▶ 重庆工商大学商务策划学院“商策文库”资助
2012年度重庆市社会科学规划项目（2012YBZX009）资助
电子商务及供应链系统重庆市重点实验室专项基金项目资助

反垃圾邮件信息过滤技术研究

FAN LAJI YOUJIAN XINXI GUOLU JISHU YANJIU

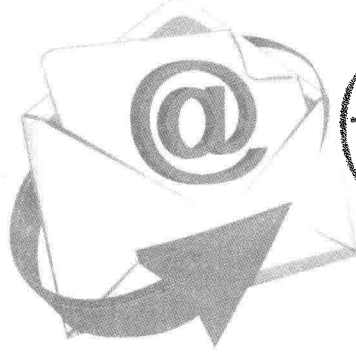
詹　川◇著

电子科技大学出版社

▶ 重庆工商大学商务策划学院“商策文库”资助
2012年度重庆市社会科学规划项目（2012YBZX009）资助
电子商务及供应链系统重庆市重点实验室专项基金项目资助

反垃圾邮件
信息过滤技术研究

FAN LAJI
YOUJIAN
XINXI GUOLU JISHU YANJIU



詹　川◇著

电子科技大学出版社

图书在版编目（CIP）数据

反垃圾邮件信息过滤技术研究 / 詹川著. -- 成都 : 电子科技大学出版社, 2016.5

ISBN 978-7-5647-3598-2

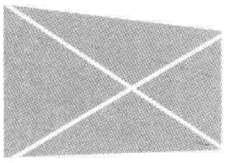
Ⅰ. ①反… Ⅱ. ①詹… Ⅲ. ①反垃圾邮件信息过滤技术研究 Ⅳ. ①TP393.098

中国版本图书馆CIP数据核字（2016）第093148号

反垃圾邮件信息过滤技术研究

詹 川 著

出　　版：电子科技大学出版社（成都市一环路东一段159号电子信息产业大厦）
策划编辑：张　琴　汤云辉
责任编辑：汤云辉
主　　页：www.uestcp.com.cn
电子邮箱：uestcp@uestcp.com.cn
发　　行：新华书店经销
印　　刷：成都蜀通印务有限责任公司
成品尺寸：170mm×240mm　　印张 8　　字数　125　千
版　　次：2016年5月第一版
印　　次：2016年5月第一次印刷
书　　号：ISBN978-7-5647-3598-2
定　　价：32.00元



前言 Qianyan

伴随着 Internet 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。但是随之而来的垃圾邮件也越来越泛滥，占用了有限的存储、计算和网络资源，耗费了用户大量的处理时间，影响和干扰了用户的正常工作、生活和学习。如何有效地治理垃圾邮件问题是全世界共同面临的一道难题，也是互联网上目前急待解决的问题。

本书从技术的角度出发，在全面、系统学习和总结了国内外反垃圾邮件领域的最新成果的基础上，深入、全面地研究了反垃圾邮件信息过滤技术，取得了以下若干创新和成果。

本书的主要创新和贡献包括以下几个方面。

1. 归纳总结了当前垃圾邮件采用的新的抗过滤的方法和手段。垃圾邮件发送者为了让垃圾邮件逃避各种垃圾邮件过滤，不断变化更新欺骗过滤器的方法和手段，目前简单的过滤方法已经无法有效地过滤垃圾邮件。本书在学习了国内外相关资料和研究了大量近期垃圾邮件样本后，归纳总结了当前垃圾邮件发送者常采用的欺骗手段和方法，及其它们的特点，以便有的放矢，更有效地反垃圾邮件。

2. 提出了一种基于内容的 MNNB 垃圾邮件过滤算法。MNNB 算法应用 Markov 链改善了 Naïve Bayes 垃圾邮件过滤算法中的词条之间相互独立的缺陷，并假设句与句之间是独立的，来简化算法的计算量。实验显示 MNNB 算法提高了 Naïve

