LAJI YOUJIAN
SHIBIE YU CHULI JISHU YANJIU

垃圾邮件
识别与处理技术研究

李志敏 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

垃圾邮件识别与处理技术研究

A Research of Spam Identification and Its Disposal Method

李志敏 著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

图书在版编目（CIP）数据

垃圾邮件识别与处理技术研究 / 李志敏著. —北京：北京理工大学出版社，2015.12

ISBN 978-7-5682-0660-0

Ⅰ.①垃… Ⅱ.①李… Ⅲ.①电子邮件-安全技术 Ⅳ.①TP393.098

中国版本图书馆 CIP 数据核字（2015）第 225299 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / （010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68948351（其他图书服务热线）
网　　址 / http：//www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 北京泽宇印刷有限公司
开　　本 / 710 毫米×1000 毫米　1/16
印　　张 / 11.5
字　　数 / 205 千字
版　　次 / 2015 年 12 月第 1 版　2015 年 12 月第 1 次印刷
定　　价 / 45.00 元

责任编辑 / 陈莉华
文案编辑 / 陈莉华
责任校对 / 周瑞红
责任印制 / 李志强

前　言

自电子邮件产生以来，它给人类的交流方式带来了革命性的改变。人们可以在任何时间、任何地点接收邮件。然而，电子邮件在给人们带来信息交流便捷的同时，也被大量滥用，导致垃圾邮件的产生。从垃圾邮件诞生之日起，人们便开始遭受各种各样垃圾邮件的困扰，如今垃圾邮件问题愈演愈烈，已经严重影响正常邮件甚至整个网络的运行。清除垃圾邮件浪费了邮件运营商、用户的大量时间和金钱。因此，如何识别和准确过滤垃圾邮件，已成为近几年热门的研究课题。

目前垃圾邮件过滤技术研究集中在基于内容解析和基于行为解析两个方向。基于内容的过滤技术是最传统、最基本的过滤技术；基于行为的识别过滤技术是对传统过滤技术的有效补充。这两类方法均有各自的优缺点，在邮件过滤中起着不同的作用。然而，中文文本、图像信息的垃圾邮件的自身特征还有待探求，使用单一的方法在垃圾邮件处理中效果并不理想，甚至有的问题无法解决，需要改进和创新。在解决实际问题过程中，往往是多种方法配合使用。鉴于此，本书在协作过滤技术和图像型垃圾邮件过滤技术等方面进行比较深入的探究。

本书共分5章。第1章垃圾邮件过滤概述，其内容包括电子邮件系统结构，垃圾邮件定义、特征、历史、现状、分类，以及垃圾邮件过滤技术的发展；第2章基于内容的垃圾邮件过滤技术研究，阐述了基于内容的垃圾邮件过滤技术工作过程，重点论述了分词、文本表示、特征选择、分类的概念及相应理论，并提出相关建议及改进措施。第3章基于邮件行为的垃圾邮件过滤技术研究，针对基于内容过滤技术的局限性，对基于行为识别过滤技术与人工免疫算法进行了研究，并揭示其规律性。考虑实际工作中垃圾邮件过滤的复杂性，采用基于内容或基于行为的单一过滤技术很难解决问题，所以本书安排了第4章、第5章不同于基本过滤技术的创新内容。第4章基于P2P协作的垃圾邮件发送行为识别技术研究，将混合型P2P结构应用于垃圾邮件过滤技术中，设计了基于超节点的Anti－Spam P2P网络结构并研究了该网络的查询策略；研究了垃圾邮件发送行为识别平台，改进了消息查询、转发以及结果返回机制；设计了基于JXTA的垃圾邮件识别

器。第5章图像型垃圾邮件过滤技术研究，详细分析了图像型垃圾邮件的特点，并深入探讨其检测方式，提出了一种改进的文本区域定位算法——ECTL；提出并实现了两种有效的图像型垃圾邮件检测算法。

本书是作者多年来从事计算机网络安全工作的研究成果。在编写过程中得到了南华大学赵治国教师的指导和帮助，也得到湖南机电职业技术学院副院长李玉民教授的指点，在此谨致谢意。

