UNIX 中文信息处理系统实现的方案和技术*

孙玉方

(中国科学院软件研究所)

THE IMPLEMENTATION STRATEGIES AND TECHNOLOGIES OF UNIX CHINESE INFORMATION PROCESSING SYSTEMS

Sun Yufang

(Institute of Software, Academia Sinica)

ABSTRACT

Based on the development projects of UNIX Chinese processing information systems for many years, The principles and tasks of UNIX Chinese processing are introduced. The several different development strategies are compared, and implementation technologies used are presented.

摘　要

本文从作者多年开发UNIX中文信息处理系统的实践出发，系统地介绍了UNIX中文处理的基本原理和主要工作，比较了系统实现各种的方案，并且介绍了实现时采用的主要技术。

§1. 引　言

UNIX操作系统最初是针对单一的语言——英语处理能力开发的。然而随着它近十年来的迅猛发展，它在许多非英语国家里也得到了越来越广泛的应用。比如，在我们国家，从1983年起引进和生产了大量运行UNIX系统或其变种的微型机、小型机及工作站。随着这些机器的应用，一个十分尖锐的问题摆在计算机厂家及软件开发者面前：在UNIX上开发

* 1989 年6 月25 日收到。

中英文兼容的信息处理系统，以便可以更方便地和计算机进行中文人机对话。正是在这样的形势下，我们在国内率先开展了UNIX中文系统的研究和开发，至今已有六、七年的历史。

这几年我们已经在多种机器上完成了基于UNIX的中英文信息处理系统。本文从我们的的开发实践出发，对有关的问题进行一次系统的总结。下面，第二节讲述UNIX系统中文处理的基本原理和主要工作；第三节讲述实现中文处理系统的几种方案；第三节介绍实现时采用的主要技术；最后是一些体会。

§2. UNIX系统中文信息处理的基本原理与工作

2.1 UNIX系统中文处理流程

UNIX系统中文信息处理的基本原理与在别的系统(比如DOS、VMS等)上采用的原理是类似的。其中最基本的部分就是解决中文的输入与输出问题，如图1所示。

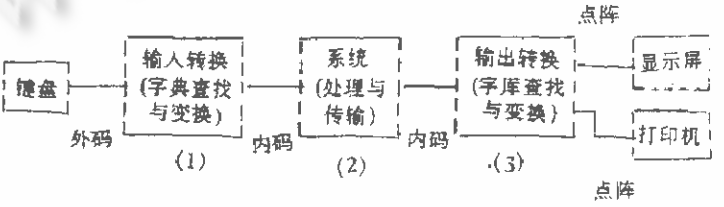

图1 中文输入/输出基本流程

从键盘上键入的表示各种输入方案的代码(称为外码)，必须经过输入转换模块(1)，变换成系统能统一处理的内部形式(称为内码)；内码在系统(2)中进行处理、在内外存之间进行传输；必要时，内码经过输出转换模块(3)变换成汉字点阵在显示屏或打印机上输出。如果把键盘与输入转换模块结合，把输出转换模块与显示屏结合，两者组合在一起就成了中文终端；如果把输出转换模块与普通打印机结合在一起就构成了中文打印机。也可以完全由软件实现输入转换和/或输出转换。在某种情况下，也可以由硬件(比如采用汉卡)来储存输入变换用的数据(字典)及输出变换用的数据(字库点阵)，甚至把输入变换及输出变换程序也放在卡上，全部由硬件完成。当然如果只用汉卡存放字典字库数据，而由系统来实现转换程序也是可以的。

2.2 UNIX系统结构与中文信息处理的基本工作

1. UNIX系统主要结构

上面所说的输入和输出变换主要涉及的是系统基本核心部分，就整个UNIX系统来说这仅是常驻内存的一小部分。整个UNIX系统分成核心层、shell层、实用程序层，上面还有用户应用程序、用户等，如图2所示。