Bayes 算法的准确率和查全率，并且由于该算法不需要分词，对过滤不同语言的垃圾邮件具有更好的适应性。

3. 提出了一种基于内容的 LVQ 神经网络过滤算法。LVQ 神经网络算法是先把邮件细分成具体的类别，然后再根据用户的定义，把具体的类别规约成垃圾类邮件和正常类邮件。LVQ 神经网络算法克服了垃圾邮件具体类别宽泛、特征离散的问题，提高了垃圾邮件识别的准确度，并且该算法可根据用户对垃圾邮件范围的不同定义，来划分垃圾邮件和正常邮件。

4. 提出了一种基于特征的近似垃圾邮件检测算法——ASD 算法。针对网络中存在大量重复、近似的垃圾邮件，利用 ASD 算法生成的特征，高效地查询收到邮件。ASD 算法以句为单位，作为 SHA1 函数的参数，计算其哈希值，然后将获得的哈希值排序，生成每个已知垃圾邮件的特征。比较新邮件的特征与已知垃圾邮件特征的近似度，来判断该邮件是否为垃圾邮件。

5. 构建了一个基于 URL 垃圾邮件快速过滤的模块。当前相当一部分垃圾邮件简单地给出某“黑网页”的 URL 地址，起到间接宣传广告的作用，而能有效地逃过现有的垃圾邮件过滤方法的过滤。针对此类垃圾邮件，采用基于 URL 的过滤，能有效过滤此类垃圾邮件，是其他垃圾邮件过滤算法的有效补充。

6. 构建了一个基于邮件服务器端的、多层次的垃圾邮件过滤系统——SpamSweeper。SpamSweeper 系统集合了 DNS 反向查询、公有、私有黑白名单、询问 / 响应、基于 URL 的过滤、基于特征的 ASD 算法、基于内容的 LVQ 神经网络算法和 MNNB 算法多种方法，各种方法之间相互协作、互相补充，形成一个准确、快速、高效、易管理和满足不同个性化要求的反垃圾邮件过滤系统。

本书共分八章。

第一章为概述，首先介绍了垃圾邮件的危害性，说明解决治理垃圾邮件问题的迫切性和重要性。然后，介绍了垃圾邮件发展的历史，给出垃圾邮件的定义。接着，从多方面分析了垃圾邮件的组成、用户对垃圾邮件的态度，并介绍各

国垃圾邮件的立法情况。再接着绍了电子邮件的工作原理，分别介绍了 SMTP、POP3、IMAP4、MIME 协议、邮件格式以及 Open Relay 的原理。最后，总结了本书的主要内容和贡献，描述了本书的总体结构。

第二章为反垃圾邮件技术的介绍，本章全面总结分析了当前的各种反垃圾邮件技术，指出各自的优缺点，并详细描述了垃圾邮件发送者针对目前的过滤技术，为逃脱过滤所采用的新方式。

第三章为基于 MNNB 算法的垃圾邮件过滤，本章在 Naïve Bayes 算法的基础上结合 N-gram 和 Markov 理论，提出了 MNNB 算法来进行垃圾邮件过滤。

第四章是基于神经网络 LVQ 的垃圾邮件过滤，我们利用 LVQ 网络的特性，构建了基于 LVQ 神经网络的垃圾邮件过滤模型，来有效地、个性化地对垃圾邮件过滤。

第五章是针对大量重复的垃圾邮件，我们提出基于特征的 ASD 算法来有效地对重复的垃圾邮件过滤。

第六章描述了一个基于 URL 垃圾邮件过滤模块，主要是针对难于从内容上判断的、短小、含有 URL 地址的垃圾邮件。

第七章介绍了我们构建的多层次、多方法、基于服务器端的垃圾邮件过滤系统 SpamSweeper。

第八章对本书的工作进行了总结，并对今后的研究工作进行了展望。

本成果得到重庆市高等学校“特色专业、特色学科、特色学校”项目建设计划中市场营销特色专业建设计划、市场营销国家级特色建设计划、工商管理特色学科专业群建设计划、重庆市高校市级重点学科——工商管理学科建设计划资助，是重庆工商大学商务策划学院“商策文库”成果之一。

