国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
一九八五年第四期　总第五十二期　No. 4 1985 Sum. 52

汉字信息处理技术的发展趋势

王鸿谷

摘　要　本文讨论了中文信息处理技术发展中的几个问题，特别是关于中文信息处理系统的体系结构方式问题。这些问题是与用户密切相关的。

一、引　　言

随着计算机应用在国内逐步推广，特别是从单纯用于科学计算扩展到各式各样的非数值应用领域。计算机处理中文信息的功能越来越受到人们的重视。人们认识到：没有汉字功能，计算机就无法在我国普及应用，自从1974年国内有计划地正式开展“汉字信息处理系统”的研究以来，无论是在基础理论方面，还是在输入技术、码制研究、中西文兼容技术、设备研制生产和标准化、系统构成等方面都取得了迅猛的发展，使得日本、香港、台湾、美国从事这方面研究的人士不得不刮目相看。

二、汉字的基础理论研究

国内开展汉字信息处理研究的初期就开始了基础理论方面的研究。比如，不但对汉字的使用频度，而且对统计频度的方法进行了科学的研究。这从量的概念上为汉字信息处理系统的设计奠定了基础。对于汉字字形、字元（字根）、笔划的研究不仅为汉字的笔形、字形类编码输入方案提供了理论分析，也为汉字自动识别技术的发展作了有益的准备工作。对汉语语法、语音、语调、上下文关系等方面的深入研究为语音输入、汉语自然语言处理等工作打下了基础。从1983年“中文信息处理国际研讨会”的论文来看，约三分之一是属于基础理论。这说明了该方面的研究工作绝不是可有可无的。

比如，对于汉字使用频度的科学统计表明，国家标准（GB-2312-80）汉字集的使用频度：一级汉字（3755个）已达到99.96%，加上二级汉字（3008个）已基本达到一般的通用要求。

这些统计给2610任务中的汉字库设计提供了依据。2610任务中汉字库采用三级库管理。这是因为86/330A系统虽然提供了较大的寻址空间，但既然一级汉字使用频度已达99.96%，且2610任务的使用环境比较专门化（汉字域不大），因此就没有必要将二级汉字也存入内存。为解决少数二级汉字调用的快速响应问题（二级汉字放入磁盘，调用时耗费访盘时间较多）以及个别一、二级汉字以外的字的使用，可以建一个活

一九八五年元月五日收到

字库，再加上造字功能（能造出一、二级以外的字）。通过动态的频度统计，可以将当时使用得较多的一级以外的字（包括造的字）保存在一个文件中，关机时进行优选，开机时文件进入内存。这样既可大大节省内存开销，又可以保证使用少数一级以外的汉字时不致耗费过多的访盘时间。

目前，在汉字基础理论方面的研究正朝着系统化、方法上的科学化以及紧密联系应用的方向深入发展，在汉字“结构理论”和“音韵理论”方面的研究有可能使汉字信息系统的技术为之一新。

三、输 入 技 术

目前汉字输入技术正沿着键盘输入、字形输入和声音输入三个方向发展。

键盘输入一般又粗略地分成整字法（如汉字大键盘）和编码法。现已有400种以上的编码方案，其中上机通过试验的并已正式采用实施的有几十种。它们大致分成：(1)流水码（如电报码、气象码等），(2)拼音码，(3)音形码，(4)字形码（字根、笔划等），还有各种混合方式。从键盘的键数看有各式各样的大、中、小键盘。由于汉字基础理论的研究以及逐步建立了对于编码方案好坏的科学考察原则，现在编码技术方面正在朝更加科学化、高效化的方向发展，并在逐步进行优选。