2. 开发UNIX中文信息处理系统的主要工作

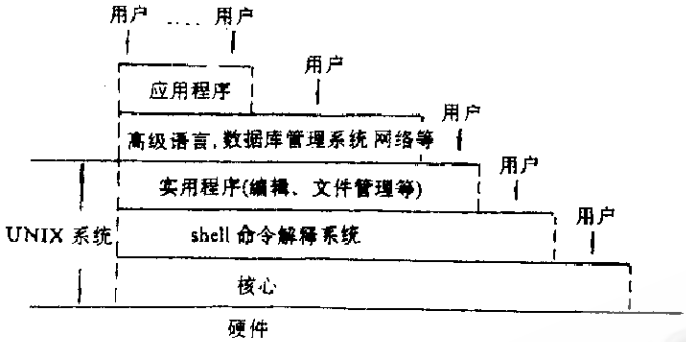

图 2 UNIX 系统基本结构

要做到中英文完全兼容，就必须对UNIX的整个系统进行必要的扩充和改造。简略说来，这些工作包括:

(1) 基本输入输出和核心改造

UNIX系统中汉字的基本输入输出主要涉及到系统核心。如果要把字典字库及其查找程序的全部或一部分放到内存中，则必须对原核心进行较大的改造。如果由硬件设备即通过加接汉字终端和汉字打印机来完成外码到内码以及内码到字库点阵的转换工作，则核心改造所要作的仅仅是解决在个别UNIX变种上由于UNIX核心中字节的高位在输入/输出时常常用作奇偶校验以及特有的延迟标志所引起的矛盾。

(2) 命令解释系统shell的改造

在UNIX上用户使用各种软件和工具都以shell命令调用形式出现，普通用户首先面对的就是shell命令。此外，UNIX中的shell不仅仅是一般的命令解释系统而且是一种程序设计语言。要使UNIX在用户面前呈现的是中英文兼容的系统，就必须对整个系统包括shell解释子系统和有关的命令本身进行扩充和改造。

(3) 各种实用软件的汉化

UNIX上有众多的实用软件，如电子邮件系统、正文编辑系统、目录和文件管理、状态询问和控制等等。这些实用程序中有些在shell汉化之后已经具备了处理汉字信息的能力。但另外一些则必须作较大的改造和扩充，比如与屏幕控制有关的编辑程序vi、ex、edix等。由于汉字是以两字节为单位进行处理的，不同于单字节ASCII码的处理；而且屏幕操作还涉及到许多有关终端的控制特性，所以要实现汉字处理其改造工作量就很大。

(4) 外层软件的汉化

如前所说，通常UNIX的外层还包据高级语言、数据库管理系统和网络通信系统等等许多软件。这些软件都要在某种程度上进行扩充和改造才能满足汉字处理的需要。比如语言中要能处理汉字字符串，或者由于

与数据库管理系统的密切关系，用户还要求变量名等标识符汉化，这就要求改造语言编译程序，使之能接收汉字标识符。

此外，国内企事业管理和办公自动化广泛地利用数据库管理系统和网络通信系统。为了能在数据库系统中使用汉字数据库名、记录名、域名，或在其中处理汉字字符串，以及在网络通信系统中可以处理汉字信息，等等，都需要对原英文版本进行改造。

另外，电子图表、字处理系统等等也是重要的应用程序，需要汉化以适合一般商业、计划或管理人员使用。

§3 不同实现方案

3.1 方案产生的依据

我们在UNIX 中文处理系统的实现过程中曾采用过多种不同方案。这些方案的产生取决于硬件环境、软件版本及运行条件，同时还要考虑到厂家及用户的要求。

国内到目前为止，运行UNIX 的硬件系统主要是超级微机、小型机、超级小型机和工作站。如，以Intel 芯片为CPU 的16 位机、IBM PC/AT 及与之兼容的286 机、国产0530 机；32 位机则包括各种386 机和IBM PS/2-80 以及与之兼容的国产0540 机。以Motorola 芯片为CPU 的M68000 型及M68020 型超级微机和32 位工作站。以AT&T WE32xxx 芯片构成的32 位3B 系列机。以PDP-11 为代表的小型机和以VAX 为代表的32 位超级小型机。

在这些机种上运行的UNIX 及其变种的版本主要有AT&T UNIX V7、System III 和System V, Microsoft 和SCO 的XENIX, BSD, Sun OS 和DEC 的ULTRIX, 等等。

实现UNIX 系统中文处理系统时，有的厂家及用户只要求接中文终端和中文打印机能输入/输出汉字信息即可。有的厂家和用户则要求接汉卡。还有的要求不添加任何硬件，完全通过软件方法来处理中文信息。

