基于行为的垃圾邮件检测技术

秦 逸
(南京大学计算机科学与技术系 南京 210093)

摘 要 电子邮件作为一种重要的信息交互手段,引发了诸如垃圾邮件、恶意邮件、隐私泄露等一系列严重的问题。垃圾邮件检测是近年来出现的一个研究领域,其目的在于判定一封邮件是否为垃圾邮件。提出了一种基于邮件行为的垃圾邮件检测技术BJMD,介绍了邮件行为检测的主要思想和算法过程。通过在实际邮件集上的实验和分析,给出了新方法的一些性能评判。

关键词 垃圾邮件,行为模型,邮件检测,数据挖掘,电子邮件

Behavior-based Junk Mail Detection

QIN Yi
(Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

Abstract As an important information exchange tool, emails contribute to serious problems such as spam-mails, malicious mails and private information leaks. Junk mail detection is a recently emerging filed, which aims to provide a efficient monitoring means on deciding whether a mail is a junk mail. We provided a behavior-based method of junk mail detection. We showed the behavior models used in our method and gave a introduction of the algorithm processes. We implemented our method on a real mail set and gave some comparisons with content-based junk mail detection to get some experiences in mining mail information.

Keywords Junk mail, Behavior model, Mail detection, Data mining, Email

1 简介

电子邮件作为当代社会最重要的通讯手段之一,极大程度地提高了使用者的交流和工作效率。除了日常通信之外,电子邮件还被广泛运用于文件交换领域。正是后一种用途使得不法用户把电子邮件作为传播有害信息、电脑病毒的平台。垃圾邮件是不正当使用电子邮件的一个主要组成,是指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。垃圾邮件对于人们正常使用网络的影响是不容小视的。中国互联网络信息中心(CNNIC)发布的《第十七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85.3封电子邮件,其中垃圾邮件占了57.5封,垃圾邮件数量已经超过了正常邮件数量,这给正常邮件的使用者带来了极大的不便,造成了巨大的经济损失。可以说,垃圾邮件严重困扰着亿万互联网用户的正常邮件通讯,大量占用了互联网上有限的存储、计算和网络资源,降低了网络使用效率,耗费了用户大量的处理时间。而且,垃圾邮件也逐渐成为病毒在互联网上传播的一个主要途径。因此,研究高效的垃圾邮件检测技术具有重要的意义。

目前,对于垃圾邮件的处理主要是基于内容的邮件检测技术。这种技术通常采用数据挖掘技术实现一个针对垃圾邮件的分类器,将其用于垃圾邮件的判定。基于内容的垃圾邮件检测技术的优势在于能够充分利用邮件中含有的所有信息,能够达到较高的检测率和正确率。但是这种方法需要使用大量有代表性的邮件集对分类器进行训练,方法的整体效率不高。同时由于在判定一封邮件为垃圾邮件时,该邮件实际上已经经过了完整的传播过程,因此这类方法还是造成了对网络资源的消耗。

本文提出了一种基于行为的垃圾邮件检测方法BJMD(Behavior-based Junk Mail Detection),其核心思想是使用邮件行为模型来刻画垃圾邮件和一般邮件的区别,使得能够在不使用邮件正文数据的情况下,根据发送行为判定其是否为垃圾邮件。由于不涉及邮件的实质内容检测,因此该方法不要求一定需要在邮件接收端进行,使得垃圾邮件有可能在其完成传播过程前被发现并且拦截,从而节省宝贵的网络带宽资源。同时该方法达到了判定过程与邮件内容的无关性,从而在一定程度上保证了用户的隐私权。

2 相关工作

目前垃圾邮件检测的主流技术是基于内容的垃圾邮件检测,其核心思想是使用数据挖掘手段,从预先获取的训练邮件集中学习垃圾邮件的特征规律,用于指导后续邮件的判断。根据特征规律获得的方式,可以将基于内容的垃圾邮件检测方法分为两类:基于规则的特征规律检测和基于概率统计的特征规律检测。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来推出结果。

