第四章 电子邮件内容安全

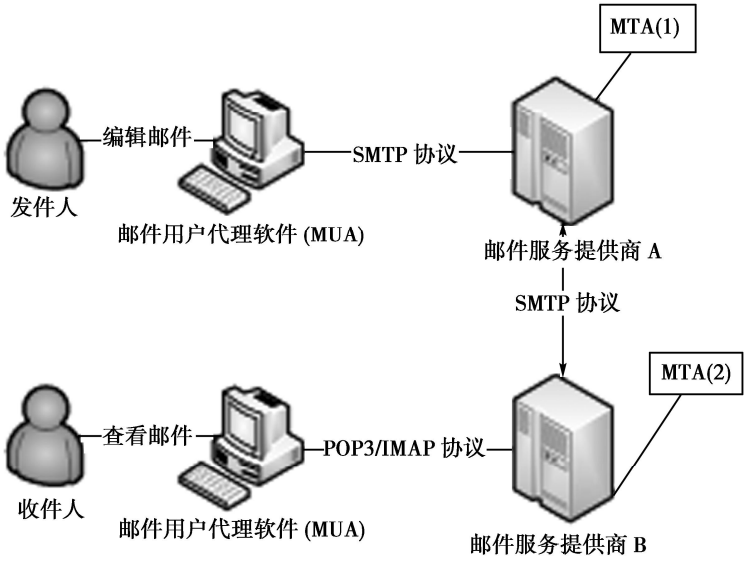
哈爾濱工程大學
HARBIN ENGINEERING UNIVERSITY

4.1 电子邮件概论

- 电子邮件是 Internet 上渗透最广泛、最受欢迎、认知度最高的应用之一
- 通过网络电子邮件系统，用户可以用非常低廉的价格、以非常快速的方式与世界上任何一个角落的网络用户联系。这些电子邮件可以是文字、图像、声音等各种方式。

4.1.1 电子邮件的通信原理

- 电子邮件的发送包括三个重要的组件，即邮件用户代理 MUA、邮件传输代理 MTA和邮件投递代理 MDA。



- 电子邮件不是一种“端到端”的服务，它利用了存储转发的机制。

4.1.2 电子邮件的格式标准

- RFC822协议电子邮件格式
- 电子邮件内容分为基本的两部分：

信头（Header）和信体（Body）。信头由一系列的字段（Fields）组成。信体是发送给收件者的数据（包括文本或文件），可以包含20多个不同的字段，但是并不是所有字段都是必需的。

一个空行（由回车符和换行符组成）将信头与信体分开，也就是说空行标记了信头的结束、信体的开始。

```
From: sender@example.com
To: receiver@example.com
Subject: test email
Date: Sun, 1 Apr 2012 00:00:00 +0800
Message-ID: 4EE645750A9014AF2751912E

Hello, World!
```

4.1.2 电子邮件的格式标准

信头字段	含义	信头字段	含义
From	邮件作者	Subject	**主题**
Sender	发件人	Comments	**备注信息**
Reply-To	回复地址	Keywords	**关键字（用于搜索）**
To	收件人	In-Reply-To	**被当前邮件回复的邮件ID**
CC	抄送地址	References	**基本等同** In-Reply-To
BCC	密送地址	Encrypted	**邮件加密类型**
Message-Id	**邮件的唯一标识**	Date	**发送日期和时间**

4.1.2 电子邮件的格式标准

- MIME 协议邮件格式
- 随着电子邮件的广泛使用，邮件系统不仅需要传输各种字符集的文本内容，而且还需要传送各种非文本文件
- MIME 扩展了 RFC 822 标准，使得二进制数据能够直接合并到一个标准的 RFC822 消息中，为此增加了五种新的信头字段

4.1.2 电子邮件的格式标准

信头字段	字段说明
MIME-Version	发送方用来对消息进行编码的 MIME 的版本
Content-Type	标识了 MIME 消息中封装数据的类型信息
Content-Transfer-Encoding	嵌入的二进制数据编码方式，RFC2045 指定了 5 种方法：7bit（标准的 ASCII 编码）、8bit、binary、Quoted-Printable、Base64。其中最常用的是 Base64 编码，将 3 个字节的二进制数据编码为属于 ASCII 字符集的 4 个字节
Content-Description	用于在邮件消息的文本中标识数据的 ASCII 描述
Content-ID	用来在使用多目录内容的情况下，以一个唯一的标识代码去标识一个 MIME 会话