字形输入是图形识别的课题之一。目前，对于印刷体的字形识别已取得很大进展，有了一些实验性成果；对于手写体的识别则处于摸索阶段。在字形输入设备方面目前常用的有：(1)电视摄象机加上 A/D 转换；(2)激光扫描；(3)基于固体图象器件的输入设备。虽然因基础理论研究得不够和输入硬设备工艺、技术、性能上不去，因而字形输入技术的发展遇到了重重困难，但是它的优越性、生命力是不容置疑的。对某些专门化的大量处理汉字的系统来说，它是解决高速输入问题的唯一途径，也是解决未来智能计算机视觉问题的一项指标。日本在1977年研制成功每分钟读入 6000 个印刷汉字的 OCR 装置。国内近年也取得了每秒 3 字的成果，以及用象限端点和用转动惯量特征识别 6000 汉字的技术。可以相信，在加强汉字基础理论，图形识别理论研究的基础上，我们依靠作为汉字的主人对汉字深切理解的优势，奋发努力，是一定能在这方面取得很大进展的。

声音输入对用户来说是最方便不过的了，但实现它难度也很大。目前国内仅有少数几个单位开展这方面的研究，有的在试制语音识别装置，以及对特定的人有限的话的声音识别方面取得了一些初级的成果。我们在看到困难的同时，也要注意到汉字具有单音节易于识别这一优越性，只要坚持努力，是能够赶上英文或其它西文的声音输入处理技术的。

在2610任务中，由于使用环境要求可靠、准确、且要考虑到经济性，因此，目前只能考虑键盘输入方案（当然，也可以是主机传来的汉字信息）。又由于要求简便、易学且考虑到多种使用人员及使用場合，因而在8086主处理器上配制了大键盘、系统键盘，在 ANC 选件中配置字符键盘和中汉字键盘。除了整字输入而外，在编码输入方案中我们先选用了 GB 码、拼音码（辅键选同音字）和26键 5 笔字形编码输入方式。此外，我们还留出两个口，等待采用将来优选的更好的输入方式。因为各输入处理程序皆模块化互相独立，因此增加一种或去掉一种方式将会是很方便的。

四、汉字系统结构形式的研究

目前，我国汉字系统的结构形式逐渐发展成三种形式，也可以说发展的三个阶段，即通常所说的（从虚拟机的角度看）应用级，系统级和设备级。

所谓“应用级”就是利用原设备和系统软件，作不大的改动，而主要通过在应用级软件上建立有关汉字信息的I/O驱动程序来实现汉字的输入输出。这种方式系统效率低，使用不方便，常常造成一系列矛盾。因此这种初期发展的方式现已面临被淘汰的趋势。

所谓“系统级”就是在操作系统一级上作修改。将操作系统的内部I/O驱动模块中增加一些处理汉字I/O的驱动程序，从而解决了中西文兼容问题，提高了系统效率，如CP/M，PC—DOS、RSX—11M等著名的操作系统都已完成了扩充和改造，UNIX的改造、扩充也取得了初步的成功。

所谓“设备级”就是主机只以约定的代码与设备通讯，而将汉字信息的I/O及其它一些处理工作统统放到设备中去进行。这种设备现称之为插接兼容式汉字终端（或其它汉字I/O设备）。需要汉字功能时，将它们接上，不需要就拆开。粗略地讲，只要主机和设备间在接口、机内码和控制码三方面协调一致，这种设备可以方便地和任何主机相連。显然，这样就大大提高了汉字系统结构形式的水平。

“应用级”系统结构无疑是落后的，但是对于“系统级”和“设备级”这两者的关系应怎样看待，人们却争论不休。这里不妨也作一点粗浅的探讨。

(1) 系统级把字库放在主机内，主机花大量时间忙于字形处理，而输出速度最多只有单纯输出汉字代码的十六分之一。其次，通过查表方式进行输入码转换，常常要耗费大块的内存空间。这显然对于主机的效率是不利的。将字库、换码表，处理程序放到外设去做不但是完全可能的，而且系统总效能肯定要高得多。因此，所谓“不扩充操作系统的就不是‘眞正’的汉字系统的”说法是一种以静止的观点来看待事物发展，是站不住脚的。