而从实现中文信息处理系统的角度来说则可以把硬件主要分为两大类，一类是AT、286、386 及其兼容机和图形工作站，另一类是普通超级微机(主要是以Motorola 芯片为CPU 的)、小型机和超级小型机。前一类硬件系统有一个特殊的主控台或者是高分辨率图形监视器，它们在系统核心中单独有一套驱动程序，不同于一般的终端驱动程序，因而不能用普通终端代替之。这样，对于第一类硬件中的主控台其汉字输入及显示的变换就必须通过系统核心来解决，而不能由汉字终端简单地代替。

解决第一类硬件系统中的主控台或图形监视器中文输入及显示问题，有的厂家或用户要求采用汉卡硬件形式，而有的则希望完全由软件方法进行。

至于软件版本，第一类硬件系统中以Microsoft 及SCO 的XENIX 以及图形工作站的软件版本(如Sun OS) 为代表。而第二类硬件系统中运行的主要是AT&T UNIX 系统。

3.2 实现中文输入输出的两类方案

从2.2中可知，UNIX中文基本输入/输出(包括显示)，主要涉及核心层。核心层上面的shell层实用程序和通用实用程序层的中文处理主要涉及汉字内码表示与原系统的冲突和矛盾问题，而与汉字的输入变换及输出变换没有直接联系。不管采用何种方案，这些层次的中文信息处理都是一样的。

针对中文的输入/输出变换这一基本问题，主要的实现方案有两类，一是加接汉字终端或汉字打印机，完全由硬件完成。另一是在核心中作改动由软件实现。

汉字终端的基本工作原理是一方面把汉字的各种外码输入直接由硬件变换成原先定义好的、系统可接受的等长内码传送给系统，另一方面是把内码变换成相应的汉字点阵在显示屏上显示出中文字符来。汉字打印机的基本工作原理是把内码由硬件变换成相应的汉字点阵在打印机上打印出汉字字符来。所以它的工作原理与终端的屏幕显示类似，只是一是显示屏，另一是打印机而已。

由于汉字的输入变换和/输出变换全部由硬件完成了，所以系统内核几乎不作什么改动或改动很少。这种方案我们称为“接插兼容”式或“外部设备级”。

但是正如前一小节所说的，在有些硬件系统中主控台或图形监视器不是普通终端代替得了的。为了在主控台或图形监视器上输入/显示中文，就必须通过改造UNIX核心来解决。即需要把中文输入/显示中文的模块放在核心中。这种方案我们称为“软件改造”式或“操作系统级”。

对于这类方案还可以细分为二种方法:

一种是采用汉卡加软件的方法。采用这种方法，可以把输入转换和输出转换用的数据(字典、字库)甚至把相应的程序都放在汉卡上，全部用硬件实现。也可以把数据(这一般是定死的)放到卡上，而把程序放在核心中，即部分用硬件，部分用软件实现。若全部用硬件实现，则要解决原有软件与新加硬件之间的协调和连接问题。如果部分还是用软件实现的，则要解决这部分与新加的部分硬件，以及这部分软件与原系统软件之间的连接和协调问题。

采用汉卡也往往存在一些问题。比如，汉卡往往很难做到使DOS软件及UNIX(XENIX)软件完全协调一致，以及与英文系统完全兼容。此外，汉卡生产厂家不一，因此一个厂家生产的汉卡，往往只能用在一个厂家的产品上。不能保证软件的完全可移植性。

另一种是采用全“软”方法。即不但是汉字处理程序而且连数据都由软件处理。这种方法的好处是可移植性极好，完全摆脱具体汉卡的限制。缺点是核心空间大增，如果算法不好则还对速度有较大的影响。好在这只是与主控台有关，对连接在多用户卡上的其他用户没有什么太大影响。

3.3 两类方案的比较

下面我们从几个方面来比较两类方案各自的长处和不足。

1. 系统自身的负担

由于第二类方案在操作系统中或多或少要加入有关程序和数据，所以在存储量上提出了额外要求(30kb–500kb)。此外，字典字库的运行也占用了系统的时间。如采用第一类方案，汉字的基本输入和输出全由设备来完成，则在内存及外存方面基本没有提出更多要求。而且由于字典字库查找处理由设备来承担，主机运行也不受什么影响。

2. 实现的难易程度

显而易见，第二类实现，要求开发者对原系统的结构，功能有十分清晰的了解，并掌握机器代码、汇编代码和相应的修改手段，所以实现方面具有相当的难度。而如由设备来完成有关处理就没有上述众多的要求。

