

文章编号: 1003 - 0077 (2006) 04 - 0088 - 06

Linux 下维、哈、柯文多语种图形化处理平台的设计与实现^{*}苏国平¹, 缪成², 夏国平¹

(1. 北京航空航天大学, 北京 100083; 2. 中国科学院新疆理化研究所, 新疆 乌鲁木齐 830011)

摘要:针对维吾尔文字、哈萨克文字、柯尔克孜文字 (以下简称“维哈柯文”) 的特点以及进行维哈柯文、西文等多语种混合处理时的特殊需求, 本文通过对 Linux 的 II8N 体系中 NLS (National Language Support) 研究分析, 提出了基于 Linux 的多语种图形化处理平台的设计目标与总体架构。该平台由维哈柯文本地化环境、维哈柯文显示、自适应维哈柯文输入和维哈柯文打印输出等 4 个子系统的十余个模块组成。本文详细介绍了各子系统主要模块的实现技术。通过在 redhat linux 8.0、turboLinux 上测试表明, 该平台在桌面环境、编辑软件、网络浏览、数据库软件、多媒体软件、图形处理软件等应用中均能较好的实现维哈柯文、汉文、西文的混合输入、显示、编辑、排版、打印等功能。

关键词: 计算机应用; 中文信息处理; 多语种; 图形化处理平台; Linux**中图分类号:** TP391**文献标识码:** A**Design and Implementation of Multilingual GUI Processing Platform
Based on Linux for Uighur, Kazakh and Khalkhas**SU Guo-ping¹, MIAO Cheng², XIA Guo-ping¹

(1. Beijing University of Aeronautics & Astronautics, Beijing, 100083;

2. The Xinjiang Technical Institute of Physics & Chemistry, CAS, Urumqi, 830011, China)

Abstract: According to the lingual characteristics of Uighur, Kazakh and Khalkhas (abbreviated as UKK in the following) and the special requirements for supporting those minority languages with Chinese and English at the same time, in this paper we presents the design goals and general framework of multilingual GUI processing platform under Linux environment based on the analysis and research of national language support in the system of II8N. The platform consists of four sub-systems, including localization, display, auto-adaptation input and printing of UKK, which are made of more than ten modules. The implementation of these modules is introduced in detail. Our test shows the platform can support inputting, display, editing and printing of UKK, Chinese and Western Languages in common applications smoothly under Redhat Linux 8.0 and TurboLinux environment.

Key words: computer application; Chinese information processing; multilingual; GUI processing platform; Linux**1 维哈柯文的语言特征和设计难点**

维吾尔文字、哈萨克文字、柯尔克孜文字 (以下简称“维哈柯文”) 均系阿拉伯语系的拼音文字, 它们与汉字和西文有很大的不同。首先是书写方向相反, 汉字和西文的书写方向是从左到右, 一般将这种输入的字符由左向右依次排列的文字称为右向输入文字, 而维哈柯文的书写

^{*} 收稿日期: 2005 - 04 - 11 定稿日期: 2005 - 10 - 24

基金项目: 国家 863 计划资助项目 (2003AA1Z2110)

作者简介: 苏国平 (1957—), 男, 教授级高级工程师, 博士生, 研究方向为软件工程。

方向是从右向左,因此也称为左向输入文字。其次维哈柯文的同一个字母依在单词中的位置不同会有四种变形。即首写形、中写形、尾写形和独写形。在文字输入时要根据字母在词中的位置来确定使用何种形式。第三、维哈柯文之间也有部分差别,维、哈、柯文三种文字的字母数和字母组成也不相同。依据维哈柯文的以上特点,设计维哈柯文图形化处理平台需要解决的技术难点有:(1)设计合理的实现方案,设置正确的本地化环境,实现图形化界面下汉文、英文、维哈柯文可以正确混合显示、排版和打印;(2)可以同时输入汉文、英文和维哈柯文。且维哈柯文输入法要有较大的弹性,以适合不同场合的需要;(3)在保证汉文、西文从左到右的输入的同时,解决维哈柯文行文方向相反的问题,实现维哈柯文的从右到左的输入;(4)按照维哈柯文字母在词中出现的位置,自动选定正确的字母形式,使得每个字母都可以准确相连。

2 多文种平台设计目标和总体设计方案

在Linux系统中,一般是通过搭建符合本民族语言特征的i18N体系中NLS(National Language Support)来建立程序的本地化运行环境,见图1。NLS通过建立符合民族语言特征的本地化环境(LOCALE)、输入方法(M)、字体(FONT)和消息机制(MESSAGE)为不同地域、不同语言环境的应用提供本地化支持。因此,建立多文种平台就是搭建合适的NLS。