4.1.3 电子邮件传输协议

- 电子邮件传输协议是由若干 RFC文档规定的。

◆SMTP 协议

- SMTP 协议，是互联网上传输电子邮件的标准协议，用于提交和传送电子邮件，规定了主机之间传输电子邮件的标准交换格式和邮件在链路层上的传输机制。

◆POP3 协议

- POP3 协议是互联网上传输电子邮件的第一个标准协议。它提供信息存储功能，负责为用户保存收到的电子邮件，并从服务器下载取回这些邮件。POP3 为客户机提供了用户名和口令，规范了对电子邮件的访问。

◆IMAP 协议

- 使用IMAP协议（目前已经使用第4版），用户可以有选择地下载电子邮件，甚至只下载部分邮件

4.1.4 电子邮件的内容安全

- 电子邮件的内容安全，可能出于不同的动机或目的：有的可能是好奇或恶作剧，有的出于商业目的，有的则别有用心。其中有些已经属于犯罪行为。
- 垃圾邮件（spam mail）又称UBE（unsolicited bulk e-mail），即未经接受者同意而大量散发的电子邮件。

4.2.1 垃圾邮件的特性

（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；

（5）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。

4.2.2 垃圾邮件产生的原因

- （1）SMTP 协议自身存在的缺陷

SMTP 协议建立在收发邮件双方互相信任的基础上，假定人们的身份和他们所声称的一致。因此，SMTP 协议并没有包含要求用户进行身份认证的内容，所以任何用户都可以使用服务器发送邮件；而且，SMTP 协议也没有规定如何对邮件头中所填写的发件人地址和回复地址作合法性检验。

- （2）商业原因

垃圾邮件一直以来都被认为是最经济有效的广告形式，是开拓市场的有力工具，电子邮件的低成本、高产出、覆盖范围广、发送不受限制、追查难度大等因素使得许多不法的商业分子有机可乘。

