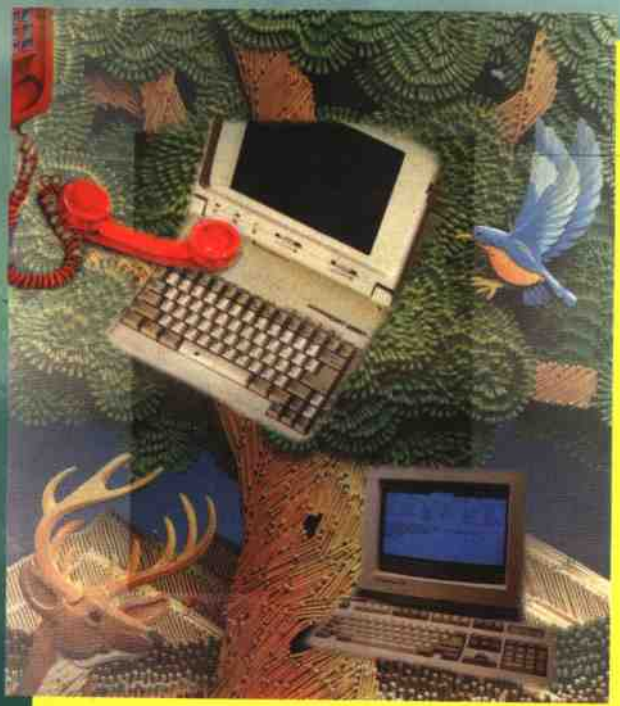
中小学生信息科学知识

汉字信息处理

主　编：李广建
副主编：邓俊强　曹正斌



北京科学技术出版社

中华万有文库

总顾问 费孝通
总主编 季羡林
副总主编 柳斌

科普卷·中小学生信息科学知识

汉字信息处理

编著 耿骞

北京科学技术出版社

中华万有文库

总顾问　费孝通
总主编　季羡林
副总主编　柳　斌

《中华万有文库》编辑委员会

主　任：刘国林
秘书长：魏庆余　和　奏
委　员：（按姓氏笔画为序）

王寿彭	王晓东	白建新	任德山
刘国林	刘福源	刘振华	杨学军
李桂福	吴修书	宋士忠	张　丽
张进发	张其友	张彦民	张晓秦
张敬德	罗林平	封兆才	和　奏
金瑞英	郑春江	胡建华	侯　玲
祝立明	贾　斌	顾士熙	高文元
游铭钧	章宏伟	常汝吉	彭松建
韩永言	葛　君	鞠建泰	魏庆余

总　序

本世纪初叶，商务印书馆王云五先生得到胡适之、蔡孑民、吴稚晖、杨杏佛、张菊生等30余位知名学者、社会贤达鼎力相助，编纂出版了《万有文库》丛书。是书行世，对于开拓知识视野，营造读书风气，影响甚巨，声名斐然，遗响至今不绝。

1000多年以前，南朝齐梁学者钟嵘在《诗品》中以"照烛三才，晖丽万有"来指说天地人间的广博万物。今天，我们全国各地的数十家出版发行单位与数千名作者以高度的历史责任感，联袂推出《中华万有文库》，并向社会各界读者，特别是青少年读者做出承诺：

传播万物百科知识，营造有益成功文库。

我们之所以沿用《万有文库》旧名，并非意图掠美。首先，表明一个信念：承继中国出版界重视文化积累、造福社会、传播知识的优秀传统，为前贤旧事翻演新曲，把旧时代里已经非常出色的事情在新时代里再做出个锦上添花。其次，表明我们这套丛书体系与内容的鲜明特点。经过反复论证，我们决定针对中小学生正在提倡素质教育的需要和农村、厂矿、部队基层青年在提高文化与科学修养的同时还要提高劳动技能的广泛需要，以当代社会科学与自然科学的基础知识为基本立足点，编纂一套相当于基层小型图书馆应该具备的图书品种数量与知识含量的百科知识丛书。万有的本意是万物。百科知识是人类从自然界万物与社会万象之中得到的最重要的收获。而为表示新旧区别，丛书之名冠以中华。这就是我们这套丛书的缘起与名称的由来。

《中华万有文库》基本按照学科划分卷次，各卷之下按照内容分为若干辑，每一辑大体相当于学科的一个二级分支，各卷辑次不等；各辑子目以类相从，每辑 10 至 20 种不等，每种约 10 万字左右，全书总计约 300 辑 3000 种。《中华万有文库》不仅有传统学科的基本知识，而且注意吸收与介绍相关交叉学科、新兴学科知识；不仅强调学科知识的基础性与系统性，而且注重针对读者的年龄特点、知识结构与阅读兴趣而保持通俗性和趣味性；不仅着眼于帮助读者提高文化素质与科学修养，而且还注重帮助读者提高社会生存能力与劳动技能。