由于作者水平有限，书中存在的缺点和错误，恳请读者批评指正。

作　者

目　录

第 1 章

垃圾邮件过滤概述

伴随着互联网的普及，电子邮件以其快捷、方便、低成本的特点已成为互联网上最重要、最普及的应用之一，但是随之而来的垃圾邮件也越来越泛滥，并占用了有限的存储、计算和网络资源，降低了网络使用效率，影响了互联网的正常使用，干扰了用户的正常工作、生活和学习，耗费了用户大量的处理时间，给全球的邮件用户和邮件运营商带来了不可估量的损失。

1.1 电子邮件简介

电子邮件作为一种信息交流手段，在当今社会的异步通信中扮演着重要的角色，人们通过网络进行信件的收发，不仅可以使对方接收到文字等传统媒体信息，还能接收到图片、声音、影像等多媒体信息。因此在互联网普及的今天，电子邮件已经成为网络中较普遍的一项应用，使人们的交流沟通方式产生了极大的改变。由于邮件系统已经在商业、学校、政府及其他企事业单位等群体中得到了广泛的应用，因而邮件系统的服务质量对人们日常生活有着重要的影响。邮件过滤技术是保证邮件系统高效、稳定运行的重要组成部分，其过滤手段是基于电子邮件的运行原理[1]进行研究与应用的。

1.1.1 电子邮件的结构

电子邮件的结构与普通邮件相类似，主要由两部分构成，即收件人的地址和信件的正文。在电子邮件中，所有的地址信息称为邮件头（Header），而邮件的内容称为邮件体或正文（Body）。在邮件的末尾，还有可选的部分，可以用于进一步注明发件人身份的签名（Signature），在后续扩展的邮件规范中也可以通过

添加附件的形式来利用电子邮件发送多媒体信息。

邮件头由多行文字组成，其内容包括如下部分：收件人，即收信人的 E-mail 地址；抄送，即抄送者的 E-mail 地址；暗送，即在抄送的基础上隐藏抄送人的 E-mail 地址；主题，邮件的主题，一个邮件头部分所包含的内容与 MUA（Mail User Agent，邮件用户代理）有着密切的联系，若 MUA 的功能丰富，所支持的邮件头内容也就较多，在 MUA 发送与接收邮件的过程中还可能加入 MUA 的软件信息，从而使电子邮件支持更多的附加功能。

E-mail 的正文通常是文字表述的信息，在传统的电子邮件中只对 ASCII 码提供支持，因此为保证收件人能够正常地阅读电子邮件的内容，尽量不要加入装饰性及不常用的文字符号。E-mail 签名的位置在邮件的末尾，用以标示邮件发送者的名称信息。由于人们对电子邮件需求的增加，人们对原有的邮件协议进行了扩展，现有的邮件协议支持传递程序、图形及其他一些计算机二进制文件。由于早期的 E-mail 系统只支持文本方式，因此为了使邮件系统具有兼容性，在发送电子邮件时附属的二进制文件需要首先转换为文本文件的形式，收件人在接收后再转换成二进制形式。MIME（Multipurpose Internet Mail Extensions）的应用使得电子邮件中不仅能加入附件，而且也真正实现了在电子邮件发送过程中嵌入视频、音频、图片等多种格式的文件，只要收件人使用与 MIME 兼容的 E-mail 软件，就会自动将附带的多媒体文件进行解码、格式化和演播。

垃圾邮件过滤过程充分利用了邮件结构的特点：对邮件头的发送地址、接收地址及 MUA 附加信息等一系列的内容进行检验，通过垃圾邮件与非垃圾邮件的显著区别特征来进行过滤；在文本类型的邮件中通过对文本信息的处理、挖掘来判定垃圾邮件，对于具有良好隐蔽特性的图片多媒体邮件体采用了光学识别技术；对病毒附件的识别则采用了病毒引擎来进行防护。总之，反垃圾邮件技术抓住了邮件结构的特点，对每种邮件的特征进行了分析，对相应类别的垃圾邮件提出了特定的、有效的解决方案。

