少数民族语言域名关键技术研究 *

王艳峰[a,b], 陈 涛[a], 李洪涛[a,b], 阎保平[a,b]
(中国科学院 a. 计算机网络信息中心; b. 研究生院, 北京 100190)

摘 要: 研究如何帮助少数民族人民利用本民族语言上网获取信息成为一个重要的现实问题。主要研究了支持少数民族域名的关键技术问题,提出了组成藏文、蒙文和维文三种代表性少数民族语言域名的字符集、组成规范和注册解析关键技术框架;研究了相关工程问题,实现了可运行的少数民族域名注册和解析系统。这些研究和实现工作对利用互联网保护和弘扬少数民族文化具有重要意义。

关键词: 少数民族语言域名; 多语种域名; 互联网寻址技术
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0145-03
doi:10.3969/j.issn.1001-3695.2010.01.043

Research on key technologies of minority language domain name

WANG Yan-feng[a,b], CHEN Tao[a], LI Hong-tao[a,b], YAN Bao-ping[a,b]
(*a. Computer Network Information Center, b. Graduate School, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract: Research on how to help minority people to use own language to access Internet for information becomes an important challenge. This paper focused on the key technologies of minority language domain name system. It proposed character sets and norms of three typical minority languages' s domain name, these were Tibetan, Mongolian and Uigur language. It also illustrated the technology framework of minority language domain name and addressed other relative engineering problem. This research work is of great significance for protecting and promoting minority culture of China using Internet technology.

Key words: minority language domain name; internationalized domain name; Internet addressing technology

1 多语种域名简介及少数民族语言域名研究必要性

域名注册及解析服务作为因特网的基础服务随着互联网的普及变得越来越重要,域名最初只能用英文标志,但是随着使用不同语言的网民越来越多,能用多种语言标志网站资源成为现实的需求。IETF 为了适应多语种域名的需要,发布了 RFC 3454、3490、3491、3492[1~4] 等文件指导多语种域名技术的发展。近年来,各国的互联网域名研究机构积极研究本国域名体系,并推动本国的官方语言成为多语种域名体系中的一员。

藏族、蒙族和维族等少数民族使用本民族语言撰写了相当数量的网页,同时在互联网上积累了大量少数民族文化信息资源。随着少数民族网民数量的增多,能用少数民族语言访问网站资源成为合理的要求。本文研究了支持少数民族语言域名的技术框架,并结合藏文、蒙文和维文的语言特点提出相应的域名参考字符集,阐述了提供少数民族语言域名服务的关键技术问题,供其他研究者参考。

2 藏文、蒙文和维文语言特点及域名用参考字符集

提供少数民族语言域名服务首先要解决的问题就是哪些字符可以用来组成少数民族语言域名。这个字符集的完整性必须能够满足网络资源标志的需求,同时又要符合相关的多语种域名国际标准、少数民族语言特点和业已形成的域名使用习惯。标准的域名用字符集是少数民族语言域名注册、解析和查询服务的基础。

2.1 藏文语言的特点及藏文域名字符集

藏文由吐蕃大臣吞米· 桑布扎在公元六世纪创立。现行藏文包括有 30 个辅音字母、4 个元音符号和形体别致的标点符号。藏文属辅音文字型,辅音字母又依其在音节书写中的位置和作用分为基字、上加字、下加字、前加字、后加字和再后加字。元音符号不能单独出现,必须加在辅音字母的上面或下面。藏文的书写是自左向右横写,部分复合声母采用字母上下叠写的方法表示。字母组合比较复杂。藏文共有 600 多字组合字符,梵音转写藏文字符则多达 5 000 以上。音节之间用音节符隔开,例字如图 1 所示。

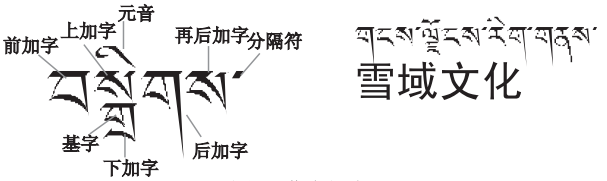

图1 藏文例字

藏文现行编码标准主要有两个:a)1997 年颁布的《信息技术 信息交换用藏文编码字符集基本集》收入藏文编码[5] 字符 192 个,俗称小字符集,该标准与 Unicode 兼容;b)另外一种字符集将藏文垂直组合(字丁)作为一个处理单元进行编码,整个藏文处理过程就与汉字处理几乎完全相同,该编码与 Unicode 不兼容,如信息交换用藏文编码字符集扩充集 *A*