目录 Mulu

第一章 绪 论

1.1 研究背景

随着 Internet 的普及，电子邮件由于其方便、快捷、低成本等特点逐渐取代了传统的通信方式，成为现代社会主要通信方式之一和互联网上最重要、最普及的应用之一，大大方便了人们生活、工作和学习。据估计，2003 年全世界电子邮件数量达到 4420 亿封。而近年来，一些公司、团体或个人为了商业利益或政治目的，在未经邮件用户同意的情况下，利用电子邮件发送大量商业广告以及各种不良信息，形成影响极坏、后果严重的垃圾邮件问题。垃圾邮件的泛滥不仅极大地浪费了网络资源，占用了用户的电子邮箱空间，降低了网络使用效率，影响了互联网的正常使用，侵犯了用户的个人权利，甚至还影响到青少年的健康成长。

垃圾邮件目前已经成为世界各国共同面临的棘手问题。据来自 Ferric 调研公司 2003 年发布的一份调研报告，量化了垃圾邮件每年给企业带来的损失，美国企业在此方面损失 89 亿美元；欧洲企业损失 25 亿美元；另外美国和欧洲的互联网服务提供商每年也因此损失 5 亿美元。AOL、雅虎以及 Hotmail 等 ISP 所处理的邮件中，垃圾邮件的数量已经过半。据英国贸易工业部官员称，垃圾邮件现在占到全球电子邮件流量的 40%。更有甚者，据韩国信息保护振兴院统计，韩国国内电子邮件 80% 为垃圾邮件，其中 60% 含有淫秽内容[1]。垃圾邮件的肆虐使得电子邮件系统受到严重挑战，许多网民准备减少或放弃电子邮件这一现代通信方式，来避免垃圾邮件的骚扰。

在中国，据中国互联网络信息中心 2004 年 1 月公布的第十三次《中国互联

网络发展状况统计报告》显示，中国网民平均每周收到 13.7 封电子邮件，其中垃圾邮件有 7.9 封，垃圾邮件数量超过了正常邮件数量占 57.6%[2]。中国是仅次于美国，受垃圾邮件危害最严重的国家。中国邮件用户 2003 年收到的垃圾邮件总数为 460 亿封，处理垃圾邮件浪费的时间为 15 亿小时，2003 年垃圾邮件浪费中国的 GDP 高达 48 亿元[3]。中国互联网协会秘书处处长李欲晓介绍，2001 年 8 月，英国网络公司 uxn.com 宣布，将屏蔽来自中国的绝大多数电子邮件，其黑名单包括了中国电信大部分的 IP 地址以及来自新浪、网易、搜狐、163 邮箱、263、21cn 等 84 个网站的电子邮件。理由是：这些地址是日益增长的、来自中国垃圾邮件的来源。2003 年 7 月，中国互联网反垃圾邮件小组收到美国反垃圾邮件组织的信件，告之他们收到的境外垃圾邮件中有 80% 来自中国，他们“将拔掉连接中国的插头”。中国的垃圾邮件问题已经发展到了必须引起高度重视的程度，如不尽快妥善解决，将不仅有损我国的国际形象，成为其他国家进行封杀的对象，还严重影响我国产业的发展，对我国实施以信息化带动工业化、促进经济发展的政策方针产生极其不利的影响。

如果中国不能及时、有效解决和处理日益泛滥的垃圾邮件问题，将会给我国带来多方面的危害。

首先，垃圾邮件会带来网络和信息安全隐患，扰乱社会秩序。境内外敌对势力通过垃圾邮件传播各种反动信息，对我国的政治和社会稳定构成很大的威胁；部分不法分子利用发送垃圾邮件的方式散布各类虚假广告，或者从事国家明令禁止的传销行为等，严重扰乱了市场经济秩序；在发送垃圾邮件之前，发送人会通过各种不正当的途径获得接收人的邮箱地址和其他个人信息，这也必然会对收件人的个人隐私构成严重的侵犯。