4.2.3 垃圾邮件的危害

（1）占用网络带宽，浪费网络资源，干扰邮件系统的正常运行。

（2）浪费用户的时间和上网费用。

（3）对网络安全形成威胁。

4.2.4 垃圾邮件发送手段分析

（1）对邮件内容及发件人信息进行伪装，吸引收件人点击查看。

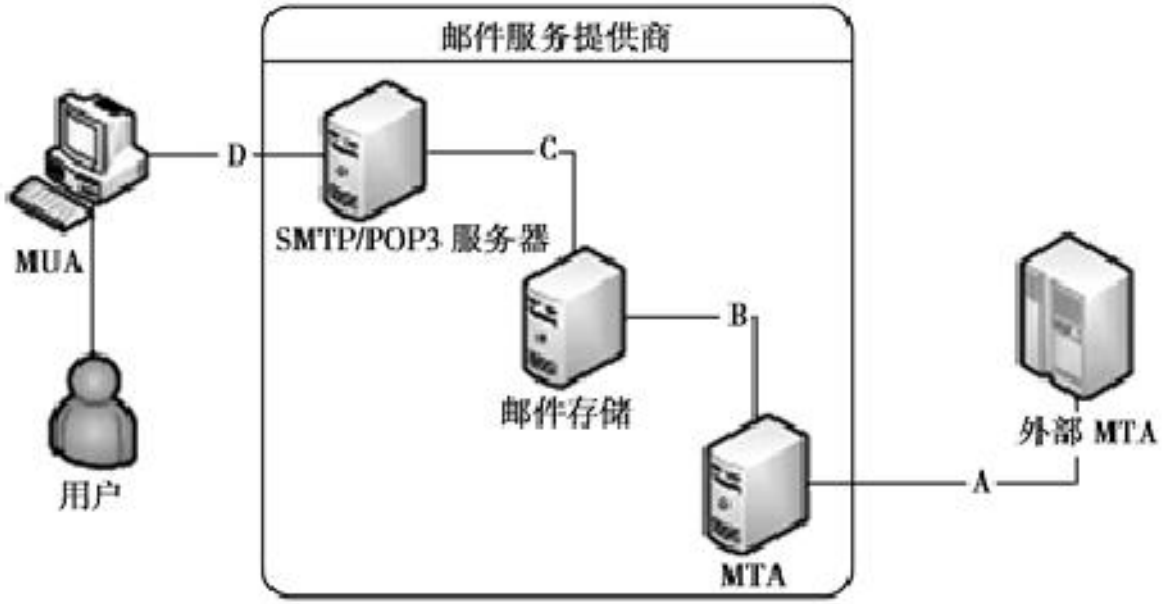
（2）以图片代替文字内容，将要传送的内容以图片的形式附在邮件中。

（3）内容加噪，文本信息中被掺杂了大量“噪声”来干扰反垃圾邮件系统

（4）采用动态或伪装 IP 甚至受病毒感染的“僵尸网络”来发送垃圾邮件，以躲避反垃圾邮件策略中对来自相同大量发送邮件行为的统计和分析。

4.2.5 反垃圾邮件技术

- 邮件服务器可以从逻辑上分为以下几个模块，即：MTA（邮件传输代理）、MDA（邮件投递代理）、邮箱（邮件存储库）、POP3/IMAP 服务器（邮件接收服务器）



4.2.5 反垃圾邮件技术

- 垃圾邮件过滤技术的发展主要经历了以下阶段：

1. 规则过滤、地址列表和统计过滤技术
2. 行为识别模式
3. 电子邮件认证技术

4.3.1 垃圾邮件的特征分析

- 垃圾邮件过滤技术需要通过对电子邮件行为或内容的分析，提取出具备垃圾邮件特性的特征，对符合该特征的邮件进行过滤和拦截。

◆1. 通信特征

◆2. 信头特征

◆3. 信体特征

4.3.2 垃圾邮件的预处理技术

- 邮件预处理技术包括邮件分词、邮件表示和特征选择技术三个部分。
  - 1. 邮件分词
  - 2. 邮件表示
  - 3. 特征选择
    - （1）建立训练邮件样本集；
    - （2）对每封邮件建立特征向量
    - （3）对每个特征计算互信息量
    - （4）定义特征向量。

4.4 垃圾邮件的过滤技术

- 根据过滤技术实施位置的不同，可以将垃圾邮件过滤技术分为客户端邮件过滤和服务器端邮件过滤技术
- 在依据位置的垃圾邮件过滤方式划分中，较为理想的过滤方式是基于服务器端的过滤，这不仅可以使用户免受垃圾邮件的骚扰，而且本地主机也能减少对邮件的处理量，节约处理器资源和带宽流量。

4.4.1 基于黑白名单的过滤技术

- 1. 用户黑白名单技术

使用用户黑白名单机制可以快速准确地过滤掉垃圾邮件，而且可以把经常错误过滤为垃圾邮件的用户邮件快速分辨出来，减少误判率。

- 2. 网络黑白名单技术

优点：减少用户的工作量和设置难度，降低一定的误报率；

缺点：有的RBL提供方提供的黑名单过于强硬，范围过广。

- 3. 分布式自适应黑名单技术

垃圾邮件是大量重复发送的，服务器上会有大量相同的邮件，而正常邮件包含相同内容的可能性很小。因此，分布式自适应黑名单技术基于这一点来区分垃圾邮件。

4.4.2 基于关键字的过滤技术

- 根据电子邮件的信头及内容区域查找邮件中是否包含关键字库中的关键字。
  - 优点：简单直接地进行过滤。
  - 缺点：容易出现误判。为了保证有效，管理员必须经常维护更新关键字库。垃圾邮件制造者也可以通过同音、拆字、生僻字或者将文字制作成图片来避开系统的过滤。

