孙星明　殷建平　陈火旺　著

中文信息处理技术

国防科技大学出版社

前　言

中文信息处理是汉语语言学和计算机科学技术的融合，是随着计算机从纯科学计算转向其他应用领域而兴起的一门学科。从20世纪60年代开始，中文信息处理逐渐成为我国推广应用计算机的关键领域之一。到80年代初期，围绕着DOS操作系统推出了许多中文DOS平台，它们提供了汉字输入、输出等功能。我国第一批高新技术企业，如联想、方正、四通等正是凭借其独有的中文信息处理技术而崛起的。进入90年代后，中文信息处理解决了计算机处理汉字的一些基本问题，中文已经不是计算机在我国推广应用的障碍。外国公司通过吸纳我国人才，也已掌握了基本的中文信息处理技术。从Windows 3.2开始，微软所有的操作系统内核以及一些主要的应用软件都已支持中文。因

此，中文信息处理的成就反过来却威胁着它自身在我国国内的地位。在国内，人们对中文信息处理的前途产生了忧虑。

但是，Internet 的发展又重新唤起了人们对中文信息处理的迫切需求。一方面，中文信息处理应当帮助人们更方便、更有效地利用Internet上几乎是无限的中文信息，例如：中文的全文检索、自动文摘、智能搜索、自动分类、机器翻译等。另一方面，应当使计算机或联网的信息家电与人之间的中文信息交流达到高效、智能，甚至自然的程度。而要达到这些要求，光靠对现有技术的改进是不够的，它涉及到人工智能、知识库、数据库、模式识别、Internet网络技术以及汉语言文字学等领域知识，需要这些相关学科研究者的共同努力才能取得突破。

中文信息处理可分为基础研究和应用研究两方面，人们往往容易忽略基础研究，但恰恰是基础研究在很大程度上决定了中文信息处理今后的前途。中文信息处理基础研究包括汉字研究、汉语词汇研究、汉语语法研究、汉语语

义理解研究、汉语篇章结构研究、汉语语料库建设等。这些工作既复杂又繁重，是中文信息处理的薄弱环节。

作者在中文信息处理领域进行了一段时间的研究，在该领域积累了一定的研究成果，本书将作者的一些研究成果进行整理、分类，其中也引入了作者的不少新思想、新方法和新技术。我们希望本书作为中文信息处理的基础研究能弥补中文信息处理的有关薄弱环节。

本书是国家教育部科学技术研究重点项目，国家教育部骨干教师资助项目和湖南省自然科学基金课题的部分研究成果。感谢国家教育部对课题“汉字结构知识的表示、挖掘及其应用研究”和“面向特定领域的软件开发方法与环境”，以及湖南省自然科学基金委员会对课题“开放式软件中汉字信息的深层次处理方法研究”给予的资助。

在本书的写作过程中，许多老师、同事和同学提出了不少有益的思想，有些同事和学生协助编写了实现本书部分思想的源代码。值本

书完成之际，作者谨向他们致以衷心的感谢！

虽然作者在本书中提出了一些“新”的思想、方法和技术，但是，这些还不一定很完善，我们衷心希望有更多的同行与其他各界读者多给我们提出宝贵的意见和建议，帮助我们进一步完善已有的研究成果。

作者

目　录

第一章

汉字结构知识的表示

1.1　引言

汉字是一种非字母化、非拼音化的象形文字。尽管汉字是由横、竖、撇、点、折等有限种笔画写成的，但因为汉字的笔画既没有确定的大小和组合方式，也没有固定的表音或表意内涵（如“己、已、巳；日、曰；未、末”），汉字的笔画不相当于西文的字母。

汉字的最小构成单位是笔画，由数十种笔画组成了数百种结构块，再由这数百种笔画结构块按照一定的方位关系组成了数万个汉字，我们将这些笔画结构块称为“部件”。不同数量、不同功能的部件依照不同的结构方式组合形成汉字。部件的数量、功能和组合方式（位置、置向、交接法）是每个汉字区别于其他汉字最重要的属性，汉字的信息量主要由部件及其组合来体现。因此，一般说来，汉字结构可分为三级：笔画、部件、整字。

