

21

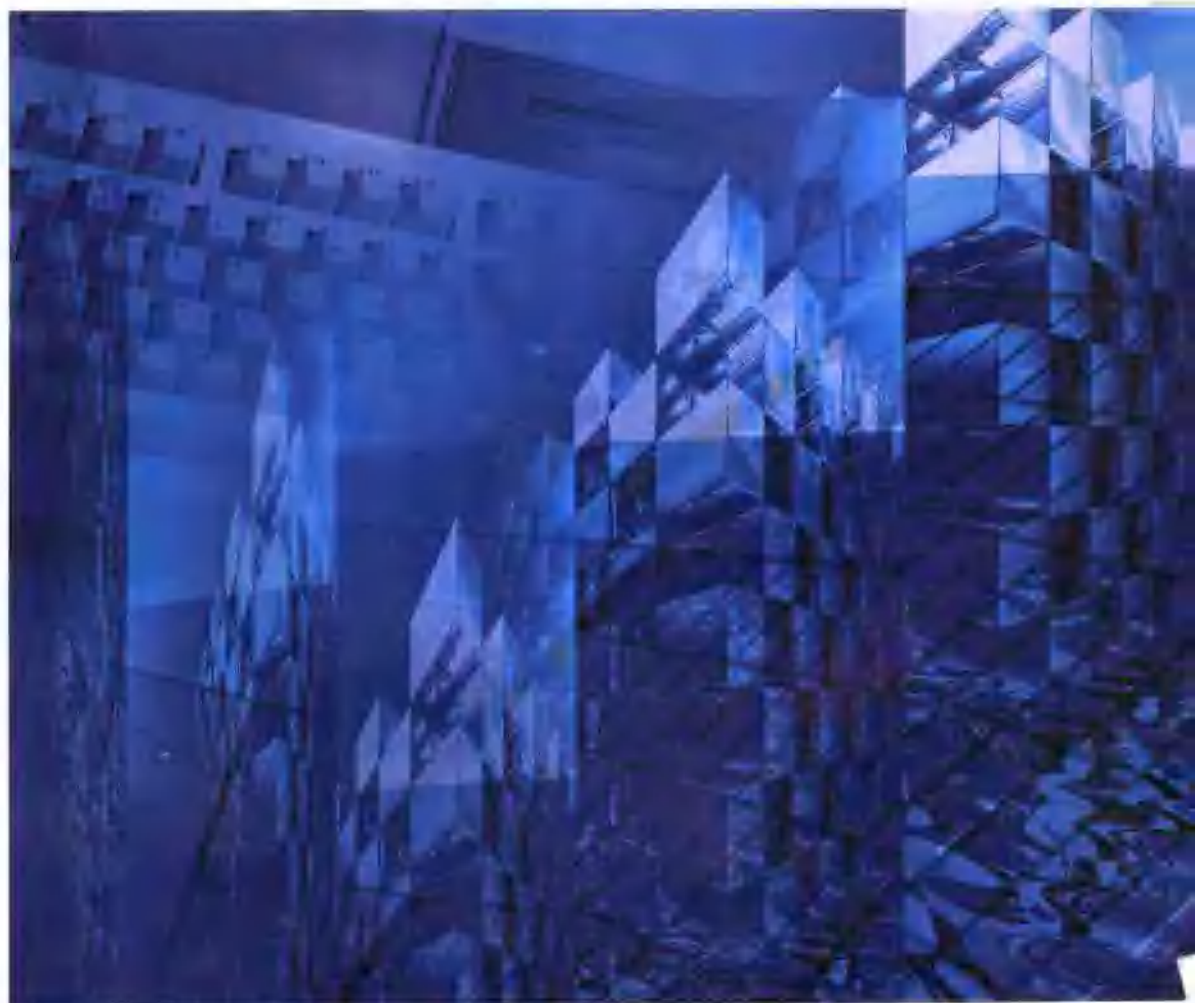
21世纪高等院校十一五规划教材

内蒙古自治区计算机教材编委会 组编

主编 嘎日迪 斯日古楞

蒙古文信息处理技术 及自然语言理解

内蒙古大学出版社



蒙古文信息处理技术 及自然语言理解

内蒙古自治区计算机教材编委会 组编

主 编 嘎日迪 斯日古楞
副主编 赵小兵 阿拉坦巴根那
编 著 阿拉坦巴根那 嘎日迪 林民
斯日古楞 雪梅 赵小兵

内蒙古大学出版社

图书在版编目(CIP)数据

蒙古文信息处理技术及自然语言理解/嘎日迪,斯日古楞主编. —呼和浩特:内蒙古大学出版社,2006.7

ISBN 7-81074-978-1

I. 蒙... II. ①嘎... ②斯... III. ①蒙古语(中国少数民族语言) - 文字处理系统 ②蒙古语(中国少数民族语言) - 自然语言理解 IV. ①TP391.1 ②TP18

中国版本图书馆 CIP 数据核字(2006)第 082200 号

蒙古文信息处理技术及自然语言理解

主 编 嘎日迪 斯日古楞

副主编 赵小兵 阿拉坦巴根那

编 著 阿拉坦巴根那 嘎日迪 林民 斯日古楞 雪梅 赵小兵

内蒙古大学出版社出版发行

内蒙古军区印刷厂印刷

开本:787×1092/16 印张:9.75 字数:182 千

2006 年 7 月第 1 版 2006 年 7 月第 1 次印刷

ISBN 7-81074-978-1/TP·38

定价:12.50 元

内蒙古自治区计算机教材编委会

主 任 李东升 梁希侠(常务)

副主任 满 达 叶新铭 包 那 裴喜春 杨国林

委 员 丁彦武 王润文 乌格德 玉 柱 叶新铭

包 那 李东升 李东魁 李燕华 李玉霞

刘东升 刘利民 刘 实 寿永熙 杨国林