其次，严重危害我国互联网的发展。垃圾邮件占用了大量的传输、存储和运算资源，不但造成网络资源浪费，而且一旦垃圾邮件占到互联网总数据流量的近三分之一，就会造成巨大的存储需求；垃圾邮件还损害了 ISP 的市场形象，造成

无形资产流失；国外邮件服务商封杀中国邮件服务器 IP 地址一事，就致使中国用户蒙受了不可估量的损失。

再次，垃圾邮件损害了用户的利益。垃圾邮件具有数量多、反复性、强制性、欺骗性、不健康性和传播速度快等特点，严重干扰了用户个人的正常生活，浪费用户的时间、精力和金钱。

可见研究反垃圾邮件的方法有着深远的社会意义和巨大的经济价值。互联网社会为了有一个清洁的环境，单位企业为了减少损失、提高效率，用户为了远离垃圾邮件的骚扰都急迫需要有效的反垃圾邮件方法。

1.2 垃圾邮件的历史

垃圾邮件是互联网的副产品，起源于美国。垃圾邮件问题真正得到人们关注是在 1994 年美国的“绿卡事件”后。

Canter 和 Siegel 是两位从事“绿卡”方面的律师，为申请“绿卡”的用户提供咨询。起初他们把广告贴到几个新闻组，宣传可以为用户代写绿卡申请信，每封收费 100 美元。这个信息对绝大多数美国人是没有用的。在 1994 年 4 月 12 日他们雇用了一个程序员写了一个脚本把广告贴到 USENET 的每一个独立新闻组，USENET 是当时最大的在线会议系统，具有几千个新闻组。成千上万收到该信的用户对该事件进行了投诉，表示了不满。然而当时 Canter 和 Siegel 所用网络的 ISP 却无法有效过滤这些垃圾邮件，后来只好停止了这两人的账户。Canter 和 Siegel 又通过别的 ISP 如法炮制他们的广告活动。后来他们俩还出了一本叫“如何在信息高速公路上发财”的书，详细介绍了如何在新闻组上收集邮件地址、如何大量发送广告邮件。在经历“绿卡”事件后，人们开始用 spam 一词来描述在新闻组对同一内容多次重复的滥用行为，现在 spam 已成为垃圾邮件的专用名词。

绿卡事件以后，许多人效仿 Canter 和 Siegel 的做法，收取客户少量的费用，提供邮件广告服务，甚至包括一些法律严厉禁止的商业活动。Jeff Slaton 首次在广告邮件中使用假邮件地址和域名来防止真实身份被查出，并给出了自己的联系方式。他的行为激怒了网络保卫者，他们发布了第一份黑名单，把 Jeff Slaton 的联系方式和个人介绍公布在网上。1996 年，英文中出现 UBE（Unsolicited bulk email）或 UCE（Unsolicited commercial email）两词，分别描述统称的垃圾邮件和商业性质的垃圾邮件。随着垃圾邮件问题不断发展，市场上存在很多可以从网络上收集邮件地址、大量发送邮件的专业软件，同时也出现了一些反垃圾邮件的组织和站点，如著名的 Spamhaus[1]，以及专门从法律角度与垃圾邮件斗争的 CAUCE 组织[2]。1998 年 4 月 Internet 协会 ISOC（Internet society）召开了会议专门讨论垃圾邮件。1992 年 2 月发布了 RFC2502，Anti-Spam Recommendations for SMTP MTAs，标志着垃圾邮件正式成为 Internet 关注的问题。