新疆是一个多民族语言混用的地区。因此,一个标准的维哈柯、汉、英多文种Linux图形化处理平台,其设计目标应是:

(1)可以同时正常混合显示、编辑和打印维哈柯文、汉文和英文;

(2)在该平台上运行的任何应用程序应无须特殊改动就可以输入、输出维哈柯文,且符合XM规范,并可以在维哈柯文之间和维哈柯文、汉文和英文之间自由切换输入文字;

(3)系统的提示和出错信息可以用维哈柯文表示;

(4)系统图形化界面的菜单、对话框、文件名等应能选用维哈柯文、汉文、西文中的任何一种文字表示。

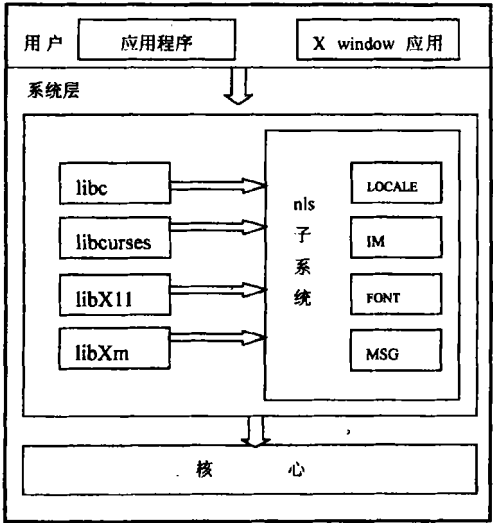


图1 Linux系统中的NLS架构图

实现此设计目标的关键就是确定维哈柯文信息的编码。因为要求该多语种平台可以同时处理维哈柯文、汉文、英文,且确保这五种文字可以混合显示、编排和打印,就必须要选择一种可以同时处理这五种文字的大字符集编码。

虽然Linux已经支持UTF-8编码,但是由于UTF-8与目前流行的中文Linux上广泛使用的GB内码不兼容,要实现汉文和维哈柯文的同时输入、显示、打印,系统就必须不断进行UTF-8与GB编码的转换,这样即增加了系统的复杂性,也影响了整个系统的效率,因此采用UTF-8编码显然不合适。

而我国以前颁布的GB2312和GBK编码也不包括维哈柯文字符的码位。基于以上考虑,本系统最终选择新颁布的GB18030编码作为系统内码。这是因为GB18030标准包括汉文、英文字符的编码,在四字节部分也包括维哈柯文字符的编码,而且维哈柯文的四字节部分在转换为Linux信息存储编码UTF-8时对应的是UTF-8编码的阿拉伯字母区,这样可以利用部分与

维哈柯文一样的阿位伯文字母及系统对阿位伯文字母的处理函数来进行维哈柯文显示处理,而且在 red hat Linux 7.1 以后系统已基本实现 GB18030 标准,可以利用 Linux 的现有本地化成果,对支持 GB18030 的本地化环境进行部分的修改,同时在中文 GB18030 的 TrueType 字库中加入维哈柯文字型,这样就编制了维哈柯文本地化环境,实现了维哈柯文字符的正常显示。这样可以在中文文化平台的基础上嵌入维哈柯文字符的处理,对中文平台做不大的改动,就实现了维哈柯文、汉文和英文的混和字符的正常显示。同时,也利用现有 Linux 系统对 GB18030 编码的支持和中文本地化数据环境,实现系统的底层函数(包括 C library、X library 等)正常处理维哈柯文信息,大大缩短了系统的开发周期。

在确定了系统内码后,根据维哈柯文多语种图形化处理平台的所应达到的设计目标,系统将平台划分为由十余个模块组成的四个子系统,即维哈柯文本地化环境子系统、维哈柯文显示子系统、自适应维哈柯文输入子系统和维哈柯文打印输出子系统。各子系统和模块关系如图 2 所示。

