基于中文域名的邮件系统研究

The Email System and its Implementation Based on Chinese Domain Name

张京鹏　（中国科学院研究生院 北京 100049）
胡安廷　（中国互联网协会 北京 100031）

摘要：本文提出将中文域名邮件系统分为邮件应用层、邮件表示层和邮件寻址投递层的 3 层结构模型。集中阐述了邮件应用层应具有的功能特性，重点讨论了邮件表示层中实现中文邮件地址而使用的 punycode 编码方案及其相关情况，分析并实现了邮件寻址投递层中中文电子邮件地址的处理。

关键词：中文域名　Punycode　邮件系统　中文邮件地址

1　引言

根据中国互联网络信息中心（CNNIC）发布的《第 19 次中国互联网络发展状况统计报告》显示，截至 2006 年底，中国网民人数达到了 1.37 亿。作为互联网上最重要的应用之一，电子邮件系统却没有得到很好的中文化，这一现象极大的制约了网民对于互联网的应用。因而中文域名下邮件系统的设计与实现对于进一步提高我国互联网应用状况具有非常重要的现实意义。

基于这种情况，本文对基于中文域名的邮件系统进行分层研究，提出将整个邮件系统划分为邮件应用层、邮件表示层和邮件寻址投递层的 3 层结构框架，并初步实现了其中的部分层次结构。

2　基于中文域名邮件系统的分层结构

邮件系统的传统研究主要侧重于从邮件系统的功能出发，将邮件系统划分为邮件用户代理 MUA（Mail User Agent）、邮件转发代理 MTA（Mail Transfer Agent）和邮件分发代理 MDA（Mail Delivery Agent），以及一系列电子邮件协议集，包括邮件传输代理协议（MTA Protocols）、邮件用户代理协议（MUA Protocols）和 MIME 协议等。然而，随着人们对于邮件系统新需求的不断提出，尤其是对于基于中文域名的邮件系统的渴望，原有基于功能划分的方式已经不能适应为了解决新问题而进行的设计与实现。为了在中文域名邮件系统设计与实现中更好的模块化邮件系统，本文对邮件系统提出了新的 3 层结构模型的划分方法，这 3 层是邮件应用层、邮件表示层和邮件寻址投递层，相关关系如图 1 所示。

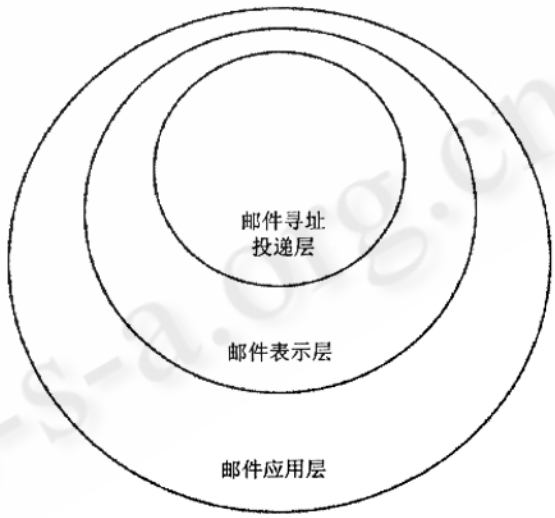

图 1　邮件系统的分层结构

2.1　邮件应用层

邮件系统中的邮件应用层，主要指用户使用邮件系统的交互界面，相应的 API 接口和一系列实用工具。该层主要实现如下功能。

（1）为用户提供实用程序操作界面或者程序接口，方便用户进行邮件的发送、查询和处理。

（2）当邮件系统出现问题时，向用户回复出错信息。

本层包容了传统的邮件用户代理 MUA（Mail User

Agent)，并使得功能目标更为明确，与实际的通信协议层次划分更为接近，更有利于明确其主要的任务是电子邮件协议族中与用户使用相关内容的运用与实现。

2.2 邮件表示层

邮件系统处理的核心是电子邮件。有关电子邮件格式[1]，请参阅文献[1]。该层主要用于协助邮件应用层使用一致的表示形式与邮件寻址投递层交换电子邮件报文。建立该层的目的是为了更好的解决现有邮件寻址投递层只支持 ASCII 字符集电子邮件问题，因为该问题使电子邮件的使用受到极大限制，同时也局限了邮件系统的功能拓展。

本层涉及的主要对象是电子邮件报文本身。由于在现有网络系统上进行任何对现有邮件系统的大规模改造和升级的前提是，必须保证互联网上最重要的服务之一——电子邮件系统——仍然正常地运行，且不给用户造成较大影响。为了升级现有邮件系统，并完成新系统对旧系统的兼容，方便用户使用新系统和未来对新系统的其他扩展，特提出增加邮件表示层。该层主要实现的功能是：

