[61] SECKER A, FREITAS A A, TIMMIS J. AISEC: an artificial immune system for e-mail classification [C]//Proceedings of the Congress on Evolutionary Computation. Canberra, Australia, 2003: 131-139.

[62] ODA T, WHITE T. Spam detection using an artificial immune system [EB/OL]. [2009-11-09]. http://terri.zone12.com/doc/academic/crossroads/.

[63] MEDLOCK B. An adaptive, semi-structured language model approach to spam filtering on a new corpus [C]//Proceedings of 3rd Conference on Email and Anti-spam. Mountain View, USA, 2006.

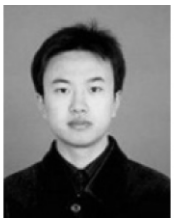
[64] MEDLOCK B. GenSpam [EB/OL]. [2009-11-09]. http://www.benmedlock.co.uk/genspam.html.

[65] ZHANG L, ZHU J, YAO T. An evaluation of statistical spam filtering techniques [J]. ACM Transactions on Asian Language Information Processing, 2004, 3(4): 243-269.

[66] ZHANG L, ZHU J, YAO T. Index of /lzhang10/spam [EB/OL]. [2009-11-09]. http://homepages.inf.ed.ac.uk/lzhang10/spam/.

作者简介:

谭　营,男,1964 年生,教授、博士生导师、博士,IEEE Senior Member. IJSIR 副编辑,IES Journal B, Intelligent Devices and Systems 副编辑,Journal of Computer Science and Systems Biology 副编辑, International Journal of KES 编委, Springer 和多个重要国际期刊的专刊的编辑,ICSI2010 大会主席,ISNN2008 程序委员会主席. 主要研究方向为计算智能、群体智能、智能信息处理、计算机安全、数据挖掘与模式识别等. 负责国家"863"计划、国家自然基金等科研项目 30 余项. 获得 2009 年度国家自然科学奖二等奖,中科院百人计划入选者. 发表学术论文 200 余篇.



朱元春,男,1985 年生,博士研究生,主要研究方向为群体智能、人工免疫系统、智能信息处理算法、计算机安全、模式识别等.