杨建省 赵俊岚 吴 敏 郑惠生 郝长胜

哈斯额尔敦德尼 胡文江 高光来 徐宝清

唐建平 梁希侠 斯日古楞 靳 荣

满 达 裴喜春 嘎日迪 薛河儒

序

内蒙古自治区的高等教育事业起步于20世纪50年代初。经过近50年的发展,我区的高等教育无论从规模上,还是质量上都取得了长足的发展。特别是近些年来,全区高等院校的招生数量成倍增长,部分院校的合并使得一些高校的办学规模迅速壮大,形成了几所万人大学。与此同时,各高校对各自的专业及课程设置都做了较大的调整,以适应当今日益发展变化的高等教育事业。面向21世纪,在科学技术日新月异,社会对人才的知识结构、层次要求越来越高的新形势下,我们的高等教育的教学水平,特别是教材建设都应有一个更新更高的要求。

回顾50年来的发展,虽然我区高等教育的教学科研水平有了较大的提高,但与之相应的教材建设的现状还不尽如人意,绝大多数主干课程的教材还沿用一些传统教材,有些甚至是20世纪七八十年代的版本。有些院校的教材选用则有一定的随机性,在几种版本的教材之中换来换去。其间,虽然部分院校也组织力量编写了一些基础课及专业课教材,但大都是各成体系,缺乏院校间的协作与交流,形不成规模,质量亦无法保证,常常滞后于学科的发展与课程的变化。这都与我区高等教育的发展极不协调。诚然,区外部分地区高校的教学科研水平比我区要高,一些教材的质量好,我们可以直接利用,但这并不能成为我们不搞教材建设的理由。好的教材还需要相应的教育资源条件与之相对应才能取得良好的教学效果,从而达到促进教学质量提高之目的。应当承认,由于经济发展的相对落后,我区高校所招学生的基础和学校的教学条件比起全国重点名牌大学相对要差一些。因而,我们高校的教材也应从实际出发,结合自己学校和学生的特点,逐步探索、建立一套适合自治区教育资源条件的教材体系,促进自治区高校教学科研水平的提高,多出人才,出好人才。