1.1.2 电子邮件系统的组成

电子邮件系统由 MTA、MSA、MUA、MDA、MAA 五部分组成，其中 MUA、MTA、MDA 为主要组成部分，另外两部分通常作为附加功能进行实现。

MUA 是一个与用户进行直接交互的邮件系统客户端程序，它通常以 C/S 或 B/S 提供了阅读、发送和接收电子邮件的用户接口。

MSA（Mail Submission Agent，邮件提交代理）在 MUA 与 MTA 之间起到校验的作用。消息发送之前需要完成相应的准备和错误检测，MSA 可以对主机名以及从 MUA 得到的信息头等信息进行检测。

MTA（Mail Transfer Agent，邮件传输代理）负责邮件的中继存储和转发（store and forward），对用户代理的请求进行监视，根据电子邮件发送的目标地址找出对应的邮件服务器，同时将信件在服务器之间传输，在传输过程中将接收到的邮件进行缓存。

MDA（Mail Delivery Agent，邮件投递代理）从 MTA 接收邮件并根据相应的地址进行适当的投递，可以投递到一个本地用户或者邮件列表。

MAA（Mail Access Agent，邮件访问代理）用于将用户连接到邮件服务器的系统邮件库，使用通过相应的协议如 POP 或 IMAP 协议对邮件进行收取。

邮件发送与传递过程中 MTA、MDA、MUA 三个代理承担着全部的功能，电子邮件的发送者首先使用 MUA 对邮件进行编辑，将编辑好的邮件通过 SMTP 协议发送到 MDA，而后 MDA 传递邮件到 MTA，通过网络中若干中继 MTA 传输到邮件接收者用户本地的 MDA，保存在服务器端的用户邮箱中[2]。最后，邮件接收者通过 MUA 相关协议接收本地邮箱中的邮件。需要说明的是，电子邮件传输部分之间的界限并不十分明确，有时一个程序模块可能既包含了 MDA 的功能又包含了 MTA 的功能，而另外一些时候又有可能是 MTA 和 MUA 的功能被组合在一起。

1.2 垃圾邮件的定义

正常邮件与垃圾邮件的区分问题，在互联网上众说纷纭，很多专家与组织都试图给垃圾邮件下一个比较准确的定义。但是，目前国际上对垃圾邮件的认定尚未出台统一标准。

（1）1997 年 10 月 5 日，国际互联网邮件协会召开的主题为《不请自来的大量电子邮件：定义与问题》报告中，就将不请自来的大量电子邮件定义为垃圾邮件[3]，即 UBE（Unsolicited Bulk E-mail）。美国弗吉尼亚州 2003 年《反计算机犯罪法》就采取了“不请自来的大量邮件”来定义垃圾邮件。这是从邮件的发送（大量）和接收（不请自来）这两方面的特征来定义垃圾邮件，更符合垃圾邮件泛滥的实际情况，不但能够涵盖目前泛滥的垃圾邮件的所有类型，也能涵盖未来可能出现的新类型[4]。

（2）2002 年 5 月 20 日，中国教育和科研计算机网公布了《关于制止垃圾邮件的管理规定》，其中对垃圾邮件的定义为：凡是未经用户请求强行发到用户信箱中的任何广告、宣传资料、病毒等内容的电子邮件，一般具有批量发送的特征[5]。

（3）2003 年 2 月 26 日，中国互联网协会颁布的《中国互联网协会反垃圾邮

件规范》中的第三条明确指出，包括下述属性的电子邮件称为垃圾邮件：①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；②收件人无法拒收的电子邮件；③隐藏发件人身份、地址、标题等信息的电子邮件；④含有虚假的信息源、发件人、路由等信息的电子邮件[6]。

目前区分正常邮件与垃圾邮件的一般惯用手段是通过对邮件的内容进行分析，采用人为制定的规则集或机器学习方法来判断、区分。但判定一封邮件是否为垃圾邮件，仅靠分析邮件字面来找出正常邮件与垃圾邮件的区别是很困难的，因为人类语言的种类众多，人对信息的感知与接受除了文字外，还有图形以及对文字本身的联想，所以很难建立一个好的、通用的并且高效的语意分析模型来分析一封电子邮件是不是一封垃圾邮件。另外，人为建立规则集的方式也不具有普遍意义，因为每个人对邮件的感受是千差万别的。所以要快速有效地区分、判定垃圾邮件，需要采取其他更有效的方式。从上述几种垃圾邮件定义来看，不难看出，正常邮件与垃圾邮件的区分就是判断该邮件是不是用户所希望得到而发送过来的邮件，正常邮件自然就是收信人希望得到的邮件。