到稿日期:2012-02-21 返修日期:2012-05-11

秦 逸(1989—),男,硕博连读生,主要研究方向为分布式计算、普适计算开发环境、人工智能技术。

基于规则的特征规律检测是通过训练得到可理解的垃圾邮件分类规则的逻辑表达式，使得垃圾邮件的检测具备了一定的可理解性。这方面的工作包括 William W. Cohen 提出的基于规则的垃圾邮件检测 Ripper[1]、Carreras 等的基于决策树的垃圾邮件过滤方法[2]、Nicholas 等引入的 Boosting“投票”多分类器检测方法[3]等。

基于概率统计的特征规律检测可以看作是前一种方法的推广，是一种寻找隐式规则的技术。由于不要求分类过程的可理解性，因此这种方法在效率和性能上都要优于基于规则的方法。这方面的主要工作包括使用 SVM(Support Vector Machine，支持向量机)的垃圾邮件检测[4,5]。PAN 等则引入了基于神经网路的 Winnow 算法，使得分类器能随着训练集的增大而快速地调整决策策略[5]。文献[6,7]则使用 Naïve Bayes 方法进行垃圾邮件检测，将邮件词汇特征、词组特征和其他属性特征作为表示邮件的特征集合。文献[15]为邮件建立了基于字符的语言模型，即通过分类器方法实现垃圾邮件的检测。

上述这些方法的成功拦截率和检测正确率都在 80% 以上，能够在测试集上较好地区分一般邮件和垃圾邮件。但是如前文所述，基于内容的垃圾邮件检测存在训练效率的问题。文献[8]指出除了基于 Bayes 的方法外，大部分检测方法的训练时间复杂度都在 $O(N^2)$ 以上，基于 Flexible Bayes 方法和基于 SVM 方法的单封邮件检测时间复杂度也达到了 $O(N)$。此外，基于内容的垃圾邮件检测还存在隐私保护、资源占用等一系列问题。

3 基于行为的垃圾邮件检测方法 BJMD

本文提出的 BJMD 方法是传统垃圾邮件检测的一个拓展，试图将垃圾邮件检测分析从文本信息获取层面上升到邮件发送行为的语义层面，目的是在更高的智能层面上给出对电子邮件分类技术的改进。BJMD 方法实际上是一种行为审计法(behavior audit)，即通过检测电子邮件的行为来确定垃圾邮件。我们事先对各个用户的电子邮件使用习惯进行建模，将垃圾邮件的范围划定在那些有悖于用户日常使用习惯的邮件上。该方法的优点在于：一是用户的隐私性得到了保证；二是预防机制所需的响应时间较少；三是由于垃圾邮件发件方无法得知收件人正常的邮件行为，使得垃圾邮件很难逃避这种检测机制。

3.1 邮件行为检测的背景

BJMD 方法将电子邮件的行为模型引入到垃圾邮件的检测中，而电子邮件的行为模型在恶意邮件(指携带有恶意代码或程序的能够自动复制传播的电子邮件)检测中已经有了一定的应用。使用行为模型对电子邮件进行检测的思想源于入侵检测，已有的电子邮件行为检测模型都是基于 D. E. Denning 的 IDES(Intrusion Detection Expert System)思想。文献[9]首次引入了行为语义的概念，将入侵行为作为一种低概率、非正常的行为加以检测。文献[10]将行为语义检测引入到恶意邮件分析领域，提出了一个有害邮件拦截系统 MEF (Malicious Email Filter)。该系统可以检测已知类型或未知类型的恶意链接，追踪恶意链接的来源并自动改进自己的检测模型以适应用户邮件行为的改变。MEF 主要是侧重于恶意邮件的检测工作，Bhattacharyya 等在其后又提出了 MET (Malicious Email Tracking)系统[11]，其添加了对恶意邮件的追踪溯源功能。在 MET 和 MEF 的基础上，Stolf 等提出了综合性的邮件行为分析工具 EMT[12]。