3. 必要性

以上两点都说明第一类方案的优越性，特别是随着半导体芯片价格的急剧下跌，国内各类汉字终端和汉字打印机的问世，在经济上已体现不出第二类方案的优越的情况下，第一类方案就更有吸引力。但是第二类方案的实现仍有必要性。

综上所述，第二类方案具有较大的灵活性，能满足已有西文设备及特殊要求的用户之需要；第一类方案，系统本身负担小，实现容易，能较快较好地满足普通用户和一般用途之需要。

§4. 主要实现途径和技术

前一节我们介绍了两大类实现方案，但是具体实现时仍可以有或者需要有不同的途径和技术。这是因为，即使解决了中文的基本输入输出，还只是第一步，还必须解决核心以上各层的中文处理问题。事实上如果整个UNIX系统再加上系统原先提供的高级语言编译、DBMS及网络通信等软件完成了相应改造，那么用户在上面构造的软件就都具备了中文处理功能。

除核心层外，UNIX的其它层次及上面的语言编译、DBMS及网络通信等软件并不直接处理汉字的输入及输出转换，所以问题就比较简单。概括说来就是如何让原有的系统能识别中文的内码表示(如两字节高位置1或别的多字节内码)，避免与英文系统的原有表示产生矛盾。所以采用的具体技术与改造核心所采用的技术有不同之处。

4.1 核心外层软件的改造技术

UNIX上层软件的改造问题相对来说比较单一。如果有源码(一般是C语言程序)，则任务就简单得多。只要分析清楚其结构、数据流向，对与中文处理有关的模块进行重点分析和改造，即可完成其工作。但如果手中只有机器目标码，困难就要大得多。一般来说，要分析结构、了解数据流向，采用静态模拟和动态调试等手段对有关模块进行改造。由于工作

是在反汇编得到的汇编代码上进行的，所以比较艰巨。如果连动态调试和跟踪工具都没有，则工作就更困难了。

从图2中可以看到，作为一个完整的中文处理系统，中文高级语言、中文数据库管理系统、中文网络通信软件等是必不可少的。这样即使UNIX系统(包括核心，shell及实用程序)可以得到源码，基本中文处理系统的开发可以在源码上进行，但是由于上述的语言、DBMS、网络通信系统来自于不同厂家，不可能都能得到源码，所以免不了要从目标码着手。上层软件和具体开发手段要因系统而异，一般方法见[1-5]。

4.2 核心改造技术

1. “打补丁”的方法

在对UNIX进行中文系统开发的早期，中文终端还不普及。用户和厂家都要求我们把中文输入/输出处理模块(图1中的(1)和(3))放到系统核心中，完全由软件来完成中文的基本I/O。此时，我们所面对的核心又只是一个经过连接装配后的可运行目标码。所以我们采用的基本方法是“打补丁”。即，第一步，通过反汇编代码来了解系统结构，弄清数据流向，重点解剖与中文处理有关的模块。第二步，用C语言编制中文输入/输出转换模块。用C语言编写的好处是便于修改的系统进行调试、修改和测试，也便于移植。第三步，把这些C语言处理模块单独进行编译加工，然后人为地用若干跳转指令和原系统连接成一个“整体”。第四步，对这种经过扩充和修改的系统进行调试、修改和测试最终得到一个可以处理中英文基本输入输出的系统。我们加入的这几个模块就戏称为“补丁”。采用这种方法时，核心空间大小受限、权限受保护、“补丁”的加入、输入方法的动态转换、字库动态淘汰等一系列问题曾给我们带来许多困难。具体实现详见[6, 7]。

2. 系统扩充和重新生成的方法

有些系统，特别是XENIX系统，通常运行在IBM PC/AT、286、386及其兼容机上。为了实现主控台的中文输入/显示，我们需要把中文处理模块的全部或一部分加到核心中。对于这种实现，即使采用汉卡，也往往需要对核心进行改造。当然这种改造就不是采用上一种“打补丁”的方法了。因为这种系统的核心，虽然没有提供源程序，但提供了.O文件，这些文件是相应.C文件编译后还未连接装配的目标模块。