纵观汉语与英语的区别，英语处理的便利就在于所有英语单词都可以由26个英文字母按前后关系拼成，而汉字是非字母、非拼音化的文字，很难找到一种方法用一些类似于英文字母的元素来表示汉字。围绕这一问题，人们提出了各种解决方案。如，在汉字键盘编码输入方面，王永民教授发明五笔字型输入法，就是企图用一些部件来组合汉字；在汉字字形设计和汉字识别方面，很多学者，如加拿大的C.Y.Suen[LI 1991][LI 2000][LIU1999][SUE1999]，台

湾的 C. W. Liao, J. S. Huang, L. H. Chen, J. R. Lieh[LIA1990][LIA1991][CHE1990], C. T. Chuang, L. Y. Tseng[CHU1995][TSE1992], 大陆的吴佑寿、张炘中、夏莹、董韫美、樊建平、赵明、姜珊[夏 1985][夏 1986][张 1987][赵 1990][吴 1990][董 1996][樊 1990][ZHA1990][姜 1995]等，就汉字的表达提出了很多有益的思想，都企图将汉字用数学方法表达出来。他们提出了汉字的有向图表示，汉字的属性关系图，汉字的相关属性关系图，汉字的二维扩展文法属性，汉字的层次模型，汉字原形方法等。张炘中、夏莹等人还明确提出了汉字表达式这一概念[夏 1986]，他们将横、竖、撇、捺、左折、右折、方、叉作为汉字基元，并扩展 BNF 范式的元符号，将汉字表示成汉字基元的数学表达式，对提高汉字的识别率和识别速度起到了一定的作用。上述方法中有的将横、竖、撇、捺等笔画作为汉字基元，有的将汉字部件作为基元，将汉字结构以基元及基元之间的相互位置关系表达出来，在汉字识别的部件分离及笔画抽取方面可以起到一定的指导作用，对提高汉字的识别率和识别速度起到了一定的作用[夏 1986]。但因为当时考虑汉字结构的数学表达的主要目的是为了识别汉字，故对有关细节考虑的比较多，使得这些有关汉字的表达方法的数据结构都比较复杂，参与运算的数据及描述数据间的关系也很复杂，因此它们绝大部分没有很好地用于汉字识别以外的中文信息处理领域。

由于汉字的信息量主要由部件及其组合来体现，将组成汉字的部件拆分出来是中文信息处理中首要的基础工作，也正是由于它的重要性，人们在相当长的时期内，自发地进行了这一工作，计算机形码的编制出现了万“码”奔腾的局面。尽管在部件拆分方面出现了如此火爆的场面，但将每一个汉字都表示成由部件组成的简单的数学表达式方面的工作还没有特别成功的报道。

在本章中，我们提出了一种全新的汉字的数学表达方法，即将汉字表示成由汉字部件作为操作数、运算符号为部件间结构关

系的数学表达式。我们选定了505个部件，部件间可通过位置相互组合生成汉字，这种相互位置关系即为运算符号。我们拟选定6种运算符号：lr，ud，ld，lu，ru，we，它们依次表示左右，上下，左下，左上，右上，全包含等关系。这些运算符有一定的优先级，括号优先运算。由此可见，这种表达方法非常接近自然，结构简单，而且可像普通的数学表达式一样按一定的运算规则处理。将汉字表示成数学表达式以后，对汉字的处理方法就可接近对英文的处理方法，从而中文信息处理的很多方面将变得比以前简单。

令$\Omega=\{$汉字$\}$，即Ω表示所有汉字的集合，一般指国标一、二级汉字库中汉字的集合；Θ表示汉字基元组成的集合；Ξ表示描述汉字基元之间的相互关系所组成的集合；(Θ,Ξ)表示Θ中的汉字基元通过Ξ中给出的关系组成的汉字或非汉字图形的集合。从文献［夏 1986］可知：

引理 1.1　设$\Theta=\{$横，竖，撇，捺，左折，右折，方，叉$\}$，$\Xi=\{$笔画之间的相对位置关系（包括角度、位置坐标等）$\}$，则有：$\Omega\ (\Theta,\Xi)$。