(1) 将用户通过邮件应用层传递的电子邮件按照相应的协议和相关电子邮件格式的要求编码，以格式化电子邮件报文。

(2) 将从邮件寻址定位层得到的电子邮件报文反向格式化，以便用户使用。

(3) 当电子邮件格式有误时，及时向各个相关层报告出错信息。

电子邮件报文主要由信封和信体组成，在信封上使用电子邮件地址来指明邮件投递的具体地址。虽然为解决扩展邮件寻址投递层仅支持 ASCII 字符集的问题，IETF 定义了多用途 Internet 邮件扩充协议 MIME (Multipurpose Internet Mail Extension)[2][3]，但是该协议没有对于中文邮件地址提出相应的解决方案，因而目前邮件地址仍然只支持 ASCII 字符集。

本层是实现基于中文域名的电子邮件系统的关键层，目前针对中文邮件地址的处理是在该层上采用 Punycode 编码[4]的解决方案，并在此基础上实现基于中文域名的电子邮件系统。

2.3 邮件寻址投递层

该层主要用于实现邮件按照电子邮件地址寻找到相应邮件服务器，并且将邮件投递到相应用户邮箱的功能。其涉及的内容包括邮件到达目的地的寻址定位与邮件路由选择、邮件投递方式选择等。主要的对象包括邮件转发代理 MTA (Mail Transfer Agent) 和邮件分发代理 MDA (Mail Delivery Agent)、邮件存储以及电子邮件协议中有关邮件寻址和投递等相关内容。

3 中文邮件地址解决方案

正如前文所述，目前的电子邮件系统仍然高效地工作着，虽然，IETF 定义了多用途 Internet 邮件扩充协议 MIME (Multipurpose Internet Mail Extension)，使得电子邮件的表现方式更加丰富多样，但是现有的邮件系统不支持基于中文邮件地址的电子邮件，无法实现电子邮件的真正中文化与本地化。在中国，绝大多数人们的英语水平有限，这就使得网民上网使用电子邮件系统存在极大的语言障碍。为了解决这一难题，通过在邮件表示层内对中文邮件地址进行转换，是很好地实现与现有系统兼容的有效方法。

中文电子邮件地址的编码方案是基于中文域名的邮件系统 3 层结构模型中邮件表示层的核心，也是实现基于中文域名的邮件系统的关键。

3.1 中文邮件地址的组成

电子邮件与普通的信件一样，需要地址信息。电子邮件地址作为多语种邮件地址 (Internationalized Mail Addresses, IMA) 的中文邮件地址由 3 部分组成，分别是多语种本地名称部分 (Internationalized Local Part, ILP 这里指中文本地部分例如用户名)、"@"符号和多语种域名部分 (Internationalized Domain Name, IDN 这里指中文域名)，按此顺序组成的序列其语法定义如表 1 所示。:

3.2 对"@"符号的处理

"@"符号是电子邮件地址的分隔符，由于中文简体编码标准 GB2312 及其扩展的 GBK 中的"@"符号与英文"@"符号不同，从而对于中文邮件地址的切分造成了一定的障碍。因而需要对"@"符号单独处理，其相关的对应编码见表 2。

目前较为现实的实现方案是将邮件应用层提交过来的中文邮件地址首先进行 Unicode 编码。并按照表 2 中的"@"的各种 Unicode 编码值进行查找并替换成英文"@"。同时将中文邮件地址分隔成 3 个部分，即中文本地名称部分 (Internationalized Local Part, ILP)、

“@”符号和中文域名部分(Internationalized Domain Name, IDN),并组成一个 Unicode 编码的邮件地址串,为下一步处理做准备。

表 1　中文邮件地址语法定义

```
Mailbox = Local-part "@" Domain
Local-part = Dot-string / Quoted-string; MAY be case-sensitive
Dot-string = Atom *("." Atom)
Atom = 1*Ucharacter
Ucharacter = atext / UTF8-2 / UTF8-3 / UTF8-4
Quoted-string = DQUOTE *qcontent DQUOTE
Domain = (sub-domain 1*("." sub-domain)) / address-literal
sub-domain = ULet-dig [ULdh-str]
ULet-dig = Let-dig / Non-ASCII
ULdh-str = *( ALPHA / DIGIT / "-" / Non-ASCII) ULet-dig
Non-ASCII = UTF8-2 / UTF8-3 / UTF8-4
```