值得欣喜的是,随着自治区教育科学水平的提高,我区高校教育领域的一些有识之士逐渐认识到,面向21世纪,来来高校之间的竞争就是学校的产品——学生质量的竞争。要想培养出高水平、高素质的学生,使我区的高校在这种竞争中立于不败之地,除各高校应努力提高自身的教学组织管理水平、提高教师的素质外,还应积极主动地加强与区内外高校的协作、交流,取长补短,走联合发展的道路,使我区高等教育的整体水平能够在较短的时间内得到提高。为此,在有利于规范高校教材体系,促进高校教育质量的提高,加强各高校教学科研人员之间的协作与交流的原则下,由自治区教育厅牵头,内蒙古大学出版社组办、资助,联合全区高等院校的有关专家、学者共同组建成立一些相关专业的教材编委会,以求编写适合我区高等教育特点的教材,逐步建立、完善自治区高等教育的教学、教材体系,并开展一些与教学相关的科研工作。我们希望,通过教材编委会这种工作模式,建设一批高质量的教材,带出一支高水平的师资队伍,培养出大批高素质的人才。

我坚信,在自治区教育厅的指导下,在编委会各位专家、学者的辛勤工作下,在各院校的相互理解、相互协作、相互支持下,我们一定能够克服发展过程中的困难,逐步推出一批高质量、高水平的教材,为推进内蒙古自治区高等教育事业做出重要的贡献。

2002年3月19日

目 录

第1章 蒙古文信息处理技术概述	(1)
1.1 蒙古文信息处理技术基本概念	(1)
1.1.1 信息	(1)
1.1.2 信息技术	(1)
1.1.3 信息技术特征	(2)
1.1.4 信息技术发展概况	(3)
1.1.5 信息社会	(3)
1.1.6 信息与数据	(3)
1.1.7 数据处理	(3)
1.1.8 数据编码	(4)
1.1.9 信息处理	(4)
1.1.10 蒙古文信息处理技术	(5)
1.1.11 蒙古文信息处理技术所涉及的领域	(5)
1.1.12 蒙古文信息处理过程	(7)
1.2 蒙古文信息处理系统	(7)
1.2.1 蒙古文信息处理系统的特点及中西蒙文兼容技术	(7)
1.2.2 汉字和蒙古文信息处理系统的组成框图	(9)
1.2.3 蒙古文和汉字的输入	(10)
1.2.4 字库	(12)
1.2.5 蒙古文信息的加工处理	(12)
1.2.6 蒙古文的输出	(13)
1.3 蒙古文信息处理技术发展简史	(14)
习题一	(15)
第2章 蒙古文操作系统	(16)

2.1	英蒙兼容蒙古文操作系统	(16)
2.2	蒙古文、汉字、西文操作系统(中分辨率)	(17)
2.2.1	蒙古文的特殊性与系统设计	(18)
2.2.2	回鹘蒙古文的键盘与打印管理模块	(20)
2.2.3	蒙古文、汉字、西文操作系统与汉字、西文、蒙古文操作系统的关系	(25)
2.3	多文种操作系统(高分辨率)	(25)
2.3.1	键盘输入模块	(26)
2.3.2	显示驱动模块	(27)
2.3.3	打印驱动模块	(30)
2.4	蒙古文 Windows 操作系统	(32)
	习题二	(33)
第3章	蒙古文字处理软件和精密照排系统	(34)
3.1	蒙古文华光、方正排版系统简介及基本设备	(34)
3.1.1	蒙古文华光、方正电子出版系统基本配置	(34)
3.1.2	华光、方正操作系统	(36)
3.2	Windows 环境下的精密照排系统和蒙古文字处理软件	(36)
3.2.1	利用字处理软件建立蒙古文文本	(37)
3.2.2	利用字处理软件编辑蒙古文文本	(41)
3.2.3	利用字处理软件编辑多文种文本	(44)
3.2.4	制作蒙古文表格	(46)
3.2.5	蒙古文文本及表格的显示	(51)
3.2.6	蒙古文文本的打印	(53)
3.2.7	蒙古文文本的查找与替换	(53)
3.3	WindowsNT/2000/XP 环境下的蒙古文字处理	(54)
3.3.1	编辑蒙古文文本	(55)
3.3.2	编辑蒙古文表格	(57)
3.3.3	蒙古文、汉字和西文的混合使用	(58)
3.3.4	编辑演示文稿	(59)