1.3 垃圾邮件的定义

垃圾邮件一般指的是大量未经用户许可，但却被强行塞入用户邮箱的电子邮件。对垃圾邮件世界上没有一个统一明确的定义，存在多种定义。中国互联网协会在 2003 年 3 月制定的反垃圾邮件规范中，它给出了一个明确的垃圾邮件的范畴。以下四种情况属于垃圾邮件：

1. 收件人事先没有提出要求或者同意接受的广告、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。

1 http://www.spamhaus.org

2 http://www.cauce.org

垃圾邮件的常见内容包括：赚钱信息、成人广告、商业或个人网站广告、电子杂志、连环信等。

垃圾邮件一般具有以下特性：同一内容多次重复发送；同一发件人特定时间段非正常通信；不合法的地址；来自国际公开 RBL 列表的 IP 请求。

在许多文献中把通过邮件四处散发的病毒邮件也称为垃圾邮件，我们没有把这种病毒邮件归入本书研究的垃圾邮件范围之内。因为我们认为虽然病毒邮件与垃圾邮件都具有大批量的向未知用户发送的特点，但是垃圾邮件在内容上具有强烈的广告宣传性，一般不带附件，而病毒邮件几乎没有任何内容，通常都带有病毒的附件。这种病毒邮件通过基于内容的过滤很难识别，但是通过对邮件附件进行病毒特征扫描，能有效地检测出该类邮件。我们觉得把病毒邮件归为病毒来处理更加合理。

1.4 垃圾邮件的组成

联合国贸易与发展会议（UNCTAD）发表的 2003 年电子商务与发展报告中统计了世界垃圾邮件来源，如图 1-1 所示，美国是全球最大的垃圾邮件制造者，全球网民收到的垃圾邮件有一半多源自美国，同时它也是最大的受害国，而中国是第二大垃圾邮件发送国，占全部的 5.6%。中国排第二主要是因为互联网在中国得到飞速发展，而相关的反垃圾邮件法律和有效的反垃圾邮件技术措施没有及时跟上，以致垃圾邮件泛滥。由于中国的垃圾邮件泛滥，曾经出现国外集体封杀中国的邮件服务器，拒绝接收来自中国的电子邮件，把中国的邮件服务器列在黑名单中，使中国成为信息“孤岛”的惨痛教训。因此对于中国来说，反垃圾邮件的重要性不言而喻，而这个任务艰巨，需要全社会、政府、单位和个人共同努力。英国、巴西和加拿大分别依次排在后面，各占 4% ~ 5% 左右。从总体来说，垃圾邮件主要来自母语是英语的国家。

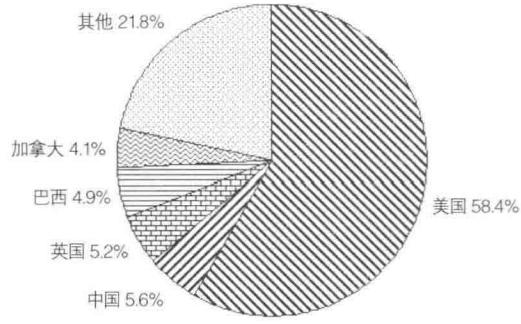


图 1-1　垃圾邮件的来源国

根据上海艾瑞市场咨询公司调研统计中国垃圾邮件的状况，从用户收到的垃圾邮件内容来看，主要由网上购物、IT 产品推销、网上赚钱、情趣用品、订房 / 订票 / 旅游、政治相关、销售商业数据、色情暴力及其他类别组成，如图 1-2 所示。其中网上购物、IT 产品推销和网上赚钱占前三位，分别占 18.6%、13.4%、12.2%；接着是情趣用品、订房 / 订票 / 旅游、政治相关、销售商业数据、色情暴力，它们比例相差不大，在 8%~10% 左右。从内容来看其中大部分垃圾邮件主要是广告性质的邮件。