3.2 BJMD 方法的邮件行为模型

与恶意邮件检测不同，垃圾邮件一般由一台服务器负责进行大量邮件的发送，垃圾邮件自己不会进行复制传播，这导致了垃圾邮件在邮件行为上与恶意邮件存在较大的区别。当然垃圾邮件的邮件行为与一般邮件行为仍存在差异性，这使得能够使用基于行为的检测方法从一般邮件中区分出垃圾邮件。在这里我们的基本假设是与用户通信的所有邮件地址中(包括正常的电子邮件用户和发送垃圾邮件的服务器)，正常的电子邮件使用者的邮件地址占到绝对多数，从而使得我们的模型能够刻画出用户的正常邮件通信行为。

为了能够使 BJMD 方法的模型准确区分垃圾邮件和正常的电子邮件，必须分析垃圾邮件的行为在哪些方面有别于正常的电子邮件。一般认为，垃圾邮件的行为存在下述几个特征：

(1)垃圾邮件的发送行为是由垃圾邮件服务器事先确定的，所以它在相当的时间内都不会改变自身的邮件行为。

(2)垃圾邮件的传播基本上是通过大量的发送量来实现的，这些邮件的行为间存在着高度的一致性，是正常邮件用户所不能达到的。

(3)垃圾邮件对邮件接收者的行为一无所知，也不知道用户与他所联系的人的关系。垃圾邮件通常会违反用户对于某些社交圈内的关系人的通信习惯。

(4)通常情况下，垃圾邮件的行为是单向的，即不会有人回复垃圾邮件。

基于上面的分析，我们对垃圾邮件行为的建模可以从 3 个方面进行。首先建立用户的正常频繁通信群，用于刻画那些与当前用户密切程度相近的邮件地址，一个用户可以同时从属于多个通信群。其次是刻画通信频率，这实际上体现了用户与每一个频繁通信群之间的相关程度。最后出于对用户通信习惯变化的考虑，对通信频率的增量进行描述。在具体的建模方法上，我们采用用户 cliques 建模方法来构建用户的频繁通信群，采用 Hellinger 距离建模方法来构建用户的通信频率，采用增量分布式建模方法来构建用户通信频率变化的情况下其通信习惯的不变性。

用户 cliques[13]建模方法侧重于考察用户习惯与哪些通信人进行集中的通信。该方法将用户的通信人划分为若干个自然的 cliques，划分的依据是用户的历史通信记录，频繁通信的用户群被划分到同一个 clique 内(如同事、家庭成员、社区朋友等)。cliques 不仅提供了关于用户的历史通信记录，同时为判断当前用户和某一社会组织之间的关系提供了一定的条件信息。一般来说，与单个 clique 保持频繁的通信关系通常暗示该用户从属于该 cliques 对应的社会群体，而与多个 cliques 保持频繁通信则意味着该用户有可能与这些团体存在纵向的社会关系。

Hellinger 距离[14]建模方法侧重于考察用户和通信人之间联系的频率以及该频率的变化规律。通常意义下，人们使用电子邮件进行沟通的通信频率与自然语言交流的频率存在一定的可比性，这使得我们可以对这种频率进行建模。通过对通信频率进行分析可以获得通信人在当前通信集中的重要

程度。我们可以通过计算用户常用收件人的"响应比"获得不同通信人的关系地位:那些立即回复的用户通常是该团体中的重要人。

增量分布式模型[13]侧重于刻画用户在与不同通信集中通信人在发送邮件的频率上存在的偏序关系。正常的电子邮件用户与不同通信集的通信习惯存在较大的差异,如前文所述,恶意邮件在传播过程中存在随机型和均匀分布的特性,因此不会符合用户历史通信信息中存在的这种规律,从而使其在向新的受害者传播时将违反用户的正常通信行为。