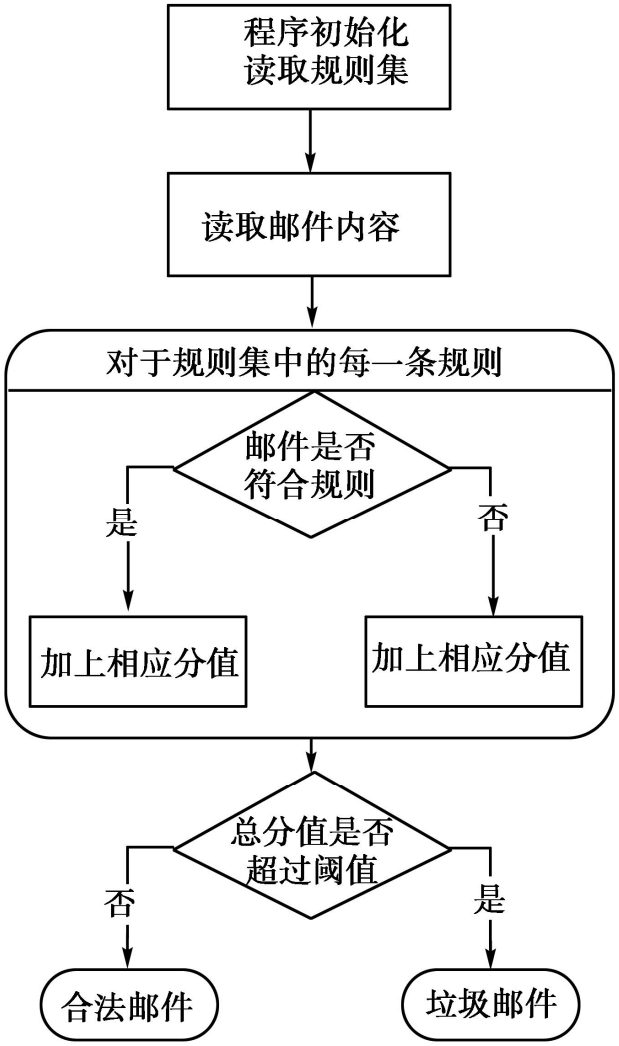
4.4.3 基于统计的内容过滤技术

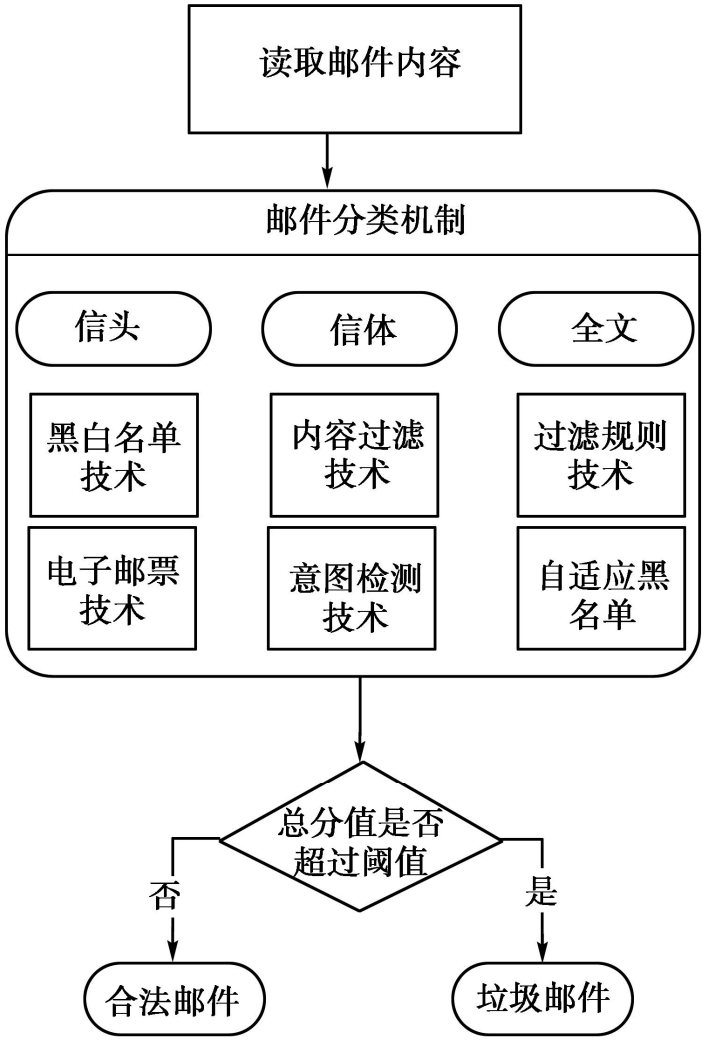
- 贝叶斯过滤技术的工作流程包括两个阶段：学习阶段和判别阶段。
- 基于统计的贝叶斯过滤技术可以在实用的过程中不断地自我学习，系统的特征库会随着已知邮件内容的变化而逐渐更新，不需要复杂的配置就可以自适应地进行过滤工作。
- 技术优点：动态，智能，时效性强，自适应性好，精度高。
- 技术缺点：需要用户干预，判别速度较慢，复杂度较高。