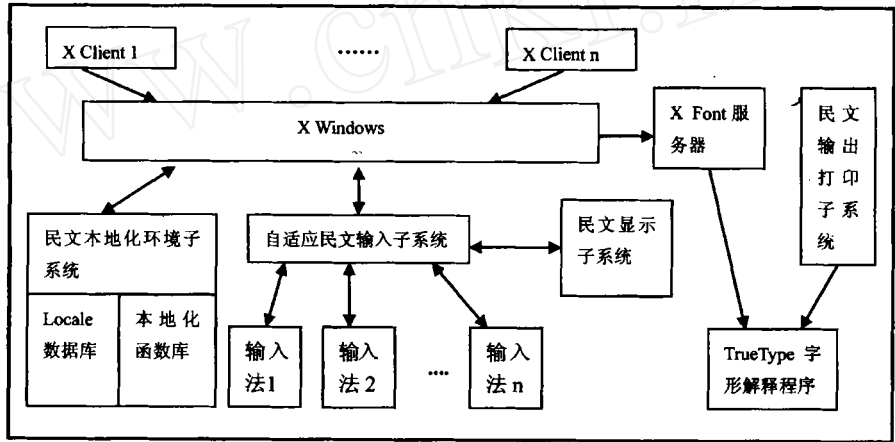


图 2 多语种平台模块结构

各部分的主要功能为:

(1)维哈柯文本地化环境子系统包括本地化函数库和本地化数据库两部分,负责给整个系统提供与维哈柯文语言特性有关的描述信息、维哈柯文图形化界面字符数据和维哈柯文字符编码处理函数集,以支持五种文字的混合编辑显示,同时用户可以在登陆时选择其中一种文字作为界面显示文字。

(2)维哈柯文显示子系统作为维哈柯文输入子系统的扩充,在应用程序输入维哈柯文、汉文和英文时,为维哈柯文与其他文字的混合排版和显示输出提供支持。

(3)自适应维哈柯文输入子系统。可以为 X Client 提供维文、哈文、柯文、汉文的输入服务。该子系统符合 XM 标准,可以与其他子系统在同一台主机上,也可以在不同的主机上,X Client 可以通过 X 协议或 TCP/IP 协议来请求维哈柯文输入子系统提供具体的输入服务。该子系统与具体的输入法及当前平台界面所使用的语言无关,支持在系统运行中自由切换维文、哈文、柯文、汉文和英文输入的功能,这使得用户可以同时对汉文、英文、维哈柯文进行混合输入。同时该系统还支持拦截 libX11.so 的函数功能,使不支持 XM 输入的应用程序接受维哈柯文输入。

(4)输入法模块提供各种具体的中文输入法(如五笔、智能拼音)和维文、哈文、柯文输入法。

(5)维哈柯文打印输出子系统为整个系统提供与打印机无关的高品质打印输出服务。

3 主要子系统的实现技术

3.1 自适应维哈柯文输入子系统

自适应输入子系统分为输入服务器模块和输入法管理模块。

3.1.1 输入服务器模块

该模块支持 XM 协议和拦截 libX11.so 两种方式向应用程序发送输入字符串。XM (X Input Method) 是 X Window 下符合国际化标准的输入法协议, 它包括 Client/Servers 模型和 Library 模型, 本系统使用了 Client/Servers 模型, 支持全部 root, on the spot, over the spot 和 off the spot 四种输入方式, 它与 X Client 通过 Inter-Client Communication Convention 进行通信。X Client 支持 XM 协议, 就可以根据环境变量 locale 和 XMODIFIERS 的值找到输入服务器模块。输入服务器为每个 X Client 建立一个输入上下文 (IC), 同时根据应用软件支持的输入方式, 调用相应输入方式服务程序段, 向 X Client 输入维哈柯文、汉文。由于所有与维哈柯文、中文输入有关的工作都由输入服务器完成, 因此, 对于应用程序来说, 这个过程是完全透明的。但是对于未按国际化/本地化规范开发的 X 应用程序, 由于不支持 XM 协议, 初始化时不查找输入服务器, 处理输入事件时就根本不可能向输入服务器发出“翻译”请求, 不能正常输入维哈柯文、汉文。

针对这种情况, 本系统采用了双通道输入方法。由于 Linux 操作系统执行程序时, 大部分都会自动链接到环境变量 LD_PRELOAD 所指向的动态函数库, 因此将 Linux 动态链接库 libX11.so 修改为 WapXlib.so, 并对某些单字节函数进行扩充, 同时把 LD_PRELOAD 设成 WapXlib.so, 这样在系统指定的本地环境下, 如果应用程序不支持 XM 协议输入, 并调用 WapXlib.so 中扩充函数处理按键事件时, 就将按键送到输入服务器, “截获”了 libX11.so 中相关函数, 按照输入服务器模块当前设置, 进行按键“翻译”, 并将结果返回, 建立第二个输入通道, 强迫这些软件接受维哈柯文、汉文的输入。通过这种输入方式, 可以使 Linux 上的绝大多数 X Client 可以正常地接收民维哈柯文、汉文的输入。

