·藏文信息处理技术·

བོད་ཡིག་ཆ་འཕྲིན་སྒྲིག་གཅོད་ཀྱི་གཞུང་ལུགས་དང་བཀོལ་སྤྱོད།

藏文信息处理的原理与应用

高定国　珠　杰◎编著

西南交通大學出版社
Http://press.swjtu.edu.cn

西藏大学博士学位授权立项建设项目资助

 བོད་ཡིག་ཆ་འཛིན་སྒྲིག་གཅོད་ཀྱི་གཞུང་ལུགས་དང་བཀོལ་སྤྱོད།

藏文信息处理的原理与应用

高定国　珠 杰　编著

西南交通大学出版社
·成　都·

内容简介

本书共10章，1～6章主要介绍藏文信息处理的概念、藏文字符的编码方式及目前所用的几种藏文字符编码，支持藏文处理的Windows、Linux系统以及不同系统下藏文字符的键盘、语音、字符识别输入方式，藏文字形设计技术等藏文信息处理的原理；7～10章以藏文信息处理的原理为基础，介绍了藏文信息检索、藏文信息抽取、藏文文本分类和机器翻译等藏文信息处理的应用。

本书可以作为高等院校藏文信息技术、计算机科学与技术、电子信息技术等相关专业的高年级本科生或研究生的教材或参考书，也可以作为从事藏文信息处理、藏语计算语言学、数据挖掘和人工智能研究的相关人员的参考书。

图书在版编目（CIP）数据

藏文信息处理的原理与应用 / 高定国，珠杰编著.
—成都：西南交通大学出版社，2013.6
（藏文信息处理技术）
ISBN 978-7-5643-2207-6

Ⅰ. ①藏… Ⅱ. ①高… ②珠… Ⅲ. ①藏语－文字处理系统 Ⅳ. ①TP391.1

中国版本图书馆CIP数据核字（2013）第035108号

藏文信息处理技术

藏文信息处理的原理与应用

Zangwen Xinxi Chuli de Yuanli yu Yingyong

高定国　珠　杰　编著

*

责任编辑　李芳芳　张　波
封面设计　墨创文化
西南交通大学出版社出版发行
四川省成都市金牛区交大路146号　邮政编码：610031　发行部电话：028-87600564
http: //press.swjtu.edu.cn
成都蓉军广告印务有限责任公司印刷

*

成品尺寸：210 mm×285 mm　印张：16.5
字数：411千字
2013年6月第1版　2013年6月第1次印刷
ISBN 978-7-5643-2207-6
定价：36.00元

前　言

从20世纪80年代起，北京、上海、西藏、甘肃、青海等地的一些院校及科研机构纷纷开始了藏文信息处理的研究，研制开发了许多藏文信息处理系统，推动了藏文信息处理技术的发展。藏文信息处理技术得到了党和国家领导人的高度重视，取得了较好的成绩。西藏大学一直从事藏文信息处理技术的研究与教学工作，这次以西藏大学藏语计算语言学博士点建设为契机，把十几年来在藏文信息处理技术方面的研究文档和讲义、教案整理成册，形成了本书。

本书共10章，1～6章由高定国编著，7～10章由珠杰编著。本书是在西藏大学博士点授权学科“藏语计算语言学”立项建设项目、藏文信息技术教育部创新团队[藏文信息技术“长江学者和创新团队发展计划”创新团队（IRT0975）]、“计算机及藏文信息技术”国家级教学团队、国家自然科学基金项目“基于虚词的藏语基本句型的形式化研究”（61063015）和“藏文Web信息的社会网络动态演化机理研究”（61262058）等项目的资助下所完成的成果之一。西南交通大学信息学院的李天瑞教授审读了全书，并提出了很好的修改意见，本书也得到西藏大学藏文信息技术研究中心、工学院领导、同仁的帮助，编写过程中借鉴了很多同行的研究成果，在此一并表示感谢！

“藏文信息处理技术”已列入国家“新闻出版改革发展项目库”。

由于编著人员水平有限，加之时间仓促、可参考资料少，书中难免存在不妥之处，恳请广大读者批评指正。

高定国

壬辰年三九寒冬在西南交通大学镜湖宾馆

2013年1月

目　录

第 1 章

概　论

藏文信息处理技术作为计算机技术与藏语言文字相结合的一门交叉学科，随着信息处理技术的发展，近年来得到了快速的发展。语言文字信息处理作为计算机应用技术的一个重要分支，伴随着计算语言学、心理学、数学以及计算机科学的发展，已经成为新世纪信息技术中的一个重要研究领域。无论是藏文信息处理还是语言文字信息处理，都离不开“信息”这个概念。

1.1　信　息

“信息”一词在英文、法文、德文、西班牙文中均是“information”，日文中为“情报”，我国台湾称之为“资讯”，我国古代用的是“消息”。作为科学术语最早出现在哈特莱（R. V. Hartley）于 1928 年撰写的《信息传输》一文中。20 世纪 40 年代，信息论的创始人香农（C. E. Shannon）给出了信息的明确定义。此后许多研究者从各自的研究领域出发，给出了不同的定义。具有代表意义的表述如下：信息论的创始人香农认为“信息是用来消除不确定性的东西”。这一定义被人们看作是经典性定义并加以引用。控制论创始人维纳（Wiener）认为“信息是人们在适应外部世界，并使这种适应反作用于外部世界的过程中，同外部世界进行互相交换的内容和名称”。它也被作为经典性定义加以引用。经济管理学家认为“信息是提供决策的有效数据”。物理学家提出了“信息熵”的概念，用信息熵描述系统与环境交流信息的程度。电子学家、计算机科学家认为“信息是电子线路中传输的信号”。我国著名的信息学专家钟义信教授认为“信息是事物存在方式或运动状态，以这种方式或状态直接或间接的表述”。美国信息管理专家霍顿（F. W. Horton）给信息下的定义是：“信息是为了满足用户决策的需要而经过加工处理的数据。”简单地说，信息是经过加工的数据，或者说，信息是数据处理的结果。