4.4.4 基于规则的内容过滤技术

- 基于规则内容过滤技术是通过检查邮件特征规则来进行垃圾邮件的判断。通过这些特征积累出一系列的判断规则，然后通过设定好的规则匹配一封新邮件是否为垃圾邮件。
  - 优点：方便，容易调整；
  - 缺点：规则需要不断进行更新。

4.4.4 基于规则的内容过滤技术





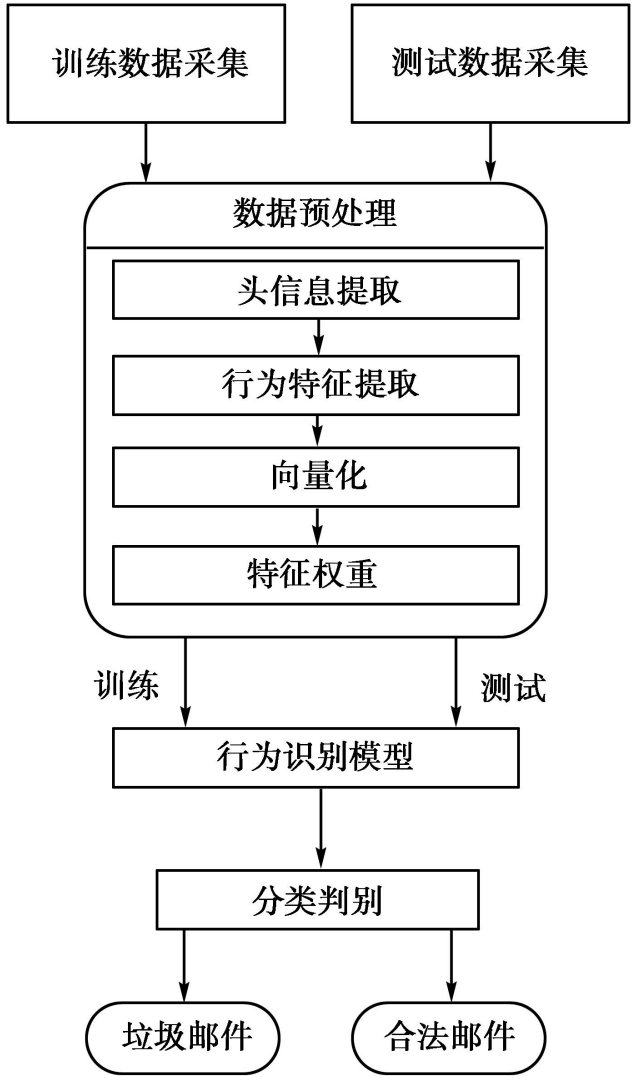
4.4.5 基于行为识别的过滤技术

- 常见的垃圾邮件发送行为有四种：
  （1）邮件滥发行为
  （2）邮件非法行为
  （3）邮件匿名行为
  （4）邮件伪造行为

- 正确判别垃圾邮件的关键问题在于对邮件发送过程中的通信信息进行识别。

4.4.5 基于行为识别的过滤技术

1. 数据采集
2. 数据预处理
3. 建立行为识别模型
4. 测试



4.4.6 图片垃圾邮件的过滤技术

- (1) 文本过滤方法。
- (2) OCR方法。

- (3) 图像属性分析法。
- (4) 图像内容分析法。

4.4.7 基于过滤器反垃圾邮件局限性

1．过滤器可能会被绕过。

2．误报问题。

3．过滤器复查。

实际上，垃圾邮件过滤器技术还有待继续发展和升级。在多数案例中，垃圾邮件依然存在，依然穿过了网络，并且依然被传播。

防止垃圾邮件，必须结合当前多种邮件过滤技术，从服务器端、客户端以及网关等多方面入手，采取层层过滤的方法。