3.1.2 输入法管理模块

该模块将输入法的挂接、切换功能定义成一组回调函数接口 MPI, 通过这些接口, 系统即可以根据配置文件中数据挂接输入法程序, 用户又可以在运行时根据编辑的需要操纵 MPI, 在维文、哈文、柯文和汉文输入法模块之间自由切换, 实现维哈柯文、汉文和英文的混合输入。同时 MPI 也是输入法程序开发者开发新的输入法模块的基础。只需要为每一个输入法实现 MPI 回调函数接口定义及其输入法算法, 编译成动态链接库, 并在配置文件中加入相应文件路径, 就可以在系统运行时链接成完整、可以运行的输入法程序。

3.2 维哈柯文显示子系统

该系统由字形的自动选形模块、倒序模块、字符识别模块和显示调整模块 (如下图 3 所示) 组成。

3.2.1 字符识别模块

该模块负责判断自适应维哈柯文输入子系统输入的字符是否是维哈柯文字符, 若是 (即属于 GB18030 维哈柯文码位的编码) 就送入倒序模块, 否则直接将该字符送出显示。

3.2.2 倒序子模块

维哈柯文书写方向是从右到左, 因此,

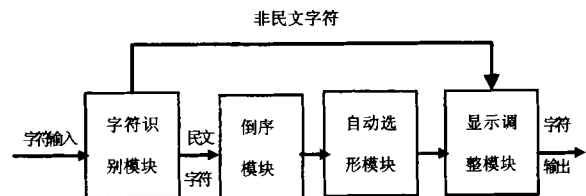


图 3 维哈柯文显示子系统模块构架图

要将字符预处理模块送来的维哈柯文字符进行一次倒序。该模块负责把维哈柯文字符串进行倒序,将左、右字符顺序对换,再将该字符送自动选形子模块。

3.2.3 自动选形子模块

该模块依据前面输入的维哈柯文字母和控制字符来确定当前维哈柯文的输入状态,在接收到维哈柯文字符数据后,依据该状态对维哈柯文字符或字符串进行选形,按照字符在词中出现的位置自动选择其对应的首写形、中写形、尾写形或独写字形编码替换该维哈柯文字符编码,使前后字母的字形能准确相连,再发送到显示调整模块。这样做输入子系统只管输入维哈柯字母,由自动选形模块进行字形的选择。

3.2.4 显示调整子模块

根据左向输出的要求,加入光标移动控制字符,使维哈柯文输入、显示符合日常书写规范,实现维哈柯文字的推挤输入方式。

3.3 维哈柯文本本地化环境子系统

该子系统主要包括本地化函数库和本地化数据库两部分,是开放系统进行本地化的基础。

3.3.1 本地化函数库

本地化函数库包括 NLS 中显示、转换、传输本地化编码(如 GB18030 编码)字符的函数,由于维哈柯文平台采用 GB18030 作为系统的基本内码,而现有 Linux 已经实现了对 GB18030 中汉文的支持,因此,系统以支持 GB18030 的 NLS 函数为基础进行维哈柯文扩充,增加处理多语种维哈柯文语言混合字符串的能力。当应用程序调用本地化函数时,可以根据处理编码的码位确定是使用原有的中西文处理过程还是使用维哈柯文处理过程,从而使调用 Linux 本地化函数库的应用程序都可以处理汉文和维哈柯文而无需进行任何改动。这一部分需要扩充的主要是 NLS 中涉及 GB18030 与国际标准的宽字节编码及其变形(如 UTF-8)之间的转换、显示函数。主要包括: `mbtowc()`、`wctomb()`、`mbstowcs()`、`wcstombs()`、`mblen()`、`putwchar()`、`getwchar()`、`getws()`、`putws()` 等等。

3.3.2 本地化数据库

本地化数据库是指 glibc 的本地化数据库(`/usr/lib/locale` 目录下的各语言编码目录)、Xlibc 的本地化数据库(`/usr/X11R6/lib/X11/locale` 目录下的各语言编码目录)以及本地化图形桌面和应用程序界面显示字符数据文件(`/usr/share/locale` 目录下各语言编码目录)这三部分,这一部分也由于系统使用了 GB18030 作为系统的内码,可以利用支持 GB18030 的中文版 Linux 成果,将其 `/usr/X11R6/lib/X11/locale` 目录下的 `zh_CN.gb18030` 中的本地化文件基本不改动就可以作为维哈柯文本本地化数据库。而 `/usr/lib/locale` 下的 `zh_CN.gb18030` 目录下的文件较多,这一部分就需要根据维哈柯文字符分类、字符显示宽度、字符串比较和理序以及各自货币、数字、时间表示格式进行改动,但由于最重要的字符集分类可以使用中文 GB18030 环境的,所以改动也不是很难。而 `/usr/share/locale` 下的更改就更加容易,主要工作就是将 .po 文件的英文条目翻译成维哈柯文字符串并编译成 .mo 文件。