每个时代，图书最大的读者群是 10 至 20 岁左右的青少年。每个时代能够影响深远的图书，是那些可以满足社会需要，具有时代特点，在最大的读者群中启蒙混沌、传播知识、陶冶情操、树立信念的优秀图书。我们相信，只要我们老老实实地做下去，经过几个甚至更为漫长的寒暑更迭，将会有数以百万计的青少年读者通过《中华万有文库》而打开眼界，获取知识；《中华万有文库》将会在他们成长的道路上留下鲜明的痕迹，伴随他们一同走向未来，抵达成功的彼岸。

天高鸟飞，海阔鱼跃。万物霜天，凭知识力量，竞取成功，争得自由。在现代社会中，任何人都没有任何理由拒绝为了获取力量而读书。这是《中华万有文库》编纂者送给每一位本书读者的忠告。

追求完美固然是我们的愿望，但是如同世间只有相对完善一样，《中华万有文库》卷帙浩大，子目繁多，难免萧兰并撷，瑕玉杂陈。这些不如人意之处，尚盼大家幸以教之。我们虚心以待。是为序。

《中华万有文库》编委会

目 录

第一章　汉字信息处理的研究内容

汉字信息处理是计算机技术与语言文字处理相结合的产物。早在计算机被发明之前，人类就在不停地创造、使用和研究语言文字。计算机的产生为人类研究语言文字提供了更先进的手段，使人们能够以更快的速度、更大的规模进行文字处理和语言研究，并得出更准确和更有价值的结果。

利用计算机进行语言文字处理的研究的先决条件是语言文字信息进入计算机，即计算机能够识别、接收、存储并输出语言文字材料。这是汉字信息处理所要解决的首要问题，是高层次的计算机语言信息研究的基础。一般来说，我们把能够完成汉字的输入、存储、输出的计算机处理系统称为汉字系统。

第一节　概述

一、文字信息处理的必要性

在现代社会生活中，我们每天都要说话、读书和写字，因而我们每天离不开语言和文字。语言和文字是人类表达和交流思想的基本的、也是最重要的工具。而其中，文字是语言

的书面表示，是符号化的语言和思维系统，也是信息的主要记录方式。文字的产生和使用是人类文明的重要标志。

人类日常使用的语言被称为自然语言（与自然语言相对的是具有严格逻辑形式的形式语言）。人类使用自然文字的传统形式是用手抄笔写的手工方式，这种方式效率很低，消耗了人们大量的时间和精力。随着现代社会的发展，人类需要进行的文字处理量越来越大，传统的方式显然不能满足现代社会的要求，因而，寻找新的文字处理途径就变得越来越迫切。

文字信息处理就是用计算机对自然语言文字的音、形、义等信息进行各种处理，即用计算机对有关信息进行识别、存贮、统计、检索、理解、生成、传输、控制和转换等加工操作，从而更有效、更准确地使用语言文字。简单地说，文字信息处理可分为以下三个方面：

1. 文字信息的输入

文字信息的输入的主要工作是通过输入设备把文字信息转换为数据代码形式，并送入计算机存储。这些输入设备可以是键盘、扫描仪或光笔等。

2. 文字信息的加工和处理

加工和处理是根据各种不同的应用，利用预先设计好的程序，对数据进行加工和处理，并得出处理结果。

3. 文字信息的输出

文字信息的输出将以数据代码形式存贮在计算机内的信息转换为文字形式，并通过输出设备输出。这些输出设备可

以是打印机或显示器等。

文字信息的处理不仅是人类提高使用语言文字效率的需要，也是计算机自身发展的迫切需要。自从1945年第一台电子计算机诞生以来，计算机应用从单纯的科学计算（军事科学、工程计算、数值统计），逐步扩展并广泛应用到过程控制和信息处理等各个领域，如数据统计、数据更新、数据查询、状态分析等，其应用的深度和广度不断加强。而在这些应用中，都不可避免地遇到文字信息处理这一问题。如在会计记帐系统中，就必须以清晰准确的文字向用户说明系统中所记录的内容，产生所需要的各种单据和报告。