另外，显然有：

引理 1.2　设$\Theta=\Omega$，$\Xi=$（空集），则有：$\Omega=(\Theta,\Xi)$。

引理1.1和引理1.2描述了两个极端情形：

1）取笔画作为基元，Θ简单，但Ξ非常复杂；

2）取所有汉字作为基元，Θ复杂，但Ξ非常简单（可以为空）。

因此，我们完全有理由尝试是否可以适当选取Θ，Ξ，使得它们都比较简单，而且$\Omega\subseteq(\Theta,\Xi)$，即$(\Theta,\Xi)$能够表示出所有汉字。本章的研究就是围绕这一问题展开。

因介于笔画和汉字之间的就是部件，因此本研究的关键是要合理地选取部件作为汉字的基元使得Ξ比较简单。

1.2 汉字部件的选取

1.2.1 部件选取的现状简介

汉字的信息量主要由部件及其组合来体现。将组成汉字的部件拆分出来是中文信息处理中首要的基础工作，也正是由于它的重要性，人们在相当长的时期内，在不同地区、不同信息处理系统中自发地进行了这一工作，如在计算机形码方面就出现了万"码"奔腾的局面，部件拆分的不规范现象也逐年增多，给文字使用和信息处理带来了混乱。这种状况既不利于计算机应用的发展，也不利于语言文字的统一规范，同时给计算机教育和识字教育造成了很大的困难。为此，国家语言文字工作委员会于 1997 年 12 月颁布了《信息处理用 GB13000.1 字符集汉字部件规范》[汉1998][邢 2001]，以此作为语言文字规范，并于 1998 年 5 月 1 日起开始实施。该规范是国家语委针对当前汉字键盘输入中汉字部件和汉字拆分上的混乱现象，组织相关学科和技术领域的专家学者经反复论证研讨，形成制定规范的理论依据和指导原则，并通过计算机对 20 902 个汉字进行部件拆分和归纳，制定出了汉字基础部件表及其使用规则。参加该规范研究的单位有北京语言文化大学、北京信息工程学院和上海交通大学等。该规范出台后也引起了一些争论。对"部件"一直就没有一个统一明确的定义，对部件的选取也没有一个统一的规范，该规范尽管确实存在一些问题，但我们认为它提出的基本思想是正确的，对规范汉字部件将具有划时代的意义。因本研究中提出的汉字表达式只是中文信息处理的基础研究，选取部件只是系统的后台部分，部件的选取只与少数几个研究人员有关而与广大用户无关，而且我们选定的部件库本身就是动态的，随着汉字数量的增加它可以出现相

应的变化，故对部件的选取及部件规范在此不作太多叙述，对有关关键问题将在 1.6 中作简要介绍。

为保证部件拆分与归纳的科学性，制订该部件规范的基本原则是“从形出发，尊重理据，立足现代，参考历史”。按照该原则，规范中提出了“交重不拆”的规定，即部件组成字时，如果出现重叠或交叉，则不应当将重叠或交叉部分拆开，而应当将它看成一个部件。也正是这一原则，使已有的汉字形码编码方案几乎无一满足该规范，因此该规范也遭到不少人的公开批评。我们认为，当时考虑汉字部件拆分的主要目的是为了输入汉字，是为了提高汉字的输入速度，故对编码是否有重码以及码长考虑的比较多。但随着计算机应用范围的日益扩大，中文信息处理出现了一些除汉字输入以外的新的重要的领域，如：互联网上汉字信息传播、中文无线通信、多媒体教学、远程教学等。在这些新的领域，人们越来越发现以往的部件不太适用新的领域，部件需要规范。本研究开始于 1998 年，当时该规范还未颁布，直到 2000 年

图 1.1　部分部件（1）

底，我们才找到该规范。但是，我们在选取部件时所确定的“以形为主，交重不拆”的方案正好与该部件规范的基本原则是不谋而合。

1.2.2 本研究中所选部件

通过大量的统计分析，我们选定了组成国标一、二级汉字的505个部件，当汉字数量增加时，部件数量将会适当增加，估计对CJK汉字部件数将不会超过600个，所选部分部件及其编号如图1.1～1.3所示。

图1.2 部分部件（2）

1.3 运算符号的定义

1.3.1 基本记号定义

由一个以上基本部件按本节将要定义的运算得到的部件称