1.3 垃圾邮件的由来、特征及危害

1. 垃圾邮件的由来

早在互联网的前身——Arpanet 建立之初，人们就已经意识到垃圾邮件过滤问题的存在。1975 年 Jon Postel 在他的文章中已经讨论了垃圾邮件产生的可能性和解决的对策。不过，当时 Arpanet 范围还很有限，仅限于一些大学和研究机构，用户素质普遍较高，因此这篇文章也并没引起足够的重视。1978 年，历史上第一封真正意义上的垃圾邮件出现了。当时的 DEC 公司新推出了一种内置 Arpanet 网络协议支持的计算机，公司的一位营销人员突发奇想，认为直接通过 Arpanet 邮件的方式来推销他们的计算机会是一种很好的方法。于是他和公司里的其他人收集了 Arpanet 用户的地址（当时 Arpanet 上所有的用户信息都是公开的）并将广告信息向所有这些用户发送。由于第一次发送操作出现了失误，导致部分用户没有被发送到，于是他们将信件又重新发送了一次。这次事件在 Arpanet 引起了强烈的负面反响，以至于 Arpanet 的运营者——美国国防通信机构（Defense Communications Agency，DCA）专门致电给 DEC 的老板进行了严厉的谴责，这使得在很长一段时间内没有人再敢尝试类似的行为。

互联网取代 Arpanet 后，首次关于垃圾邮件的记录是 1985 年 8 月一封通过电子邮件发送的连锁信。而历史上比较著名的时间点是 1994 年 4 月 12 日，美国亚

利桑那州两位从事移民签证咨询服务的律师劳伦斯·凯特（Laurence Canter）和玛撒·西格尔（Martha Siegel）把一封宣传“绿卡抽奖”活动的广告信发至6 000多个新闻组。这是互联网上第一次有人大规模地滥发广告邮件。有趣的是，这两位律师在1996年还合作写了一本书——《网络赚钱术》（*How to Make a Fortune on the Internet Superhighway*），书中介绍了他们的这次辉煌经历：通过互联网发布广告信息，只花费了20美元的上网通信费用就吸引来25 000个客户，赚了10万美元。他们认为，通过互联网进行E-mail营销是前所未有、几乎无须任何成本的营销方式。当然，他们并没有考虑别人的感受，也没有计算别人因此遭受的损失。

垃圾邮件开始引起了人们的注意和反感的同时，一些触觉敏锐的商人意识到电子邮件带来的商机。许多人开始利用电子邮件做商业广告，与发送垃圾邮件相关的一些产业也开始出现。1995年5月有人写出了第一个专门的大批量发送电子邮件的程序（名为Floodgate）。紧接着在8月份，有人拿200万个邮件地址出售。垃圾邮件越来越多地与商业联系起来。

凯特和西格尔从网上赚钱之后半年多时间，到了1994年10月27日，网络广告才正式诞生，而目前全球最著名的亚马逊网上商店成立于1995年7月，比“网上赚钱第一人”的诞生要迟15个月。垃圾邮件自诞生以来，尽管遭到了无数人的谴责，但仍有大量的效仿者。出售非法收集的电子邮件地址，尽管被人们所唾弃，却也在悄悄地赚钱。

由此看来，互联网上的赚钱史是从垃圾邮件开始的。垃圾邮件发送者从垃圾邮件中得到的好处是垃圾邮件产生的驱动力。2002年，英国色情垃圾邮件带来了20亿美元的收入。不要认为那些垃圾邮件发送者不挣钱，事实上垃圾邮件的发送已经变成了一种行业，垃圾邮件发送者不断地抓住机会，为自己赚取更多的利益。