表 2　中英文“@”符号 Unicode 编码

	英文“@”	中文“@”
Unicode 编码	U+0040	U+ff20

3.3 Unicode 中文域名的处理

现有域名体系使用的是支持 ASCII 字符集的 LDH(Letter Digital Hyphen)字符串格式域名。为了使中文域名能被现有的域名体系所使用和处理,需要对中文域名进行 Punycode 编码变换成现有域名系统可以识别的 LDH 字符串。其主要步骤包括:

(1) 中文域名繁简转换。中文具有繁体与简体之分,在实现中文域名的过程中采用表格查找法进行简繁转换以实现统一。表中包含有繁简字的一一映射信息,以及无歧异的一对多信息。其他有歧义的字在繁简转换中不直接做处理,而是另外处理。

(2) 中文域名预处理。对于 Unicode 编码的中文域名需要对其输入进行域名标准检查,即进行预处理(Nameprep),以去掉不符合标准的字符。这一步主要包括有字符映射(Mapping)、规范化(Normalization)以及对于禁用字符的禁止输出(Prohibited Output)[5]。另外在这一步中还可以把中文域名分隔符的“。”映射为“.”。

(3) Punycode 编码实现。Punycode 编码是 AMC-ACE-Z 编码的简称,该编码算法具有较好的完整性、输出数据的唯一性、编码算法的可逆性、编码算法的高效性、编码算法实现的简易性、以及在提高编码效率的同时尽量保证编码的可读性等特点。目前已经由 IETF 发布的 RFC 3492 所定义和规范,详细情况请参见文献[4]。

3.4 Unicode 中文本地名称部分的处理

为了简化对于 Unicode 中文本地名称部分的处理,本文采取了与 Unicode 中文域名类似的处理方案。按照 Unicode 中文本地名称部分的定义以及自身的特点,进行如下操作:

(1) 中文本地名称部分预处理。本文为了简化方案,假设 Unicode 中文本地名称部分均为简单中文字符串。因而只需要对该部分进行类似于中文域名预处理相同的操作,如前所述。

(2) 进行 Punycode 编码处理。为了进一步简化处理方式,在现实中采用与中文域名部分相同的 Punycode 编码方式,如前所述。

3.5 中文邮件地址处理流程

经过上述步骤,再将 3 部分组合在一起就形成了一个 Punycode 编码的电子邮件系统。其处理过程示例图如图 2 所示。

例如,中文邮件地址:“张三@互联网。中国”,经过上述步骤之后得到的 Punycode 编码形式的中文邮件地址为“xn--ehq892b@xn--blq510jgwa.xn--fiqs8s”。

4 基于中文域名的邮件系统实现

目前在互联网上使用的电子邮件系统采用“存储转发”(Store and Forwoard)工作原理,在这种工作原理工作方式下,现有的电子邮件系统高速、有效的运行着,任何对于现有系统的升级改造都会对整个电子邮件系统造成极大的影响。为了实现基于中文域名的邮件系统,最大程度的与现有系统兼容,需要在现有系统上增加中文域名邮件地址处理模块(Chinese Mail Addresses Model, CMAM)。其实现模型如图 3 所示。

4.1 中文域名邮件系统实现模型

目前采用的 Punycode 编码方案实现了对 Unicode

表示的多语种域名编码，并用 ASCII 字符集表示最终结果，从而兼容了现行的 ASCII 域名体系。该方案可以用以借鉴实现中文域名邮件系统，即在邮件应用层的应用程序中增加邮件表示层处理模块。具体说，在传统的邮件系统中，在客户端程序上增加中文域名邮件地址处理模块（Chinese Mail Address Model，CMAM），在服务器端的邮件存储系统中加载相应中文域名邮件地址对应的 Punycode 数据，结构如图 3 所示。这种基于邮件客户端的解决方案，同多语种域名技术解决方案相一致，且只需要对用户端程序升级改造，最大程度避免了邮件基础设施的变动，保障了现有系统的稳定性、可靠性、安全性，节省了升级成本。