文字信息处理的主要任务是用计算机对语言文字作各种加工，但总的来看，文字信息处理是由语言学、文字学、心理学、信息论、计算机科学、通讯技术等多门学科和技术结合而成的一门综合性学科和技术。文字信息处理是信息科学的重要分支，其发展有赖于与其相关的各个学科的发展。

二、汉字的基本特点

人类的语言文字基本上可分为两类：拼音文字和拼形文字。汉字属于后者，而目前流行的大多数文字属于拼音文字。由于拼音文字是由少数的字母通过线性排列组合而成的，如英语就是由26个字母组合而成的，所以这类文字的计算机信息处理工作相对简单，只要完成了对这些少数字母的处理，就基本完成了对整个文字集合的处理工作。例如：可借助精心设计的由基本字母组成的一个输入键盘完成文字的输入工

作；只要对每一个基本字母编了码，就可以由此生成整个字符集的编码，从而解决了文字的存储问题。与拼音文字不同，拼形文字虽然也有构成文字的基本材料，如偏旁、部首、字根等，但是这些基本材料构成文字的方式不是线性的，而是在一个二维的平面上完成的，尤其是汉字，数量庞大，笔划、结构复杂。由于拼音文字的计算机信息处理可以建立在几十个字母的基础之上，所以这类文字被称为“小字符集”文字，而像汉字这样的拼形文字不可能以一个小字符集为基础，其信息处理必须直接面对所有汉字，所以在计算机处理这类文字时，人们称其为“大字符集”文字。拼形文字所包含的复杂的构字信息为计算机处理带来了极大的困难。

关于汉字进入计算机的问题，即汉字信息处理问题，曾经存在着激烈的争论。一种观点认为，汉字进入计算机只是个方法问题，只要认真研究，一定能找到解决汉字信息处理的满意方案；而另一种观点认为，拼形文字系统完全不适于计算机处理，为计算机所增加的任何汉字处理能力，无论做得多么好，都是在时间复杂性和空间复杂性上的额外开销，因而都只能是权宜之计。真正彻底解决问题的办法是走汉字拼音化的道路。一些国外的大公司经过对汉字信息处理的大量研究后，也曾断言：要么不用汉字，要么不用计算机。

汉字源于象形文字，是人类最古老的文字之一。我国是汉字的发源地，使用汉字有几千年的历史。与其它各种早已绝迹的象形文字，如古埃及的文字和苏木尔楔形字相比，汉

字几千年来虽有变化，但一直盛行不衰，不因社会体制和地区语言的不同而变化，始终是我国统一的文字，是我国人民最为通用的信息记载工具。汉字为保留和促进中华民族文化作出了巨大的贡献。目前全世界使用汉字的人越来越多，地区越来越广，使用人数已占总世界人口数的30%左右。同时，汉语是人类最精美、最简洁、最丰富的语言之一。在联合国各种语种的文件中，汉语文件最简短，表达含义最准确。因而，放弃使用汉字，对中华民族而言，是不可想象的。

古老的汉字能置身于信息社会不仅仅是依靠人为的加工，主要是汉字本身的素质和结构的直接关系。汉字结构是以笔划为基础的字元有规律的组合，很类似拼音字的字母，字母易于转换成信息，同样字元也易于转换成信息。如果人们认清了汉字结构的本质规则，简易识别汉字固有的信息，就不难进行汉字信息的处理。历史已经证明了这一点。

虽然汉字构字结构复杂，但从当前汉字信息处理的成果来看，与拼音文字相比，汉字也具有以下一些优点：

1. **码短**

汉字输入码平均不大于四个字母，在国标常用字中，平均码长为2.5个汉字，而拼音文字输入码则较长，英文平均字长为6.7个字母。输入码短，打字输入速度就快。

2. **文字简炼**

中文的造句比西文精炼紧凑，一页英文稿件，翻译成中文往往只有半页纸，因而，汉字数据的存储比西文节省空间。

3. **高频字集中**

当今汉字的发展趋势是常用字不断减少，多方统计表明，现在被人们经常使用的汉字基本集中于几千个汉字这样一个相对较小的集合内，而词汇和专业术语却大大丰富，单个汉字的作用加强了。汉语表达的宏观和微观能力更强了。

汉字信息处理属于中文信息处理的范畴。严格地说，中文信息不仅包括汉字信息，也包括中华民族其它各个少数民族和语言和文字信息。从意义和信息处理功能上分析，汉字信息处理包括对汉字本身的处理。