2. 垃圾邮件的特征

通过比较Outlook Express、Webmail等标准客户端发出的邮件，我们可以发现垃圾邮件的十大重要特征：

（1）没有X-Mailer信头，或者使用特殊的X-Mailer信头。我们看到Outlook发出的E-mail的信头里面有标明发送客户端的一段X-Mailer：Microsoft Outlook IMO，Build 9.0.2416（9.0.2910.0）。而垃圾邮件或者没有，或者有特征信头，例如，X-Mailer：EhooSend…。

（2）客套话。比较礼貌的垃圾邮件，会带有客套话，例如，“请随手删除”“不好意思打扰了”“打搅……原谅……”。

（3）伪造发件人。MTP 命令里面的 Mail From 和信体里面的 From 不一致，发件人的 E-mail 地址不是真实存在的 E-mail 地址。

（4）通过 ADSL 等动态 IP 发送。外部用户通过动态 IP 直接连接到 MX 服务器，而本地没有提供 SMTP 服务，IP 也不可能进行反向解释。

（5）发送工具签名。一些共享的群发工具会在发出的每一封信里面有自己的签名。

（6）不使用标准的 MIME 格式。邮件格式不标准，但是 Outlook Express 可以兼容。

（7）发送时间超过当前时间。有些垃圾邮件为了在客户端保持一个排列第一的位置，会将发送时间强行修改到一个超前的时间。

（8）经过很多的服务器转发。在信头有三个以上的 Recived。

（9）信体内容带有特殊的 HTML tag。为了嵌入更多的内容和脚本（script），垃圾邮件往往会使用一些正常邮件不会使用的 HTML tag（标签）。如 iframe、frameset、object 等。

（10）信件内容带有取消订阅或者不再接受此类邮件的描述。这往往是发送者更大的陷阱，一旦点击相应链接，就会收到更多的垃圾邮件。

3. 垃圾邮件的危害

（1）占用大量网络带宽，浪费存储空间，影响网络传输和运算速度，造成邮件服务器拥堵，降低了网络的运行效率，严重影响正常的邮件服务。

（2）泛滥成灾的商业性垃圾信件，每 5 个月数量翻倍，国外专家预计每封垃圾邮件所抵消的生产力成本为 1 美元左右。我国开始被其他国家视为垃圾邮件的温床，许多 IP 地址有遭受封杀的危险，长期下去可能会使我国成为“信息孤岛”。

（3）垃圾邮件以其数量多、反复性、强制性、欺骗性、不健康性和传播速度快等特点，严重干扰用户的正常生活，侵犯收件人的隐私权和信箱空间，并耗费收件人的时间、精力和金钱。

（4）垃圾邮件易被黑客利用，危害更大。2002 年 2 月，黑客先侵入并控制了一些高带宽的网站，集中众多服务器的带宽能力，然后用数以亿计的垃圾邮件发动猛烈攻击，造成部分网站瘫痪。

（5）严重影响电子邮件服务商的形象。收到垃圾邮件的用户可能会因为服务商没有建立完善的垃圾邮件过滤机制，而转向其他服务商。

（6）妖言惑众，骗人钱财，传播色情、反动等内容的垃圾邮件，已对现实社会造成严重危害。

1.4 垃圾邮件的现状

毫无疑问，电子邮件已经成为现代生活中使用最多的通信工具之一，几乎每一个电子计算机使用者拥有一个或多个电子邮件信箱。电子邮件已经改变了人们的生活交流方式。正是这样一种免费、方便易用的联系工具，使得人们受到垃圾邮件的严重困扰。普通用户对于每天不请自来的邮箱“垃圾”除了删除还是删除，除了感到厌烦外，还浪费大量的下载时间和带宽。另外，垃圾邮件的泛滥严重损害了电子邮件服务供应商的服务质量和正常业务开展。更严重的是，伴随垃圾邮件传播的计算机病毒、色情和政治反动内容正在造成无法估量的社会影响，影响人们工作、生活、社会稳定与安定团结。

数量巨大的电子垃圾邮件，造成各方面资金上的巨大浪费。据统计，2002年垃圾邮件造成的巨大损失为90亿~100亿美元，2003年成倍增加，2004年超过200亿美元的浪费，这些还不包括垃圾邮件对个人用户所造成的时间和金钱上的损失。

