邮件解析介绍

黄一夫

Email

电子邮件，是一种从一个作者到一个或者多个接收者的数字信息交换手段。现今的电子邮件系统是以存储并转发模型为基础的。电子邮件服务器接收，转发，投递并且存储消息。用户和他们的电脑都不需要同时在线，他们只需要简单地连接到一个电子邮件服务器，就可以收发消息。电子邮件从最初仅文本格式（7-bit ASCII and others）的通信媒介，经过了从 RFC2045 到 RFC2049 的标准化过程，已经扩展到现在能够携带多媒体内容附件的电子邮件。这些 RFC 标准总称为多用途互联网邮件扩展（MIME）。

互联网电子邮件信息格式定义在 RFC5322 中，包含两个主要的部分：

一．邮件头——结构化的字段例如 From，To，CC，Subject，Date 和其他关于电子邮件的信息。

1. From：电子邮件地址，和可选的作者的名字。在很多的电子邮件客户端都是不可变的，除非通过改变帐号设置。
2. Date：这封消息被写时的当地时间和日期。和 From 字段相同，很多电子邮件客户端将会自动地填充它。接收客户端可能格式化地显示它，并且将其转换为接收者的时区的表达。
3. Message-ID：也是自动生成的字段，用于防止重复投递并且被 In-Reply-To 字段所引用。
4. In-Reply-To：回复信息的 Message-ID。用作将信息链接到一起。这个字段只用于回复消息。
5. To：电子邮件地址，一个或者多个，和可选的消息接收者的名字，用于指出首要的接收者。
6. Subject：一个简短的消息主题的总结，某些缩写经常用于主题中，包括”RE:”,”FW:”。
7. Bcc：添加到 SMTP 投递表单但是通常不列在消息数据中，对其他的接收者保持不可见。
8. Cc：指出那些将接收到信息的拷贝的邮件地址，许多的电子邮件客户端会依靠你是在 To 字段还是在 Cc 字段中出现来不同地标记收件箱中的邮件。
9. Content-Type：关于消息如何将被显示的信息，通常是 MIME 型。
10. Precedence：通常有’’bulk”，”junk”，或者”list”的值，用作指出自动的”vaction”或者”out of office”，响应不应该被返回到这封邮件。
11. Received：记录被邮件服务器生成的以前的消息，顺序相反。
12. References：正在回复的邮件的 Message-ID 和之前回复的邮件的 Message-ID。
13. Reply-To：应该被用于回复消息的地址。
14. Sender：实际发送者的地址，代表在 From 中列出的地址。
15. Archived-At：一个直接的关于个人电子邮件信息的的归档格式链接。

二．邮件体——基本的无结构化的文本内容，有时在结尾包含一个签名块。这和通常信件的体是相同的。

1. Content encoding

   电子邮件最初被设计为 7-bit ASCII。许多电子邮件软件是 8-bit clean，但必须承担起它能与 7-bit 服务器和邮件读者沟通的职责。MIME 标准出台字符集符和两个内容传输编码，使非 ASCII 数据传输：

quoted printable 的大多是 7 位的内容与范围之外的字符；
base64 针对任意的二进制数据。

2. Plain text and HTML
   许多现代的图形化电子邮件客户端允许使用 plain text 或者 HTML 对信息的主体进行用户定制。HTML 电子邮件信息经常也包含一份自动生成的 plain text 拷贝，基于对兼容性的需求。为使 HTML 在电子邮件中合适地传送，一个额外的邮件头必须在发送的时候被指定： “Content-type:text/html”，大多数的电子邮件程序都自动地发送该邮件头。
3. Content—Type
   MIME 中的邮件体部分比早期的单文本文件复杂很多，实现时要根据其 Content—Type 类型决定具体访问方法。Content—Type 一般包括文本的类型、文本使用的字符集等，文本为复合类型时还包括分隔不同部分的分界字符串(boundary)。RFC2046 定义 Content—Type 顶层有 5 种离散类型和两种复合类型。离散类型是 text(文本信息)，image(图像信息)，audio(音频信息)，video(视频信息)，application(其它信息)。复合类型是 multipart(多部分信息)和 message(压缩信息)，顶层类型又有其子类型。正常邮件中，正文文本多是 text 类型，其它类型大多作为附件方式存在。由于邮件中可能多种类型并存，且不同复合类型会相互嵌套，造成了邮件体分析时的复杂情况。

邮件头和邮件体由一行空格隔开。字段名和值严格使用 7-bit ASCII 字母，Non-ASCII 值可能使用 MIME 编码进行表示。

接收电子邮件后，电子邮件客户端应用程序保存消息到操作系统中的文件系统中。一些客户端作为单独的文件保存单个消息，而其他客户端使用往往专有的不同的数据库格式。所使用的特定格式通常是表示由特殊的文件扩展名指出。
.eml 被很多的邮件客户端，包括 Microsoft Outlook Express，Windows Mail 和 Mozilla Thunderbird 所使用。文件是在 MIME 格式下的 plain text，包括电子邮件头和信息内容，还有在一种或者多种格式的附件。EML 格式是微软公司在 Outlook 中所使用的一种遵循 RFC822 及其后续扩展的文件格式，并成为各类电子邮件软件的通用格式。

Javamail

Javamail 是一种 Java API，用于通过 SMTP,POP3 和 IMAP 接收和发送消息。JavaMail，顾名思义，提供给 Java 开发者处理电子邮件相关的编程接口。它是 Sun 发布的用来处理 email 的 API。它可以方便地执行一些常用的邮件传输。我们可以基于 JavaMail 开发出类似于 Microsoft Outlook 的应用程序。虽然 JavaMail 是 Sun 的 API 之一，但它目前还没有被加在标准的 java 开发工具包中（Java Development Kit），这就意味着在使用前必须另外下载 JavaMail 文件。除此以外，你还需要有 Sun 的 JavaBeans Activation Framework（JAF）。JavaBeans Activation Framework 的运行很复杂，在这里简单的说就是 JavaMail 的运行必须得依赖于它的支持。
JavaMail 包中用于处理电子邮件的核心类是：
Session,Message,Address,Authenticator,Transport,Store,Folder 等。解析电子邮件主

要用到了前三个类。
使用 javamail 解析的大致流程：

1. 构建出解析 email 的类，以类 MimeMessage 为内核进行封装。
2. 从需求出发，封装类 MimeMessage 的成员方法，并且构建出新的方法。
3. 调用设计好的方法获得电子邮件的字段，内容和附件，分别插入数据库或者写到本地文件中，同时要采用相应的解码方式进行解码。

主要的成员方法有：getSendDate()，getFrom1()，getMailAddress()，getSubject()等

jConnect

jConnect for JDBC 是 Sybase 对 Java JDBC 标准的补充。它提供在多层和异构的环境中 Java 开发者对本地数据库存取的需求。jConnect 提供高质量的本地数据库存取并且对所有的 Sybase 产品都支持
用法：

1. 注册 jConect-6.0 驱动
   DriverManager.registerDriver((Driver)Class.forName("com.sybase.jdbc3.jdbc.SybDriver").newInstance());
2. 从相应的 url 中获取连接
   Connection con = DriverManager.getConnection("jdbc:sybase:Tds:localhost:2638?charset=eucgb","DBA","sql");