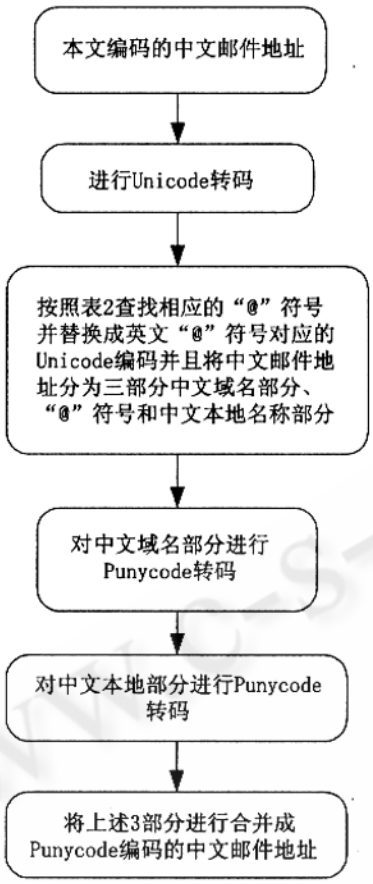

图 2　中文邮件地址处理流程

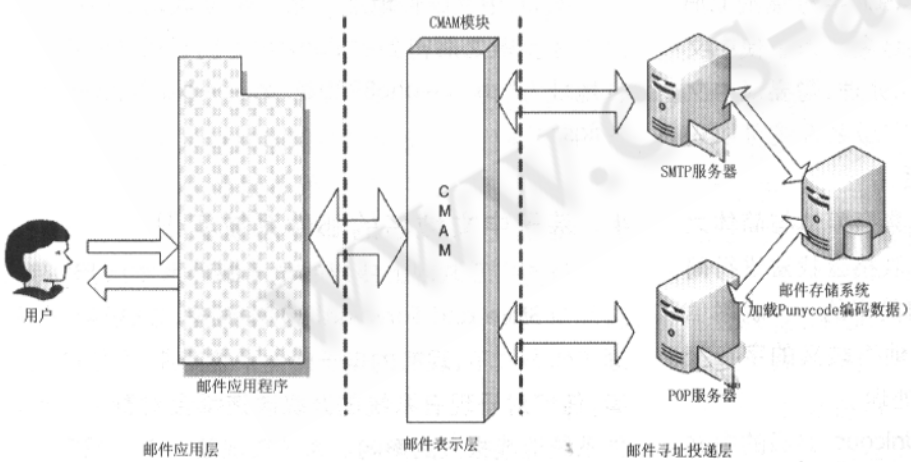

图 3　中文域名邮件系统实现模型

4.2　中文域名邮件系统示例

本文设计实现了基于中文域名的邮件系统。该系统可以支持多个中文域名；支持浏览器方式的 Webmail、邮件客户端软件访问；也支持用户使用浏览器通过 Webmail 访问方式直接进行中文进行账户注册、登录等操作；收件人可以直接使用中文邮件地址；发件人字段以中文显示，并可直接回复等特性。其建构过程如下：

（1）搭建 qmail 服务器的搭建。安装 qmail + mysql + vpopmail，配置邮件账户，测试 SMTP 和 POP3 是否正常。Vpopmail 支持虚拟域名，配置邮件账号时在用户名中加入域名信息。SMTP 如果需要认证，则在邮件应用程序中加入相应的设置。

（2）增加中文域名及中文用户。增加中文用户，配置该中文域名邮件账户，在实践中保证在显示给用户的时候使用 Unicode 编码，而在其他时候使用 Punycode 编码形式。

（3）加载转码库。本文使用 CNNIC（中国互联网络信息中心）提供的 idn – conv – linux –1.0. tar. gz 程序包，该程序包提供了用于转码的接口示例程序。

（4）二次开发 sqwebmail 。对 sqwebmail 就登录、新建邮件等子模块中涉及中文域名邮件地址的地方调用相应的转码接口，并对 sqwebmail 模块进行再编译、安装和调试。

（5）二次开发 vqregister 。vqregister 模块可以为

（下转第 53 页）

(上接第 48 页)

用户注册成功后发送确认邮件、预设用户的密码、限制密码程度等等。其二次开发的过程同 sqwebmail 类似。

(6) 设置客户端。如果客户需要使用 MS Outlook/Outlook Express,本文建议下载安装 CNNIC(中国互联网络信息中心)推出的中文上网官方版软件,以支持中文域名邮件。

5 结束语

本文提出的 3 层结构模型工程化划分方法,为基于中文域名的邮件系统的设计和实现提供了一种解决方案,这对人们在下一步针对不同分层上的相关课题研究打下了基础,将对相关邮件协议中的带格式的邮件用户名、中文简繁体用户名以及带路由信息的中文邮件地址进行处理和讨论,改进基于中文域名的邮件系统性能有现实意义。

参考文献

1 David H. Crocker. Standard for the Format of ARPA Internet Text Messages. Internet RFC 822, August 1982.

2 Ned Freed, Nathaniel S. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies. Internet RFC 2045, November 1996.

3 Ned Freed, Nathaniel S. Borenstein. Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types. Internet RFC 2046, November 1996.

4 Adam M. Costello. Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA). Internet RFC 3492, March 2003.

5 Paul Hoffman, Marc Blanchet, "Nameprep: A String – prep Profile for Internationalized Domain Names", June 24, 2002, http://www.ietf.org/internet – drafts/draft – ietf – idn – nameprep – 11.txt.