3.3 BJMD方法的算法过程

BJMD方法检测垃圾邮件的过程分为3个步骤:频繁通信群建立、用户通信动态性考察以及引入知识的检测结果复验。

(1)频繁通信群建立

采用了第3.2节中的用户cliques建模方法,结合Hellinger距离为用户的一般通信行为建模。具体的做法是对用户的日常进行通信的邮件地址进行分类,将其中的地址根据用户的邮件收发行为划分为若干个地址群,地址群的划分是依据用户是否给该群内的用户群发过电子邮件。群内通信模型实际上刻画了用户的一般习惯邮件收发行为,而非平常邮件的发送行为会被系统检测到并交给后续的检测机制进行检测。

(2)用户通信动态性考察

在邮件行为检测中,必须考虑不符合用户一般通信规律的非垃圾邮件出现的可能性。另一方面,用户也有主观改变邮件收发习惯的可能。我们通过对用户通行动态性的考察来判断这类邮件是否有可能是垃圾邮件。具体采用了收件人频率来定量地描述用户的邮件发送行为,使用Chi Square公式和Hellinger距离来判断收件人频率的改变情况。

Chi Square公式如下:

$$Q=\sum_{i=1}^{k}\left(\frac{X(i)-np(i)}{np(i)}\right)$$

式中,$X(i)$为在观察窗内第i个收件人的观测数,$p(i)$为第i个收件人的收件人频率。

Hellinger距离公式如下:

$$HD(f_p[\,],f_t[\,])=\sum_{i=0}^{n-1}\left(\sqrt{f_p[i]}-\sqrt{f_t[i]}\right)^2$$

式中,$f_p[i]$和$f_t[i]$为用户p和用户t的收件人频率。

(3)检测结果复验

通过频繁通信群和用户通信动态性考察,已经可以大致判定哪些邮件有可能是垃圾邮件。为了进一步提高系统的检出率和检测正确率,BJMD方法引入了结果复验机制。第3.2节中的特性告诉我们垃圾邮件的行为存在单向性和时间密集性。在结果复验中,我们对可能的垃圾邮件的前后数封邮件再次进行检测,当其满足第3.2节中的垃圾邮件行为特性时,就能够以较大的概率确定一封邮件是否为垃圾邮件。

4 实验分析

为方便定量地评价垃圾邮件检测的效果,我们引入拦截成功率、拦截错误率和拦截未成功率3个概念。通过概率论可以知道,一种二值预测方法的预测结果与实际值之间的关系可以由表1描述。结合电子邮件行为检测的实际情形,表1中实际值、预测结果的取值解释如下:

(1)实际值True指电子邮件为垃圾邮件;

(2)实际值False指电子邮件不为垃圾邮件;

(3)预测结果True指检测方法判定电子邮件为垃圾邮件;

(4)预测结果False指检测方法判定电子邮件不为垃圾邮件。

表1 二值预测方法预测结果与实际值之间的关系表

预测结果 \ 实际值	True	False
True	True Positive(TP)	True Negative(TN)
False	False Negative(FN)	False Positive(FP)

我们利用表1中的术语进一步给出下面的定义。

定义1 电子邮件行为检测的拦截成功率为:

$P=TP/(TP+FN)$

P的物理含义为检测出的垃圾邮件占总垃圾邮件的比例。

定义2 电子邮件行为监测的拦截错误率为:

$T=FN/(FP+FN)$

T的物理含义为检出的非垃圾邮件中实际为垃圾邮件的比例。

定义3 电子邮件行为检测的拦截未成功率为:

$F=FN/(FN+TP)$

F的物理含义为检测方法未检测出的垃圾邮件占总垃圾邮件的比例。

4.1 实验数据集

选取的实验数据是某高校某实验室17位邮件用户(研究生)在2010年4月25日到2011年3月17号之间的8个时间段(每个时间段长5小时)内接收的344封邮件信息。为了能够使用邮件行为检测,同时获取了这17位邮件用户在同时段内的发件信息。

经过人工筛选,有效邮件一共有331封(无效邮件的情况包括不完整邮件、系统回退邮件),其中包括垃圾邮件127封。该有效邮件集被随机划分为学习集和测试集(4∶1)两个部分。学习集中包含265封邮件,其中有垃圾邮件102封。测试集中包含66封邮件,其中有垃圾邮件25封。