习题三	(60)
第4章 蒙古文信息处理现状与展望	(61)
4.1 引言	(61)
4.2 我国蒙古文语料库研究现状	(61)
4.2.1 蒙古语复合词的确定和自动切分知识软件系统	(63)
4.2.2 蒙古文同形多义词知识软件系统	(64)
4.2.3 蒙古文附加成分知识软件系统	(64)
4.2.4 蒙古文动词处理知识软件系统	(64)
4.2.5 校对知识软件系统	(64)
4.2.6 特殊处理知识软件系统	(65)
4.2.7 蒙古文词类标注规则系统	(65)
4.2.8 蒙古文数据库系统	(65)
4.3 我国蒙古文机器翻译技术研究现状	(68)
4.3.1 研究方法	(68)
4.3.2 电子词典介绍	(70)
4.3.3 EMMT 系统规则库	(70)
4.4 我国蒙古文识别技术研究现状	(74)
4.5 我国蒙古文辅助教学软件研究现状	(76)
4.5.1 “蒙汉英多媒体电化教学软件系统”功能简介	(76)
4.5.2 “小学数学蒙古文多媒体教学系统”功能简介	(78)
4.6 结束语	(84)
习题四	(85)
第5章 蒙古文 unicode IME 输入法介绍	(86)
5.1 蒙古文键盘表	(88)
5.2 与蒙古文国际标准编码名义字符及控制字符相对应键盘布局表	(89)
5.3 自由变体选择符	(90)
5.4 蒙古文元音间隔符	(94)

5.5 窄宽度无间断空格·····	(95)
5.6 零宽连接符·····	(96)
5.7 零宽禁连接符·····	(96)
5.8 单个字符的录入方式·····	(96)
5.9 复合元音的录入方式·····	(101)
习题五·····	(102)
第6章 自然语言处理·····	(103)
6.1 自然语言处理与蒙古文信息处理·····	(103)
6.2 自然语言处理的基本问题·····	(105)
6.3 自然语言处理新趋势·····	(111)
6.4 自然语言处理的其它几个难点·····	(112)
习题六·····	(113)
第7章 自动分词·····	(114)
7.1 自动分词概述·····	(114)
7.1.1 分词规则·····	(114)
7.1.2 自动分词原则·····	(115)
7.2 词典系统·····	(116)
7.2.1 词典体系简介·····	(116)
7.2.2 分析词典·····	(117)
7.2.3 概念词典·····	(119)
7.2.4 动态词典·····	(120)
7.3 机械分词方法·····	(121)
7.3.1 机械分词方法简介·····	(121)
7.3.2 机械分词方法的局限性·····	(122)
7.4 歧义字段处理·····	(124)
7.4.1 歧义处理知识·····	(124)
7.4.2 一体化分词·····	(125)

7.4.3 分词规则.....	(126)
7.5 未登录词的处理	(127)
7.5.1 未登录词识别方法综述.....	(127)
7.5.2 中文姓名的自动识别.....	(128)
习题七.....	(130)
第8章 自动标注	(132)
8.1 词性标注与概念标注	(132)
8.1.1 词性标注与概念标注.....	(132)
8.1.2 歧义消除.....	(133)
8.2 隐马尔可夫模型.....	(135)
8.3 标注算法的一个具体例子.....	(136)
8.3.1 词类分类标准.....	(136)
8.3.2 模型的选择.....	(137)
8.3.3 词典管理.....	(139)
8.3.4 面向文本解释的标注.....	(140)
习题八.....	(140)
第9章 机器翻译	(141)
9.1 概述	(141)
9.2 汉语到蒙古语机器翻译系统.....	(142)
9.3 蒙古语到汉语机器翻译系统.....	(143)
习题九.....	(144)