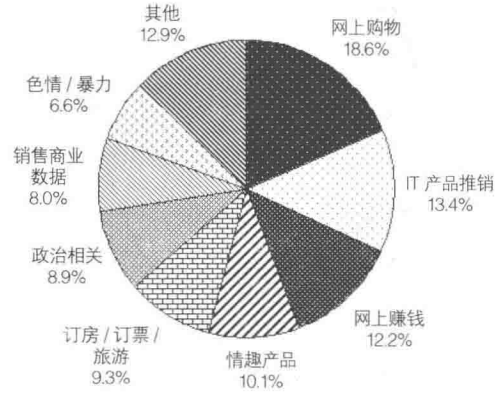


图 1-2　垃圾邮件的类别

对于中国网民来说，从收到的垃圾邮件语言种类来看，中文和英文垃圾邮件占全部的 98%，其他语言的垃圾邮件对于中国网民来说，收到的相当少，可

以忽略不计，如图 1-3 所示。同时也说明垃圾邮件发送者有明确的目的性，从语言上表现出强烈的区域性。在中英文垃圾邮件中，分别以中文简体和英文为主体，占 45.8% 和 41%。因此在中国反垃圾邮件具有鲜明的特色，反英文垃圾邮件跟反中文垃圾邮件同等重要。对于基于内容的垃圾邮件过滤系统，要求对中英文语言都有很好的过滤效果。

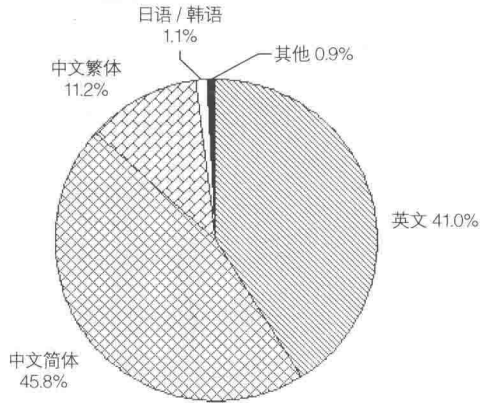


图 1-3 垃圾邮件的语言类别组成

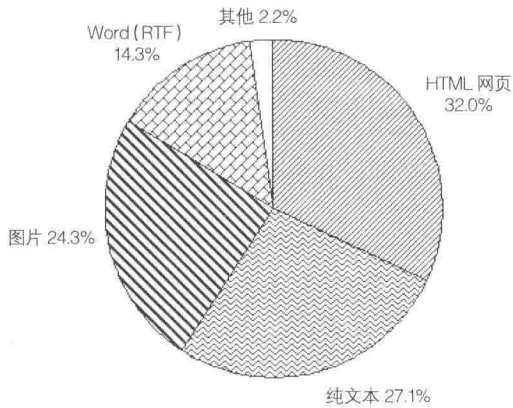


图 1-4 垃圾邮件的文件格式

从垃圾邮件文件格式来看，垃圾邮件中采用最多的是 HTML 网页，占 32.0%，而且这个趋势还在不断增加，如图 1-4 所示。因为 HTML 网页格式具

有更加丰富的表现形式，可以在其中嵌入声音、图片、动画等。垃圾邮件发送者喜欢用这种方式，更重要的原因是采用 HTML 格式可以更容易地、有效地欺骗垃圾邮件过滤器的检查。其次是纯文本占 27.1%。排在第三位的是图片格式，占 24.3%，它将生成的文档以图片格式发送，比如境外反动的法轮功组织经常使用该方式向国内网民发送反动的垃圾邮件。目前还没有有效的过滤方法从图片内容来辨别该图片是否为垃圾邮件，它涉及图像处理方面的难题。Word 格式的垃圾邮件占 14.3%，排在第四位。以上 4 种格式占全部的 97.8%，因此在垃圾邮件过滤前，特别是基于内容的垃圾邮件过滤，需要垃圾邮件预处理，剔除不同格式造成文本不同。