1. 国内情况

自2002年3月以来，作为互联网上最重要的应用之一的电子邮件服务受到国内社会的广泛关注，其原因是互联网用户收到越来越多的垃圾邮件、电子邮件服务商收到越来越多关于垃圾邮件的投诉。垃圾邮件已经成为影响用户对互联网正常使用的障碍，成为一个影响互联网健康发展的负面因素。

为此，国内外的互联网运营商、邮件服务提供商采取了大量的技术和管理手段试图改变这种状态。令人遗憾的是，我们看到，垃圾邮件仍然保持着一种增长的势头。

据中国互联网信息中心（CNNIC）2003年7月底公布的统计数据表明，中国互联网用户户均每周收到垃圾邮件8.9封，正常邮件7.2封，垃圾邮件已经高出正常邮件。

来自中国互联网协会的统计数据显示，2003年，国内的邮件服务器共收到1 500亿封垃圾邮件，尽管其中60%~80%被服务器过滤掉，但至少有470亿封最终流入用户的信箱。

数据还显示，2003年，每个网民平均每天收到1.85封垃圾邮件。为处理这些垃圾邮件，每个网民每天至少需要花费3.65分钟。这意味着，全国网民每年会浪费掉15亿小时的宝贵时间。

中国互联网协会主办的2006年第四次中国反垃圾邮件状况调查结果表明：

我国每年邮件运营商为过滤垃圾邮件的费用投入就有 1.006 7 亿元人民币之多，其中包括硬件、软件、日常运营维护以及人力的投入；普通网民用户为删除这些垃圾邮件每年给国民经济造成约 103.308 3 亿元人民币的损失，由此得出垃圾邮件每年给我国国民经济造成的经济损失约为 104.315 亿元人民币[7]。

2010 年、2011 年平均电子邮箱用户每周收到垃圾邮件率为 40% 以上。

据 2014 年调查显示，近年来垃圾邮件的威胁呈上升趋势，而携带各种钓鱼软件或者木马的病毒邮件更是以每月 16% 的增幅增加，垃圾邮件的病毒率高达 47% 以上。垃圾邮件发送者不断采用新技术以逃避邮件认证服务，威胁着广大邮件用户，尤其是行业用户的网络安全。

由此可以看出，虽然广大邮件服务提供商和互联网运营者采取了各种各样的技术措施，垃圾邮件增长的势头依然旺盛。另据中国互联网协会反垃圾邮件协调小组最近的调查显示，国内拥有邮件服务器的企业普遍受到垃圾邮件的侵扰，有的企业每周收到上万封垃圾邮件，有的企业每年为应付垃圾邮件投入上百万元的费用和大量的人力，给企业造成了沉重的负担。

2. 国外情况

在研究欧洲电子邮件市场的一份报告中，2004 年在欧洲收到的 46% 的电子邮件都是垃圾邮件，2008 年这一比例上升到 71%，在之后 4 年内垃圾邮件给欧洲的企业造成 850 亿欧元以上的损失。

联合国贸易与发展大会（UNCTAD）发布的《2003 年电子商务与发展》报告中称，美国是全球最大的垃圾邮件制造者，该国产生的垃圾邮件数量占到世界总量的一半以上。同时报告称，美国也是垃圾邮件影响最大的国家。

自 2005 年 11 月以来，全球垃圾邮件的数量增长达 222%，2007 年全球用于垃圾邮件防治的费用达到 530 亿英镑以上，而 2006 年为 270 亿英镑。仅美国市场，2007 年防治垃圾邮件的费用就达 180 亿英镑，而两年前为 100 亿英镑。赛门铁克从 2007 年 8 月开始进行数据检测时，全球垃圾邮件约有 30.6% 来自欧洲，而来自美洲地区的比例为 46%。2010 年 12 月份卡巴斯基检测到的垃圾邮件数量同 11 月份相比，下降了 0.6%，但平均占全部邮件总量的 76.8%。垃圾邮件泛滥成了世界性的一个大问题。

1.5 垃圾邮件的分类

垃圾邮件分类有多种方法，为了便于讨论垃圾邮件的防治措施，这里将垃圾邮件分为三类，即中继邮件、低频垃圾邮件、高频垃圾邮件。