第1章 蒙古文信息处理技术概述

1.1 蒙古文信息处理技术基本概念

1.1.1 信息

从广义上讲,信息(information)是人类一切生存活动和自然存在所传达的信号和消息,是人类社会所创造的全部知识的总和。回顾人类发展史可以发现,人类在进入文明社会以后之所以比以往发展得更快,就在于人类更有效地继承前人的经验和知识,而经验和知识就是信息。人们常说“一代比一代强”的道理就在这里。信息是人类的一种宝贵资源。

在现代社会,信息占据着十分重要的地位。人们无论何时何地都在与信息打交道,阅读的报纸是信息,看到的广告是信息,电视传播的是信息,上网浏览的是信息等等。当代人简直生活在信息的“海洋”中,每时每刻都在自觉或不自觉地接收或传递着各式各样的信息。但不是所有的信息都有用,人们需要从各式各样的、不断出现的、无序的信息中收集和整理出有用的信息,并加以管理、传播和利用。计算机能够帮助人类更好地管理信息,存储信息,处理信息,传播信息和利用信息。

1.1.2 信息技术

信息技术是人类继狩猎技术、农业技术和工业技术之后的又一新技术。

狩猎技术的核心是石器和语言(从被动到主动)。

农业技术的核心是以锄为代表的农业工具和文字。文字的产生,有助于人类智慧的记忆、保存和交流,冲破了时间和空间的限制。

工业技术的核心是以蒸汽机(英国发明家 James Watt, 1736-1819)为代表的动力机械。人类开始以机器生产代替手工劳动,利用蒸汽机人类第一次实现了从热能到机械能的转换,从而蒸汽机也成了人类征服和改造自然的强大的物质力量之一。产业革命的实质就是能源的利用。

信息技术是指对信息的获取、传递、存储和处理的相关技术。其核心为计算机技术和通信技术。它极大地扩展了人类了解自然和征服自然的能力。社会的进步将不断产生信息,同时也需要先进的技术来支撑信息的获取、处理和发布。信息技术从生产力变革和智力开发两个方面推动着社会文明的进步,它对人类社会的冲击比以往的社会技术更为猛烈,发展的速度更快,影响更深远。

1.1.3 信息技术特征

(1) 高智商性

信息技术是高科技,以大量的知识为依托,是一种知识密集型和智力密集型的高科技。信息产品是信息技术的物化和体现。对信息产品的研究开发,需要集中大批尖端科技人员,形成高素质的人才群体,其结果将有力推动信息技术的进步,使新信息技术、新信息产品不断出现,且出现的周期越来越短。

(2) 高速更新性

信息技术的快速、高水平发展使产品更新换代的周期大大缩短。如,电子管计算机 12 年之后才被晶体管计算机所取代,而从晶体管计算机到集成电路计算机只用了 6 年时间,从集成电路计算机到大规模集成电路、超大规模集成电路计算机的出现只用了 4 年时间。再如,DOS 系统统治几十年,WINDOWS95 刚取代之,两三年之后又被 WINDOWS98 所取代。可见,信息技术带给的产品更新换代周期的变化是十分明显的。

(3) 高投入性

集计算机技术、通信技术、微电子技术于一体的信息技术,在其研究开发的任一环节上都需要巨大的资金投入。一般涉及到三方面的费用:一是精密仪器的配置;二是尖端材料的消耗;三是复杂的开发活动。

(4) 高风险性

在传统技术领域,高投入期望高回报,例如,资金密集、劳动密集的粗放型产业更为如此,而信息技术高投入的回报率风险性很大。可从以下三个方面来看:一是研究本身具有很多的不确定性;二是从设计开发到形成产品成功率很低;三是市场环境瞬息万变,波动很大。