(2) 修改操作系统是一件不容易的事。各机种常常采用不同的操作系统，一种操作系统又常常修改(甚至有时不同版本差别很大）。因为修改操作系统又常常产生“链式反应”，要进一步修改其它相关软件，所以，在计算机技术的发展日新月异的时代里，着重发展研制“设备级”的插接兼容式汉字处理外设，显然能处于较稳定的状态和较主动的地位。

(3) 按插接兼容的思想，对于一些新的汉字I/O设备（例如声音输入设备，字形输入设备）的研制无疑会带来方便，会更有助于技术突破。

(4) 当然，插接兼容技术也面临一些困难。例如：不解决接口、机内码、控制码的协调乃至标准化问题，就不可能做到眞正的插接兼容。但这些困难是可以克服的。一旦突破，就会更加显示出插接兼容式汉字系统的优越性。只要抓紧工作，提高质量，将这类设备产品打入国际市場也是可能的。

(5) 另一方面，我们也要重视研制几种所谓“系统级”的设备，尤其是微型机。这是因为目前很多单位把微型机用于事务管理。这时使用汉字多。无论是作为单独的个人计算机或作为网络终端机，使用者希望设备件数少和所有信息显示在一个屏上。他们还

希望系统在这种应用环境中取得高效率。这时所谓“系统级”设备显示出优越性。因此，适当研制几种中西文兼容的“系统级”是必要的。目前这方面已做了不少工作。如果能全面规划，优选几种系列，就会避免浪费人力、财力的状况。

(6) 目前，有些单位开展的终端仿真工作，即对某些先进的国外终端设备进行仿真，保持原西文操作的全部功能，再加上汉字功能，这也是很有意义的。虽然，终端仿真的着眼点在于针对某些先进的终端增加汉字功能，而插接兼容的思想着眼点在于克服主机控制码不同的困难、试图研制出一种能与多种主机联机使用的汉字终端设备，但是，它们都是在终端方面做工作，使终端承担大量的中文信息处理。当前，在学术上有些争论，我个人认为：这两者是相辅相成的，而且主要的思想（在外设上做文章）是一致的。在显示终端方面，两者都会有发展，以满足不同的需要。而在打印和将来的图形输入，声输入设备方面，插接兼容式的汉字外设无疑会有更广阔的前途。

(7) 有种意见认为：对“系统级”的评价值得商榷，因为随着超大规模集成电路的发展，汉字完全可以象西文字符发生器一样生成，可以放在打印机设备内或CPT控制板上，因此“系统级”不一定要“把字库放在主机内，主机花大量时间忙于字形处理”。他们认为，“系统级”的特点主要在于从操作系统改造入手，做到中西文兼容。

这种意见是有道理的。但我们的考虑是：

首先，这种意见涉及到对目前汉字系统结构形式的划分问题。按目前汉字信息处理系统的发展实际看，似以按本节前面所述的划分为宜。当然，随着技术的推进，这种划分是可以改变的。

其次，从这些年来汉字系统及汉字终端研制的经验看，既然已经用硬件完成了汉字库或汉字发生器，那末增加一些通讯、处理、控制功能就可以实现很好的插接兼容式汉字终端设备。而在操作系统的改造上作文章，就会有前面提到的一系列弊病，利少弊多。

这方面，我校研制的HZD—7032型汉字终端就取得了很好的经验。他们在分布合理的多CPU支持下实现了以硬件为主体的汉字处理系统，功能强，联机方便，通信速率可在较大范围内选择。

概括地说，基于“设备级”思想的插接兼容式汉字设备在系统构成思想上是先进的。随着CPU、EPROM，ROM及其它硬设备价格的降低，随着存贮体的进一步微型化，其优越性将更加明显，应该重点发展。然而，根据一些实际需要，我们也应该优选发展几种“系统级”的汉字微机，不断提高性能、质量和可靠性，满足相应用户的要求。

五、网和库的汉字功能

计算机应用目前已发展到了联网（各种远程网、局部网）建库（数据库）的阶段。机、网、库联成一体，这是计算机应用的有效形式。汉字信息处理系统的研制也自然在向网和库的方向推进。