第二节　汉字信息处理研究的基础和内容

从应用角度来看，汉字信息处理研究的主要内容包括汉字输入方法、汉字编码、汉字输出技术三个大的部分。而汉字信息处理系统中这些具体方法与技术的实现要依赖对汉字自身规则及属性研究的成果，这些研究包括汉字字符集研究、汉字属性的研究和汉语词组及方法结构研究等等几方面的内容。这些研究的目的是对汉字信息的规范，并为汉字信息处理研究的核心内容——计算机汉字信息处理提供一个统一明确的基础。

一、汉字属性

汉字属性是指汉字本身所具有的一些基本特性，如字形、字量、字体等特征信息。由于汉字信息处理是一项以计算机

技术为核心的综合性的技术，为了合理地制定一些计算机处理汉字的规则，必须先研究有关的汉字的一些基本特性，即汉字属性。汉字属性基本上包括以下几个方面：

1. 汉字字量

汉字是表意文字，或称象形文字，它的每个字有其特有的形状和构造，这是不同于各种拼音文字的一大特色。在使用中，所有汉字字量的多少是一个重要问题。我国汉字字量多达五六万，而一个汉字处理系统需要收容的汉字数量及所要收录的具体汉字，即汉字集问题，要根据实际使用来确定。

2. 字形分解

汉字字形是汉字属性中的一个重要项目。汉字分解后，其基本组成部分有部首、字首、字根、笔画、位点。可以把位点看作是组成字的最小单位。分解字是为了找出汉字的结构规律，以便为汉字信息处理技术在字形存储方面提供依据。例如在建立汉字库，特别是在建立用字根合成的汉字库的过程中，字形分解显得格外重要。通过对字形分析研究，可以选取最少量的字根，合理地组成所需的汉字，从而可以改善经济性，提高效能。此外，在基于字形特征的汉字编码方法中，为了得到高性能的编码方案，更需注重字形分解的研究工作。

3. 汉字字体

在用于印刷排版的汉字处理系统中，对字体的种类要求较高。一般而言，汉字字体至少可以分为宋体、仿宋体、黑体和楷体四种，而每一种又有方体、长体、扁体的区别。

4. 使用频度

对于不同的汉字，其使用频度的差别是很大的。同一汉字在不同的专业领域使用时，其频度也有差异。因此，对于不同的专业的字频，要分别进行统计。一种综合频度是在若干有代表性的专业领域中统计出各自所用汉字的频度后，力求取其平均值提到的。根据汉字的使用频度，可以把汉字分为常用字、次常用字、稀用字和罕用字等几个等级。在建立不同种类的汉字处理系统中，必须根据使用频度来选用字库中所收容的汉字。

5. 汉字发音

每个汉字有其标准的发音。目前国内以所推广的拉丁化的汉语拼音作为汉字的发音属性。汉字发音特性的研究，对于在计算机中按音序检索汉字，或以发音特性作为汉字编码的研究工作是很重要的。

6. 汉字索引

在汉字信息处理系统中，可以从不同的角度检索汉字，如以笔画、偏旁或部首来检索，也可以用汉字发音的音序来检索，或以其它途径检索。其目的都是要以简捷的规则准确地查得某个汉字的或它的标准编码。

7. 汉字排序

与西文相比，汉字的排序是一个较复杂的问题。汉字可以用笔画的多少排序，也可以用汉字的拼音排序，或者以汉字的综合使用频度排序。无论用哪一种方法作出的汉字排序表，记忆和掌握起来都较困难。目前国内研制的大多数汉字

信息处理系统，都是采用国标码的汉字序数作为内部码的。

二、汉字词组及文法结构

除了汉字属性外，为了更有效地研究汉字信息，需要对组成的字或字组（称为词）进行研究。所谓词是指经常使用并有特定含义的单个汉字或多个汉字的组合。词的属性包括词的种类、组词字数、词的使用频率、词的含义、排序特性等。在汉字信息处理技术中，对词的研究是很重要的。在汉字输入方案中，对于使用频率特别高的词，可以用软件方法设定，用一个键位代表一个词，也可以根据需要改变某个键位所代表的词。

三、汉字信息处理要解决的问题

由于汉字的特点是字量大，字形复杂，因此，要建立一个汉字系统，就需要解决汉字的输入、存储、处理和输出等问题，这在实现上要比实现西文信息处理系统更困难。具体来说，包括以下几方面的内容。

1. 汉字输入技术

汉字输入技术包括编码输入（键盘输入）、手写汉字及印刷汉字的识别输入（光电字符扫描输入）和语音识别输入等项内容。其中编码输入是汉字输入技术的重点，而其它输入技术则是汉字输入的难点所在。

2. 汉字存储编码

汉字信息处理的每一个环节都离不开编码，汉字存储编