3.4 维哈柯文打印输出子系统

该子系统在系统打印维哈柯文时,将打印维哈柯文字符数据转换成 postscript 方式,再输出到打印机进行打印。

4 系统的测试和结论

本系统开发完成后,在 redhat linux 8.0、turbolinux 上进行了测试,在这些系统上安装本系

统以后,在 gnome或 KDE桌面环境下,实现了桌面及菜单维哈柯文。同时一些支持国际化的软件,在加入了相应的.mo文件以后,均可以正确显示维哈柯文菜单和维哈柯文消息。在编辑软件 gedit、StarOffice中,也可以正常混合输入、显示、编辑、打印维哈柯文、汉文和英文。在网络浏览器 mozilla、数据库软件、多媒体软件、图形处理软件等也均可以在其文本框中输入维哈柯文。实现了 LINUX下的维、哈、柯、汉、英多文种图形界面处理平台的功能。

参 考 文 献:

- [1] 戴庆厦,许寿椿,高喜奎.中国各民族文字与电脑信息处理[M].北京:中央民族出版社,1991,83-94
- [2] 信息交换用汉字编码字符集基本集扩充[S].北京:中国标准出版社,2000
- [3] 毛德操,胡希明. linux内核源代码情景分析(下册)[M].杭州:浙江大学出版社,2001,330-337.
- [4] 吴健,孙玉芳,李国华.“炎黄 中文平台结构设计[J].中文信息学报,2001,15(4):53-58
- [5] 董治江,吴健,钟义信.在 ICU中实现少数民族文字的处理[J].中文信息学报,2004,18:66-72
- [6] Masahiko Narita, Hideki Hiura: The InputMethod Protocol[S]. Version 1.0, X Consortium Standard, Open Software Foundation, November 1990
- [7] Valerie Quercia, Tim O'Reilly: X Window System User's Guide [M]. Volume Three, O'Reilly & Associates Inc., January 1993, 125-158

(上接第 62页)

- [10] Idomuso, Dawa, Katsuhiko Shirai, et, al Design of Mongolian speech database considering dialectal characteristic[J]. J. Acoust Jpn (E) 20, 3, 1999, Vol 20, No 3. p181-188
- [11] 伊.达瓦,大川茂树,白井克彦.蒙古语多方言语音识别及共享识别模型探索[J].中国中央民族大学学报,2001,28(4):114-121.
- [12] 伊.达瓦,大川茂树,白井克彦.蒙古语主要方言的声学 and 音律特征分析分类[J].中国民族语文,2001,1,26-32
- [13] 伊.达瓦,大川茂树,白井克彦.,等.蒙文通信的新方法[J].俄罗斯语言学学报,2005,24(1):44-50
- [14] Idomuso Dawa, Hitoshi Isahara, et, al Automatically Obtain A Corpus for Minority Language [A]. iSTEPS-2004 and Oriental COCOSDA-2004. SPLASH-2004[C]. New Delhe: The McGraw-Hill Co. 2004, Vol 2, p42-46
- [15] 伊.达瓦,大川茂树,白井克彦. Multi-sound source clustering-editing, and retrieving for meeting and audio data[A]. 日本音声学会. 日本音声学会 2004年秋季研究发表会 [C]. 日本冲绳, 日本音声学会, 2004, p201-202
- [16] 伊.达瓦,白井克彦. 语音数据发话人自动识别分类编辑系统(科技发明专利报告书)[M]. 专利公开代码 2004-53821, 日本.
- [17] 合同书. Agreement Of Cooperation Between The National Institute Of Information And Communications Technology Of Japan And Inner Mongolia MENKSOFT Software CO., LTD [M]. 2005. 10
- [18] I Dawa, Husela, Ulang et al Multilingual parallel electronic dictionary of Mongolian, English, Chinese, Japanese, and Korean [A]. ICEIC2006 CONFERENCE COMM ITTEE [C]. Ulaanbaator Mongolia MUST, 2006 (accepted paper).
- [19] 伊.达瓦,井佐原均.蒙古语多文种-多语言文本-口语语料库的建设[A].青海师范大学.第十届中国少数民族语言文字信息处理研讨会论文集[C].青海西宁:青海师范大学,2005,86-95.