目前，国内已在若干中、小型网络上实现了汉字信息通讯。2610任务中在显示处理器和ANC处理器之间设计了汉字通讯处理功能，也在考虑与主265机间的汉字通讯。这既是2610任务使用环境的需要，也是向实现网络汉字功能迈步。当然，完成高层网络软件

的汉化有大量工作等待着我们。

汉字数据库的情况就困难一些。对“汉字数据库”有各种理解。但现在应该统一到多数人的理解上来，即“能存放汉字信息（或具汉字功能）的数据库”。实现这样的数据库不但要增加很多处理汉字信息的功能模块，而更加棘手的问题是：怎样节省存贮空间？怎样保证在解决上述问题时不过份地降低系统效率？等等。

尽管面临各种困难，但考虑到：(1)建库联网几年以后会是势在必行。要想保证2610任务实现后保持较长一些时间的先进地位，在二期工作中非上数据库不可；(2)实际使用环境也是需要的；(3)由于86/330A系统实时多任务操作系统可以方便地配置，因此建数据库是有实用价值的。因此，我们决定在2610任务的二期工程中为系统配置具一定水平的汉字数据库管理系统。

六、使中文信息处理技术的步伐跟上第五代计算机的发展

目前，人们已开始注意到中文信息处理技术怎样适应第五代计算机的发展。第五代计算机完全是一种崭新的体系结构，预计在1990年左右问世。目前已有一些阶段成果，国内也有人在进行这方面的研制工作。因此，如何让第五代计算机具有良好的中文信息处理功能已是一项急迫的任务。

一种途经是人们正在做的工作，即以软件手段使第五代计算机的程序设计语言（逻辑型，或逻辑函数型）汉化或使其具有处理中文信息的能力。这种方法与已经研制成功的汉字BASIC，汉字COBOL，汉字C语言相比，没有实质性的突破。

人们也注意到，中文信息处理技术，尤其是输入处理技术，很多方面是可以采用人工智能技术的。比如，自然语处理本身就是人工智能技术的一个应用领域，因此，怎样在这方面发展、应用人工智能技术，就成了一个大课题。我们粗浅地认为，大体有两种方法来进行这方面的研究：

(1) 利用人工智能技术，特别是用第五代计算机的阶段性成果来研制高性能的真正具有智能意义的汉字终端设备、“智能”输入设备（声音输入、图象输入）。现在人们已在研究，但我们需要更先进的设计思想和更高性能的硬件。当然，这里还要注意插接兼容思想，使之能方便地与未来各种类型的第五代计算机相连。

(2) 我们自行研制设计第五代计算机，就应在整个逻辑设计中充分考虑中文信息处理，设计出这方面的功能块。例如，将目前流行的第五代计算机的体系结构中的智能接口机设计成具有强大中文信息处理功能的机器；使数据库机能很好地适应中文信息的存贮管理，使之与整个知识库协调。这也可理解为研制一种相当于目前Von—Neumann计算机的“系统级”汉字系统的第五代计算机的汉字系统。

我们期待中文信息处理技术和人工智能技术，与第五代计算机技术更紧密地结合。

七、结 束 语

以上从五个方面概述了一下目前汉字信息处理技术的发展趋势。实际上，还有不少动向也是很重要的，例如，代码体系的研究，各种I/O设备及汉字库的不断创新，字符集、字模、代码（尤其是机内码）的标准化和标准化研究等等。由于时间和篇幅关系，

就不一一谈及了。我们相信：只要重视发展策略科学化，注重经济效益，就一定能把我国的汉字信息处理技术发展到世界前列。

参 考 文 献

[1] 1983年中文信息处理国际研讨会讨论会论文集、第1，2，3卷

[2] 何克抗等著，插接兼容技术的由来和影响，计算机世界。

The Trend in the Development of Chinese Information Processing Techniques

Wang Honggu

Abstract

This paper deals with several problems in the development of Chinese information processing techniques, especially the one of architecture of the chinese information processing system. All of these problems are closely related to the user.