4.2 实验过程与结果分析

分别使用Naïve Bayes(NB)分类器和拓展后的Bayes分类器TAN(Tree-Augmented Naïve Bayes)作为基于内容的垃圾邮件检测方法的代表,与基于行为的垃圾邮件检测BJMD方法进行了比较实验。由于在进行基于邮件行为的垃圾邮件检测时的有效信息仅仅包括邮件头信息,因此我们在使用分类器时使用了3种不同的分类器输入,分别是只使用邮件头信息数据集A,使用正文文本特征信息数据集B和使用正文写作风格特征的数据集C,每个数据集可以进行数据挖掘的数据量是逐渐增大的。

数据集A包含的数据仅限于邮件头信息,包括收件人、发件人、时间、日期、大小、编码方式、主题和内容格式8个数据域。由于收件人信息、发件人信息难以进行离散化处理,实验中没有包含该数据域的分类器。另一方面,由于获取的邮件日期不能够体现用户的一般习惯,因此也没有将该项数据列入分类器的学习范畴。对其它数据进行了如下的预处理:

(1)将时间离散化为时刻;

(2)将主题处理为长度和主体编码方式两个部分;

(3)将邮件大小离散化为[0,1kb]、[1kb,5kb]、[5kb,

10kb]、[10kb,100kb]、[100kb,300kb]、[300kb 以上] 6 个区间；

(4)编码方式和内容格式两项数据本身就是离散化的，因此不需要进行预处理。

在数据集 B 中，我们在数据集 A 的基础上添加了如下 4 个数据域：

(1)内容中是否有 html 语句；

(2)内容中是否有超链接；

(3)内容中是否包含附件；

(4)邮件有无落款或问候语。

而在数据集 C 中，我们又添加了下面 4 项数据：

(1)两个换行符中的平均字符数；

(2)每个句子的平均长度；

(3)每个句子的平均分句书；

(4)邮件正文的段落数。

使用第 3.3 节所述的方法，根据 17 位邮件用户的历史通信记录为其建立了各自的邮件行为模型。

实验结果如表 2 所列。可以看出，在数据集 A 下，由于参与分类的可用数据有限，因此基于内容的分类器算法的优势难以得到发挥。NB 的分类拦截成功率不到 50%，TAN 拦截成功率也仅有 58%。在数据集 B 下，分类器的检测效果出现了一定的提高，两种分类器的拦截成功率均超过了 50%。但是 NB 的误报率仍然很高，这表明当前的数据输入还不能够完全发挥分类器进行垃圾邮件拦截的优势。而在数据集 C 下，两种分类器的检测效果又得到了大幅的提升，拦截成功率达到了 80%的水平，可以认为其基本上能够检测出绝大部分的垃圾邮件。

表 2　基于内容的垃圾邮件检测方法与 BJMD 方法的对比实验

Method	DataSet	TP	FP	TN	FN	P	T	F
NB	A	54.5%	36.4%	43.5%	63.6%	46.1%	63.6%	53.9%
	B	62.1%	43.9%	37.9%	56.1%	52.5%	56.1%	47.5%
	C	77.3%	80.3%	22.7%	19.7%	79.7%	19.7%	20.3%
TAN	A	60.6%	56.1%	39.4%	43.9%	58.0%	43.9%	42.0%
	B	68.2%	60.6%	31.8%	39.4%	63.4%	39.4%	36.6%
	C	84.8%	87.9%	15.2%	12.1%	83.3%	12.1%	16.7%
BJMD		72.7%	81.8%	27.2%	18.2%	80%	18.2%	20%

对于 BJMD 方法，在相同数据的情况下(均为数据集 A)能够达到较高的拦截成功率，基本上与分类算法在数据集 C 下的表现相同。考虑到在构建邮件行为模型所需要的数据量远小于全部的邮件正文数据量，可以认为这种方法在效率上要高于传统的分类器拦截方法。