收稿日期: 2009-04-09; **修回日期**: 2009-05-30 **基金项目**: 国家发改委 CNGI 项目(CNGI-09-03-04);中国科学院网络信息中心青年基金项目(CNIC_QN_08006)

作者简介:王艳峰(1973-),男,河北邯郸人,高级工程师,博士研究生,主要研究方向为互联网资源定位与寻址技术、下一代互联网技术(wyf@cnnic.cn);陈涛(1976-),男,工程师,博士研究生,主要研究方向为互联网资源定位与寻址技术、下一代互联网技术;李洪涛(1977-),男,工程师,硕士,主要研究方向为互联网资源定位与寻址技术;阎保平(1950-),女,研究员,博导,主要研究方向为互联网资源定位与寻址技术、下一代互联网技术.

和扩充集 *B*。

多语种域名国际标准采用与 Unicode 兼容的字符集，所选域名字符应能通过 Punycode 转码算法转换为七位 ASCII 字符串。所以本文在藏文小字符集基础上提出了藏文域名字符集，包括藏文基本字母、藏文数字和隔音符等共 75 个字符。本文去掉了不常用的梵音转写藏文字母、天文历算符号及域名禁止使用的标点符号，所保留的字符对于普通藏民用本民族语言寻址网站是完备的。具体字符集如表 1 所示。

表1　藏文域名字符集

字形	Unicode	字形	Unicode	字形	Unicode	字形	Unicode
▼	0F0B	ྏ	0F4F	ྐ	0F66	ྐ	0FA0
๐	0F20	ྑ	0F50	ྑ	0F67	ྑ	0FA1
๑	0F21	ྒ	0F51	ྒ	0F68	ྒ	0FA3
๒	0F22	ྔ	0F53	ྕ	0F71	ྔ	0FA4
๓	0F23	ྕ	0F54	ྖ	0F72	ྕ	0FA5
๔	0F24	ྖ	0F55	ྗ	0F74	ྖ	0FA6
๕	0F25	ྗ	0F56	྘	0F7A	ྗ	0FA7
๖	0F26	྘	0F58	ྙ	0F7C	྘	0FA8
๗	0F27	ྙ	0F59	ྚ	0F90	ྙ	0FA9
๘	0F28	ྚ	0F5A	ྛ	0F91	ྚ	0FAA
๙	0F29	ྛ	0F5B	ྜ	0F92	ྛ	0FAB
๚	0F40	ྜ	0F5D	ྞ	0F94	ྜ	0FAD
๛	0F41	ྞ	0F5E	ྟ	0F95	ྞ	0FB1
๜	0F42	ྟ	0F5F	ྠ	0F97	ྟ	0FB2
๝	0F44	ྡ	0F60	ྡ	0F99	ྠ	0FB3
๞	0F45	ྡ	0F61	ྡྷ	0F9A	ྡ	0FB4
๟	0F46	ྣ	0F62	ྤ	0F9C	ྣ	0FB6
๠	0F47	ྤ	0F63	ྥ	0F9E	ྤ	0FB7
๡	0F49	ྥ	0F64	ྦ	0F9F		

2.2　蒙文语言的特点及蒙文域名字符集

蒙古文有 34 个字母，其中有元音 7 个，辅音 27 个。蒙文的拼写规则是以词为单位竖写，词与词之间用空格分开。一个词内各个字母之间连着写。书写规则是采取从上到下的书序，从左到右的行序。蒙古文词性又分阳性、阴性和中性。每一个字母在字首、字中、字尾有不同的变体。