1.5 邮件用户与垃圾邮件

iUserSurvey 对中国邮件用户抽样调查发现，从 2003 年 11 月到 2004 年 3 月垃圾邮件占所收邮件的比例从 26.3% 迅速升到 60.5%，翻了一倍多；95% 的邮件用户收到过垃圾邮件，平均每周 19.3 封；平均每人每周要花费 9.6 分钟处理垃圾邮件。这些数据说明垃圾邮件问题已经广泛影响到中国邮件用户的正常活动，是目前不可回避的社会问题。

调查邮件用户对垃圾邮件的态度，有 95.6% 的用户是持反感态度，其中 57.7% 的用户明确表示不愿接到垃圾邮件。而对于不同类别的垃圾邮件，用户表现出不同的态度，如图 1-5 所示。用户最讨厌的两类分别为色情相关和情趣用品，反映出用户对此类邮件厌烦程度很高。有 1% 的用户愿意接收垃圾邮件，3.4% 的用户对垃圾邮件持无所谓的态度，广告类的垃圾邮件受用户讨厌的比例相对较小，说明某类垃圾邮件对一定的用户具有可用性。因此我们不能笼统地认为垃圾邮件是有害的，一个好的垃圾邮件过滤系统，应该满足用户对垃圾邮件个性化定义的需求。

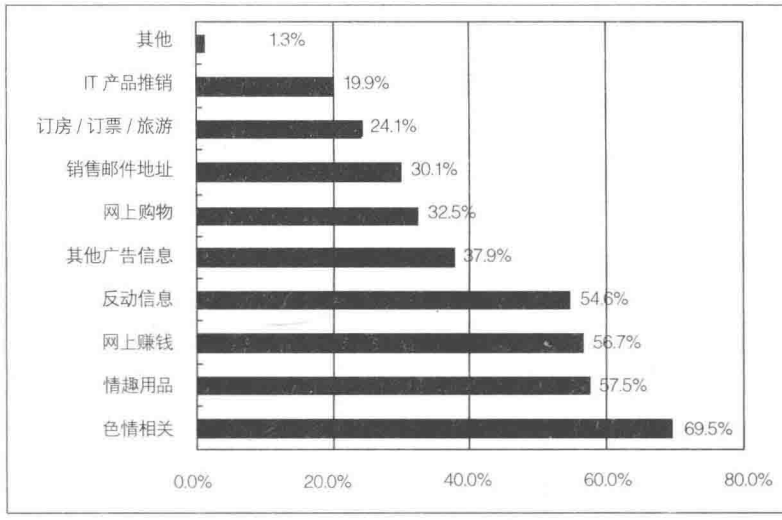


图 1-5 用户讨厌的垃圾邮件类别比例

用户收到垃圾邮件后，有 69% 的人会马上删除，30.2% 的人会在有空时一起删除。从图 1-6 可知，有 78% 的用户对垃圾邮件的内容很少查看或只查看一部分，只有 12% 的用户从不查看，还有 9.8% 是几乎都看。而近 50% 的用户或多或少的从垃圾邮件中获得过信息。这也就是为什么垃圾邮件发送者在知道垃圾邮件令人厌烦的情况下，仍然大量发送垃圾邮件。只要其中少量的邮件逃过邮件系统过滤，被用户查看，就能实现他们的广告效益，这也再次说明垃圾邮件对部分用户具有一定的吸引力和有用性。

对于如何有效地防止垃圾邮件，用户认为最有效的方法是：服务商采取技术手段、不随便公开自己的邮箱地址和建立反垃圾邮件法。如图 1-7 所示，自身防范意识只是起到辅助作用，立法则是对垃圾邮件处理上的法律依据和保证，关键还是在于技术上。因为即使你有很好的防范意识，但是垃圾邮件发送者仍然可以通过字典攻击，将垃圾邮件任意地发送到用户邮箱中。在已经立法的国家，由于商业利益或政治动机，仍然有不少违法者大肆发送垃圾邮件。对于 ISP 服务商来说，采用有效的技术手段防止垃圾邮件是当前国内治理垃圾邮件主要的方法。