出于技术融合的考虑，将基于内容的检测方法与基于行为的检测方法相结合。在判断权值为 1∶1 的情况下，实验结果如表 3 所列。

表 3　TAN 分类器与 BJMD 融合检测结果

Method	DataSet	TP	FP	TN	FN	P	T	F
TAN+	A	66.7%	69.7%	33.3%	30.3%	66.7%	30.3%	33.3%
BJMD	C	78.7%	80.3%	21.2%	19.7%	85.6%	19.7%	14.4%

由于在数据集 A 上的 TAN 分类器的性能比较差，因此在加入其进行 1∶1 权值的决策后，整个检测算法的性能出现了较大的下降。而在数据集 C 上，整个检测算法的性能得到了进一步的提升，优于参与决策的两种方法各自的表现。可以看出，这种融合两种检测方法的技术应该能为垃圾邮件检测提供更大的提升空间。

结束语　本文提出了一种基于邮件行为的垃圾邮件检测技术 BJMD，介绍了邮件行为检测的主要思想和算法过程。通过在实际的邮件数据集上的实验和分析，给出 BJMD 方法的一些性能评价。垃圾邮件检测是一项长期处于攻防状态的技术，在垃圾邮件检测技术被不断改进的同时，垃圾邮件制造者也在设计更难以被非智能方法检测出来的垃圾邮件。需要进一步研究垃圾邮件的实时检测技术以及各种方法融合的综合垃圾邮件检测技术等。

参 考 文 献

[1] Cohen W. Fast effective rule induction[C]// Machine Learning: Proceedings of the Twelfth International Conference. Lake Taho, California, Mongan Kanfmann, 1995: 115-123

[2] Carreras X, Marquez L. Boosting Trees for Anti-Spam Email Filtering[C]// Proceedings of Euro Conference Recent Advances in NLP (RANLP-2001). Sep. 2001: 58-64

[3] Nicholas T. Using AdaBoost and Decision Stumps to Identify Spam E-mail. Stanford University Course Project (Spring 2002/2003) Report[OL]. http://nlp.stanford.edu/ courses/cs224n/2003/fp/

[4] Drucker H, Wu D, Vapnik V N. Support Vector Machines for Spam Categorization[J]. IEEE Transactions on Neural Networks, 1999, 20(5): 1048-1054

[5] 董建设，袁占亭，张秋余. 基于多种核函数的 SVM 在垃圾邮件过滤中的应用[J]. 计算机应用，2008，28(2)：424-427

[6] Sahami M, Dumais S, Heckerman D, et al. A Bayesian approach to filtering junk e-mail[C]// Proc. of AAAI Workshop on Learning for Text Categorization. 1998: 55-62

[7] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An Evaluation of Naive Bayesian Anti-Spam Filtering[C]// Proc. of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000). May 2000: 9-17

[8] Androutsopoulos I, Paliouras G, Michelakis E. Learning to Filter Unsolicited Commercial E-Mail[R]. Technical report 2004/2, NCSR "Demokritos". 2004

[9] Denning D E, Neumann P G. Requirements and model for IDES-A real-time intrusion detection system[R]. Tech. Rep., Comput. Sci. Lab, SRI International, Menlo Park, CA, 1985

[10] Schultz M, Eskin E, Zadok E, et al. MEF: Malicious Email Filter A UNIX Mail Filter that Detect Malicious Windows Executables [C]// USENIX Annual Technical Conference-FREENIX Track. 2001

[11] Schultz M, Eskin E, Zadok E, et al. Data Mining Methods for Detection of New Malicious Executables[C]// IEEE Symposium on Security and Privacy. 2001

[12] Bhattacharyya M, Hershkop S, Eskin E. MET: An Experimental System for Malicious Email Tracking[C]// Workshop on New Security Paradigms(NSPW). 2002

[13] Stolfo S, 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

[14] Yang, Lo G, Cam L, et al. Asymptotics in Statistics: Some Basic Concepts[M]. Berlin: Springer, 2000

[15] 苏绥，林鸿飞. 基于字符语言模型的垃圾邮件过滤[J]. 中文信息学报，2009(2)