蒙文国际标准码 2000 年 2 月得到国际标准化组织的正式通过，Unicode 技术码位为 U + 1800 － U + 18AF。本文在此基础上提出了蒙文域名字符集，包括蒙文字母、阿拉伯数字字符和连接符，共 58 个字符。蒙文词与词之间用空格分开，而空格是组成域名的禁止字符，所以选取连接符用来代替空格，满足蒙文的分词习惯。具体字符集如表 2 所示。

表2　蒙文域名字符集

字形	Unicode	字形	Unicode	字形	Unicode	字形	Unicode
FVS1	1808	ᢍ	182D	ᢝ	183D	1	0031
FVS2	180C	ᢎ	182E	ᢞ	183E	2	0032
FVS3	180D	ᢏ	182F	ᢟ	183F	3	0033
ᠰ	1820	ᢐ	1830	ᢠ	1840	4	0034
ᠱ	1821	ᢑ	1831	ᢡ	1841	5	0035
ᠲ	1822	ᢒ	1832	ᢢ	1842	6	0036
ᠳ	1823	ᢓ	1833	ᢩ	1849	7	0037
ᠴ	1824	ᢔ	1834	ᢋ	184B	8	0038
ᠵ	1825	ᢕ	1835	ᢔ	1854	9	0039
ᠶ	1826	ᢖ	1836	ᢕ	1858	－	002D
ᠷ	1827	ᢗ	1837	ᢖ	185B		
ᠸ	1828	ᢘ	1838	ᢗ	185E		
ᠹ	1829	ᢙ	1839	ᢘ	1861		
ᠺ	182A	ᢚ	183A	ᢙ	1862		
ᠻ	182B	ᢛ	183B	ᢚ	1874		
ᠼ	182C	ᢜ	183C	O	0030		

2.3　维文语言的特点及参考字符集

现代维吾尔文是在晚期察合台文基础上形成的以阿拉伯文字母为基础的拼音文字，有 8 个元音字母，24 个辅音字母，自右向左横写。每个字母按出现在词首、词中、词末的位置有不同的形式。

《信息交换用维吾尔文、哈萨克文、柯尔克孜文编码字符集、基本集与扩展集》[5] 所规定的字符集是目前维文书面文字的标准规范，该字符集实际采纳了 ISO/IEC 10646-1：2000（Unicode 3.0）字符集规定。根据维文的使用习惯，本文提出的维文域名字符集保留了维文字母、阿拉伯数字和连字符作为维文域名的合法字符，共 45 个。具体字符集如表 3 所示。

表3　维文域名字符集

字形	Unicode	字形	Unicode	字形	Unicode	字形	Unicode
ھ	06BE	م	0645	ى	064A	2	0032
ب	0628	ن	0646	ر	0632	3	0033
ع	063A	و	0648	ز	0698	4	0034
د	062F	ۋ	067E	ف	0641	5	0035
ې	06D0	چ	0686	ك	06AF	6	0036
ا	0627	ر	0631	ح	062E	7	0037
ە	06D5	س	0633	ج	062C	8	0038
ى	0649	ت	062A	ۇ	06C6	9	0039
ل	06AD	ۇ	06C7	ك	0626	－	002D
ق	0642	و	06C8	لا	0644+0627		
ل	0643	ۇ	06CB	0	0030		
ل	0644	ش	0634	1	0031		

3　少数民族域名组成规范及注册解析技术框架

多语种域名是指非英语国家为推广本国/本族语言的域名系统的总称。同英文域名一样，少数民族语言域名也存在组成规范的问题，其中包括所用字符、域名长度、是否可以与其他语言相混合等问题。本文在尊重少数民族语言使用习惯的基础上，本着简化域名系统复杂性的原则提出如下藏文、蒙文和维文域名的组成规范。

3.1　藏文域名规范

藏文域名组成范式：

〈domain〉：：=〈entity-name〉.〈A-TLD〉.
〈entity-name〉：：=〈label〉|〈entity-name〉"."〈label〉
〈label〉：：=〈letter〉[[〈ldh-str〉]〈letter〉]
〈ldh-str〉：：=〈let-dig-hvp〉|〈let-dig-hyp〉〈ldh-str〉
〈let-dig-hyp〉：：=〈let-dig〉|"－"
〈let-dig〉：：=〈letter〉|〈digit〉
〈letter〉：：=Tibetan alphabetic character
〈digit〉：：=Tibetan ten digits