(5) 高增值性

信息技术的高风险性与预期的高额利润同时存在。信息技术的高额利润包括两个方面:一是信息技术的高价值;二是信息技术的高增值性。

(6) 高竞争性

拥有信息技术的程度反映一个企业的经营水平,反映一个国家的综合国力。所以,信息技术是政府决策者、企业经营者关注的焦点之一。

(7) 高渗透性

半个世纪以来,信息技术的发展与应用,使得社会从以物质生产占主导地位的形式逐步转变为以信息技术占主导地位的形式,新的信息技术在不断地向物质生产领域渗透。

1.1.4 信息技术发展概况

自从人类诞生到现在,在信息技术方面已经历了五次革命。

第一次信息技术革命是语言的使用。语言的产生是人类历史上最伟大的信息技术革命,其意义不亚于人类取火。

第二次信息技术革命是文字的创造。由于人脑容易遗忘事情,一旦遗忘,信息就无法取出。为了长期存储信息,人们开始创造出一些符号来代表语言。经过漫长岁月的发展,这些符号便逐渐演变成了文字。

第三次信息技术革命是印刷技术的发明。中国古代四大发明中的造纸术和印刷技术与第三次信息技术革命有密切联系。

第四次信息技术革命是电报、电话、广播、电视的发明和普及应用。它标志着人类电子时代的开始。

第五次信息技术革命开始于是20世纪60年代,其标志是电子计算机的普及与现代通信技术的有机结合。

信息技术的发展趋势可概括为四化,即宽带化、数字化、网络化和全球化。

1.1.5 信息社会

在历史上,人类社会经历了农业社会和工业社会。现在随着信息量的增加和信息技术的发展,信息社会已成为世界各国探讨的主题。信息社会是以信息活动为社会基本活动,以信息经济为主导经济,以信息技术为技术基础,以信息文化改变人类教育、生活和工作方式以及价值观的新兴社会形态。在信息社会里,信息作为继物质和能源之后的第三个资源,在社会发展中起着主导作用。在今天,物质、能源和信息是人类赖以生存的三大要素。

1.1.6 信息与数据

虽然数据与信息有密切联系,但它们是两个截然不同的概念。将它们视为等同是错误的。

数据(data)是用来反映客观世界而记录下来的可以鉴别的符号,其本身无意义。而数据所表示的含义就是信息。如,在候车室屏幕上滚动显示的一串符号“K98 开车时间 10:30”是数据,但其所表示的意义是信息,它告诉人们,“K98 列车 10:30 开”的信息。数据不光包括数字,还包括图形、图像、动画、影像、声音等。

1.1.7 数据处理

数据处理是指将数据转换成信息的过程。

信息=数据+处理

1.1.8 数据编码

数据在计算机内部是以二进制编码表示的。所谓的数据编码就是规定用什么样的二进制编码来表示字母、数字及专门符号。数字通常采用 BCD 码, 英文字符采用国际通用的 ASCII 字符编码, 即美国国家标准信息交换码, 在我国, 汉字采用 GB2312-80 标准和 GB13080-2000 标准所规定的汉字国标码。

表 1.3.2 十进制数与 BCD 码对照表

十进制数	BCD 码	十进制数	BCD 码
0	0000	5	0101
1	0001	6	0110
2	0010	7	0111
3	0011	8	1000
4	0100	9	1001

[例] 请将 $(973)_{10}$ 转换为 BCD 码。

$(973)_{10} = (1001 \ 0111 \ 0011)_{BCD}$

1.1.9 信息处理

信息处理是指获取信息、传递信息、存储信息和加工信息。

在信息处理中, 对文字信息的处理称之为文字信息处理。这里主要介绍文字信息处理中的蒙古文信息处理。

事实上, 计算机系统只能处理数据, 而数据所体现的含义就是信息。因此, 在这里所谓的信息处理就是指数据处理。表示文字信息或符号信息的数码, 称为代码 (code)。例如, 在对西文字母以及符号的处理中, 对应 26 个字母 (分为大写和小写) 和一些常用符号, 按照某种规律和约定, 编一组数码, 称之为字符代码, 如国际通用的 ASCII 码。对文字信息的加工处理就是对代码数据进行加工处理。文字信息处理过程可概括为三个阶段:

(1) 信息的输入。通过输入设备把文字信息转换成代码, 并输入到计算机中。

(2) 信息的加工处理。根据不同的用处, 借助预先设计好的程序对输入的信息进行加工处理, 从而得出结果信息。

(3) 信息的输出。通过输出设备把以数据代码形式表示的结果信息, 复原为文字。

目前, 在科学技术高度发达的西方国家, 成熟的文字信息处理技术非常普遍, 究其原因, 除了这些国家对计算机设备、技术开发得较早, 掌握得较好外, 还有一个重要的原因就是这些国家的文字都是拼音文字, 拼音文字字母数量少, 字形简单, 从而容易进行处理。

而我们所广泛使用的汉字是一种表意文字,具有字数多、字形复杂等特点。这些对汉字输入和汉字字模库的建立等带来了诸多困难。

我国是一个多民族的国家,使用着多种文字。除汉字之外,现使用着的还有蒙古文、藏文、朝鲜文、维吾尔文、哈萨克文、柯尔克孜文、彝文、锡伯文等 38 种文字。加上曾在历史上用过,现已不用了的少数民族文字就有 50 多种文字。设计开发包括少数民族在内的各民族文字信息处理系统才能真正满足国家各方面的需求。

1.1.10 蒙古文信息处理技术

所谓蒙古文信息处理是指对蒙古文代码进行加工处理。如蒙古文的输入、输出、识别和加工处理等。蒙古文字符处理是指用计算机对蒙古文字符进行操作和加工,例如,对蒙古文字符进行编码、输入、打印、存储、传输、识别等等。在蒙古文信息处理中,蒙古文字符处理是基础,是最基本的处理层次,在此基础上进行的蒙古文信息处理更为重要。它不是单个蒙古文字符或词语,而是用蒙古文文字及其集合来表征自然存在和社会活动中产生的信息,并利用计算机对这些信息进行分类、编目、存储、检索、分析、综合决策、打印输出、辅助设计以及提供咨询服务等等。目前,在我国,计算机主要用于事务处理、信息管理、办公自动化、情报检索、轻印刷、激光照排、科学研究和工程设计等领域。而蒙古文信息处理是这一系列应用的基础。若不实现蒙古文信息的计算机处理,我国的蒙古民族的计算机应用就难以推广。目前不少发达国家已开始进入现代化社会,社会现代化是以社会的信息化为标志的,没有信息化,就没有现代化。因此,蒙古文信息处理技术的研究和开发对我国蒙古民族尽快实现社会现代化有着重要的意义。

1.1.11 蒙古文信息处理技术所涉及的领域

蒙古文信息处理技术是一门跨学科、多学科交叉的边缘学科,涉及语言学(包括词汇学、词义学、语用学、语法学、识别与理解、方言学、风格学、翻译学等等)、心理学、声学、计算机科学、数学、电子学、逻辑学和人工智能等学科。而多学科交叉研究又产生了计算机语言和语言工程学等新的边缘学科。蒙古文信息处理技术是一门高技术学科,它建立于计算机、大规模集成电路设备的基础之上。蒙古文是蒙古民族的通用语言文字,蒙古民族曾在人类文明历史长河中留下过许多光辉灿烂的篇章。由此可见,蒙古文信息处理技术研究的广泛性远远超过其它学科。

蒙古文信息处理技术研究是在多层面上开展的,其基础研究、高层次研究、应用研究和产品开发彼此密切配合,相互促进,使蒙古文信息处理技术不断得到发展。蒙古文信息处理技术主要涉及以下研究领域。