在通过反汇编，人工改写还原的方法得到了系统核心中那些与中文处理有关的模块的源码，并经过了再三的测试后，我们就在C语言基础上对这些模块进行相应扩充，加进中文处理功能，但保证其对外接口(调用关系、参数数量、类型等)与原英文系统完全一致。此时我们只要把这些模块加以部分编译，就可与系统中的其它.O模块一起用make工具构成一个可处理中文信息的新核心。经过反复测试、修改，总可得到一个完整、正确、并扩充了中文处理功能的新核心。由于工作基本都是基于C语言源码完成的，所以系统可靠，而且可移植性极好。目前我们开发的

286、386 机上的中文系统基本上采用此方法，详见[8, 9]。

3. 采用源码的方法

前面所说的几种方法都是在目标码或基本没有源码的条件下所采用的。当我们与AT&T达成协议在UNIX System V上开发中文应用环境CAE时，我们的工作就全部建立在C语言源码基础上了。核心上层的shell及实用程序当然可以在源码基础上加以改造以解决汉字内码与原系统的冲突、元字符、汉字表示日期时间等的习惯以及在编辑时一个汉字跨越两行等问题。

对于核心，由于系统提供的STREAMS技术，从而使中文输入输出功能的加入来得比较方便，而且比较能保证正确性，详见[10, 11]。

§5. 结束语

一九七九年我们开始对UNIX进行了分析研究，对UNIX整个系统的全面而透彻的了解，为我们的开发工作打下了坚实的基础。

本文讲述的UNIX系统上多年来开发中文信息处理系统的原理、方案和具体实现技术，这些思想和方法是经过了实践考验的。

我们开发的所有这些系统都已成为完整的软件产品。其中有的已转给国内计算机厂家移植并装备到国产机0530、0540上，有的成批转让给国内一些部门。许多采用了我们的系统的单位已取得了明显的经济效益和社会效益。

我们在UNIX的研究和开发工作得到了国内许多领导和专家的鼓励，先后获得国家，省部级和科学院的奖励。

也因为我们在UNIX系统上进行的卓有成效的开发，使我们与美国、日本、西德、意大利、新加坡、香港等国家和地区的许多计算机厂商、大学及研究机构建立了紧密的合作关系。

由于UNIX系统的微、小、中、大，甚至巨型机上的普及，并且将成为操作系统的国际标准，因此在中国推广UNIX的使用将有极大的意义和价值。然而，在中国要使UNIX广泛使用，中文信息处理又是首先要解决的。虽然，目前国际上正在开展UNIX国际化工作，目的就是要为UNIX支持各国的自然语言处理打下一个良好的基础，但是真正要让UNIX适合中国国情，还是需要我们自己进行开发。

UNIX系统的开发对我国计算机事业的发展，对汉字信息处理及系统软件国产化有着重大意义，我们愿同广大用户、有关专家一起为完成这一使命而作出自己的贡献。

参考文献

[1] 孙卫国，孙玉方，吴健，UNIX系统shell的汉化改造，《软件产业》，1988年第12期。
[2] 陈一清，中西文兼容的屏幕编辑程序的实现，《计算机研究与发展》，24:3, 1987。
[3] 汪木兰，周晓萱，曾显满，一类机器上高级语言的汉化，《软件产业》，1988年第12期。

[4] 杨建平，中西文兼容的C-INGRES 的实现。《中文信息》，1986, 第三期。
[5] 汪木兰，周晓萱，曾昱清，中英文兼容的INFORMIX 数据库管理系统的汉化，《小型微型计算机系统》，1986 年第11 期。
[6] 孙玉方，中英文兼容的C-XENIX 系统总体设计及实现，《计算机学报》，9:4, 1986。
[7] 孙玉方，李有志，汉字处理与XENIX 核心的扩充和改造，《计算机研究与发展》，24:2, 1987。
[8] 孙玉方，陈一清，吴健，郑蕾，曾昱清，PC/AT 及其兼容机C-XENIX 总体设计及实现，《中文信息学报》，3:1, 1988。
[9] 吴健，孙玉方，曾昱清，胡先祥，文强，PC-AT 及兼容机上C-XENIX 核心的实现，《计算机工程与应用》，1989 年第1 期。
[10] 孙玉方，杨建平，陆拓实，郑蕾，UNIX 系统国际化与中文应用环境，《中文信息学报》，待发表。
[11] 孙玉方，杨建平，陆拓实，郑蕾，UNIX 系统中文应用环境的开发，《计算机学报》，13:6, 1990。