其中：〈A-TLD〉为 cn 根域名；〈entity-name〉部分包含藏文字母（辅音和元音）、藏文数字、必要的隔音符号等，合法字符取自前文提出的藏文域名字符集。各级藏文名字之间用实点（.）连接，各级藏文名字长度不得超过 20 个藏文字母。

3.2　蒙文域名规范

蒙文域名范式：

〈domain〉：：=〈entity-name〉.〈A-TLD〉.
〈entity-name〉：：=〈label〉|〈entity-name〉"."〈label〉
〈label〉：：=〈letter〉[[〈ldh-str〉]〈letter〉]
〈ldh-str〉：：=〈let-dig-hvp〉|〈let-dig-hyp〉〈ldh-str〉
〈let-dig-hyp〉：：=〈let-dig〉|"－"
〈let-dig〉：：=〈letter〉|〈digit〉
〈letter〉：：=Mongolia alphabetic character
〈digit〉：：=any one of the ten digits 0 through 9

其中：〈entity-name〉部分由蒙文字母、阿拉伯数字和连接符组成。规范字符采用前文提出的蒙文域名字符集。各级蒙文名字之间用实点(.)连接，各级蒙文名字长度不得超过 20 个蒙文字母。

3.3　维文域名规范

维文域名的书写特点是从右向左，RFC 3454 规定其字符属性具有 AL 属性，不能与具有 L 属性字符语言相混合组成域名字符串，因此维文不能与英文字符混合组成域名，但是可以与不具有 L 属性的其他字符混合，如与阿拉伯数字和连字符组成维文域名串。具有 AL 属性的语言组成域名时，各级名字开头和结尾都需要采用 AL 属性的字符。考虑以上多语种域名的技术规定和维文语言的使用习惯，本文提出如下范式形式的维文域名：

〈domain〉::=〈entity-name〉.〈A-TLD〉.
〈entity-name〉::=〈label〉|〈entity-name〉"."〈label〉
〈label〉::=〈letter〉[[〈ldh-str〉]〈letter〉]
〈ldh-str〉::=〈let-dig-hvp〉|〈let-dig-hyp〉〈ldh-str〉
〈let-dig-hyp〉::=〈let-dig〉|"-"
〈let-dig〉::=〈letter〉|〈digit〉
〈letter〉::= Uigur alphabetic character
〈digit〉::= any one of the ten digits 0 through 9

其中：〈entity-name〉包含维文字母、数字(0 ~ 9)或连接符(-)；各级维文名字之间用实点(.)连接，各级维文名字长度不得超过 15 个维文字母，且必须以维文字母开头和结尾；〈A-TLD〉为 cn 根域名。

3.4　少数民族语言域名注册和解析技术框架

与英文域名的使用不同，少数民族语言域名的使用必须得到应用程序支持。少数民族网民使用本民族语言输入法在应用程序中输入规范的少数民族语言域名后，应用程序必须利用 Punycode 算法将其转码为 ACE(ASCII compatible encoding)编码的英文域名[2]；然后调用主机解析器进行 DNS 记录查询并接收查询结果。应用程序还要根据需要将 ACE 编码的英文域名转换为少数民族语言域名，完成必要的显示工作。目前 IE7 和 Firefox2 等浏览器已经可以完成上述的转码和显示工作，少数民族语言域名可以在这两款浏览器中直接使用。

另一方面，少数民族语言域名服务机构提供注册、解析和查询服务。使用少数民族语言域名的用户应该先通过域名注册商完成购买域名服务的流程，提供域名所对应的 IP 地址等信息进行数据注册。注册后的少数民族域名转码成为对应的 ACE 英文域名存储在数据库中，并生成 zonefile 文件加载到 DNS 服务器上对外提供解析服务。图 2 是“新闻. cn”的藏文域名的注册和解析过程。网站管理者注册该藏文域名时，藏文域名将被转码为相对应的英文域名“xn--nbd3ib3jidx0juy. cn”存储到数据库中，并加载到 DNS 服务器上。网民在 IE7 等浏览器地址栏中输入藏文域名，浏览器自动使用 Punycode 转码后的域名向 DNS 进行查询并取得相应的 IP 地址。网民就可以直接用藏文访问网站了。