1. 基础理论研究

(1) 蒙古文属性研究。它包括蒙古语的使用频度、蒙古文读写规则、蒙古文排序和蒙古文索引等内容。

(2) 蒙古文信息处理交换用国家标准。它包括信息处理交换用蒙古文字符集国家标准;信息处理交换用蒙古文字符集键盘字母区布局;信息处理交换用蒙古文点阵字模集等内容。

(3) 词与词组的研究。它包括词的种类、含义、数量、同形词、多义词、动词、复合词、附加成分等内容。

(4) 蒙古文信息的输入理论研究。

(5) 蒙古文字形识别研究和蒙古文语音识别研究。

(6) 蒙古语自然语言理解。它包括用计算机对蒙古语的语法和语义进行分析、蒙古语的人机接口和自动作文摘、自动编译文稿等内容。

(7) 计算语言学和语言工程学研究。

(8) 蒙古语与其它自然语言间的自动翻译研究。

(9) 蒙古语族语言间的比较研究。

2. 蒙古文信息处理的基本环境及其实现

(1) 蒙古文、汉字、西文兼容技术。

(2) 蒙古文信息的输入、输出技术与设备。

(3) 蒙古文信息的存储。

(4) 蒙古文信息的机内处理。

(5) 蒙古文、中文、西文兼容系统软件。

(6) 蒙古文信息处理系统。

(7) 蒙古文信息处理开发平台。

(8) 蒙古文的本地通信和远程通信技术。

3. 蒙古文信息处理技术的应用开发

(1) 电子印刷排版系统的研制与开发。

(2) 情报检索和档案管理系统的研制与开发。

(3) 辅助教学、辅助设计系统的研制和开发。

(4) 蒙古文办公室自动化系统研制与开发。

(5) 蒙古文专家系统研制与开发。

(6) 翻译系统研究。

(7) 蒙古文语音识别系统和字形识别系统等模式识别技术的研究。

(8) 蒙古文信息通讯系统研制与开发。

(9) 蒙古文订票系统、公用咨询服务系统、电话号码查询系统和广告宣传系统等应用项目研制与开发。

1.1.12 蒙古文信息处理过程

蒙古文信息处理过程如图 1-1 所示, 由三大部分组成:

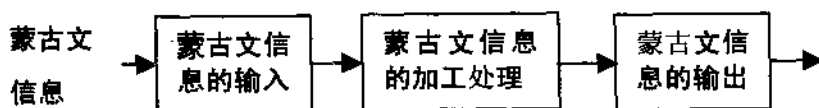


图 1-1 蒙古文信息处理流程图

1. 蒙古文信息输入

通常借助于键盘、模式识别设备、光笔输入设备等把蒙古文字符或语音所代表的蒙古文信息转换成计算机内所使用的代码(称为内部码或内码), 输入到计算机内进行加工处理。

2. 蒙古文信息加工处理

对蒙古文信息加工是借助计算机执行相关程序来完成的。蒙古文信息处理包括多种不同的处理要求, 如对文稿中蒙古文字符的增、删、改等操作; 对若干字符或整段的增、删、改等操作; 对字符串的处理, 如分类、合并、比较、排序、检索以及对齐等均属于蒙古文字符的加工处理, 是基础性工作。更高层面的蒙古文信息加工处理包括对文稿中句子进行语法和语义分析、句子和全文的理解、自动作文摘、机器翻译等。

3. 蒙古文信息的输出

蒙古文信息的输出形式有字符输出和语音输出两大类。系统将经计算机处理的结果信息以蒙古文字形或语音形式进行输出。当蒙古文信息处理系统进入公用网进行蒙古文信息交换时, 输出的蒙古文信息一般是用某种蒙古文代码来表示的。

1.2 蒙古文信息处理系统

1.2.1 蒙古文信息处理系统的特点及中西蒙文兼容技术

目前的计算机系统除了进行计算外还能够处理文字信息。但是, 计算机首先是由美国人设计开发的, 当初只考虑了对英文的处理, 没有考虑对汉字、蒙古文等其它文字的处理。从信息处理角度来看, 汉字、蒙古文等文字数字化后与西文(英文)字符没有本质区别。原则上, 现有的西文信息处理系统可以用来处理汉字、蒙古文信息。但事实上远非如此。