4　少数民族域名服务的其他工程问题

4.1　少数民族语言的编码支持问题

少数民族语言域名字符表是 Unicode 编码的子集。其表示码可以采用 UTF8、UTF16 或者 GB18030。其中 GB18030 属于变长编码，分为 1 字节区、2 字节区和 4 字节区。4 字节区定义了我国少数民族(蒙、藏、维吾尔、彝、朝鲜等)文字的编码。目前广泛使用的 Windows 2000/XP 操作系统不支持 GB18030 的 4 字节汉字的直接显示，但是主机里的 Unicode 编码汉字与外部设备的 GB18030 编码的转换工作是可以完成的。因此，GB18030 也可以担当少数民族语言域名表示码的工作。

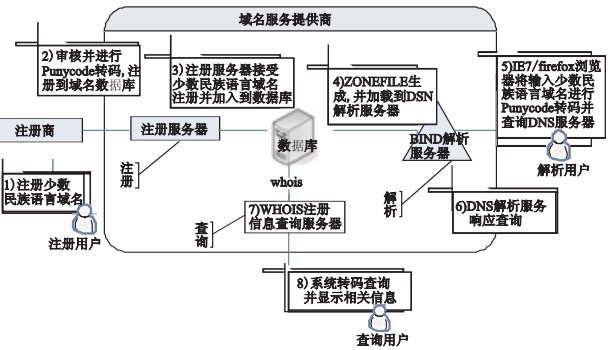

图2　少数民族域名服务系统框架图

4.2　少数民族语言的输入法支持

目前并没有专门支持少数民族语言的商用操作系统可供使用，使用少数民族语言域名需要在主流的操作系统中安装相应的少数民族语言输入法。维文、蒙文和藏文输入法可以采用 Vista 平台上的微软输入法。新疆理化所研制的多语种智能输入法也可以用于维文域名的输入。尽管还有一些如同元藏文输入法、蒙科利蒙文输入法等少数民族输入法存在，但是它们使用的内部编码不是 Unicode 编码形式，因此不能用这些输入法来使用少数民族域名。

4.3　系统开发工具对少数民族语言域名的支持

少数民族域名字符集在 Unicode 3.0 版本中得到完整的支持，Java JDK 必须在 1.4 以上才能支持域名服务系统的开发工作。Oracle9i 以后的版本已经开始对 GB18030 提供支持，使用的编码名称为 ZHS32GB18030，可以根据存储的需要将少数民族语言域名存储为 GB18030 或者 UTF8 格式。

5　结束语

随着少数民族网民的日益增加，研究推广少数民族语言上网技术日益重要。这些技术包括利用少数民族语言寻址网站、信息发布和资源搜索等，这些研究工作对保护和弘扬少数民族文化具有重要意义。目前操作系统、输入法和显示软件包对少数民族语言的支持非常薄弱，相关标准很少，需要有更多的研究人员对少数民族语言上网技术给予足够的关注。

参考文献：

[1] HOFFMAN P, BLANCHET M. RFC 3454, Preparation of internationalized strings (“stringprep”)[S/OL]. (2002). ftp://ftp.rfc-editor.org/in-notes/rfc3454.txt.

[2] FALTSTROM P, HOFFMAN P, COSTELLO A. RFC 3490, Internationalizing domain names in applications (IDNA)[S/OL]. (2003). ftp://ftp.rfc-editor.org/in-notes/rfc3490.txt.

[3] HOFFMAN P, BLANCHET M. RFC 3491, Nameprep: a stringprep profile for internationalized domain names (IDN)[S/OL]. (2003). ftp://ftp.rfc-editor.org/in-notes/rfc3491.txt.

[4] 中国国家技术监督局. GB16959—1997, 信息技术信息交换用藏文编码字符集基本集[S]. 北京：中国标准出版社，1998.

[5] 新疆维吾尔自治区质量技术监督局. DB65/2190—2005, 信息交换用维吾尔文、哈萨克文、柯尔克孜文编码字符集、基本集与扩展集[S]. 北京：中国标准出版社，2005.