从哲学的角度说，信息是事物运动的存在或表达形式，是一切物质的普遍属性，实际上包括了一切物质运动的表征。传播学研究的信息是在一种情况下能够减少或消除不确定性的任何事物，它是人的精神创造物。

（1）本体论层次的信息。在最一般的意义上，亦即没有任何约束条件，我们可以将信息定义为事物存在的方式和运动状态的表现形式。这里的“事物”泛指存在于人类社会、思维活动和自然界中一切可能的对象。“存在方式”指事物的内部结构和外部联系。“运动状态”则是指事物在时间和空间上变化，信息的载体所展示的特征、态势和规律。

（2）认识论层次的信息。主体所感知或表述的事物存在的方式和运动状态。主体所感知的是外部世界向主体输入的信息，主体所表述的则是主体向外部世界输出的信息。

在本体论层次上，信息的存在不以主体的存在为前提，即使根本不存在主体，信息也仍然存在。在认识论层次上则不同，没有主体，就不能认识信息，也就没有认识论层次上的信息。

根据近年来人们对信息的研究，信息的概念可以概括为：

信息是对客观世界中各种事物的运动状态和变化的反映，是客观事物之间相互联系和相互作用的表征，表现的是客观事物运动状态和变化的实质内容。[①]

根据全国科学技术名词审定委员会审定，信息（Information）定义为“以适合于通信、存储或处理的形式来表示的知识或消息”。

信息具有以下性质：① 普遍性；② 依附性；③ 有序性；④ 相对性；⑤ 可度量性；⑥ 可扩充性；⑦ 可存储、传输与携带性；⑧ 可压缩性；⑨ 可替代性；⑩ 可扩散性；⑪ 共享性；⑫ 时效性；⑬ 传递性；⑭ 价值相对性；⑮ 真伪性；⑯ 可处理性；⑰ 客观性；⑱ 不完全性；⑲ 可加工性。

在信息论中，信息从不同的角度有不同的分类[②]：

（1）按性质，信息可分为语法信息、语义信息和语用信息。

（2）按地位，信息可分为客观信息和主观信息。

（3）按作用，信息可分为有用信息、无用信息和干扰信息。

（4）按应用部门，信息可分为工业信息、农业信息、军事信息、政治信息、科技信息、文化信息、经济信息、市场信息和管理信息。

（5）按携带信息的信号性质，信息还可以分为连续信息、离散信息和半连续信息。

（6）按事物的运动方式，信息可分为概率信息、偶发信息、确定信息和模糊信息。

（7）按内容，信息可分为消息、资料和知识。

（8）按空间状态，信息可分为宏观信息、中观信息和微观信息。

（9）按信源类型，信息可分为内源性信息和外源性信息。

（10）按价值，信息可分为有用信息、无害信息和有害信息。

（11）按时间性，信息可分为历史信息、现时信息和预测信息。

（12）按载体，信息可分为文字信息、声像信息和实物信息。

1.2 信息处理

信息处理就是对信息的接收、存储、转化、传送和发布等过程。随着计算机科学的不断发展，计算机已经从初期的以“计算”为主的一种计算工具，发展成为以信息处理为主、集计算和信息处理于一体的工具。

进一步分析计算机信息处理的过程，可以看到，信息的接收包括信息的感知、信息的测量、信息的识别、信息的获取以及信息的输入等；信息的存储就是把接收到的信息通过存储

① 李兴国. 信息管理学[M]. 北京：高等教育出版社，2007：4-5.

② 李兴国. 信息管理学[M]. 北京：高等教育出版社，2007：4-5.

设备进行缓冲、保存、备份等的处理；信息的转化就是根据人们的特定需要把信息进行分类、计算、分析、检索、管理和综合等处理；信息的传送就是通过计算机内部的指令或计算机之间构成的网络把信息从一个地方传送到另外一个地方的处理；信息的发布就是把信息通过各种表示形式展示出来。

计算机信息处理的过程实际上与人类信息处理的过程一致。人类的处理也是先通过感觉器官获得信息，然后通过大脑和神经系统对信息进行传递与存储，最后通过言、行或其他形式传出信息。

1.3 中文信息处理

中文信息处理是指用计算机对中文的音、形、义等信息进行处理和加工。中文信息处理是自然语言信息处理的一个分支，是一门与计算机科学、语言学、数学、信息学、声学等多种学科相关联的综合性学科，信息处理技术的应用很广泛。从1980年开始，中文信息处理进入了快速发展阶段，并极大地提高了社会的信息处理效率。狭义上来说，中文信息处理分为汉字信息处理与汉语信息处理两部分，具体内容包括对字、词、句、篇章的输入、存储、传输、输出、识别、转换、压缩、检索、分析、理解和生成等方面的处理技术。从广义上来说，中文信息处理所需要处理的文字，不仅包括简体汉字、繁体汉字，也包括藏文、蒙文、壮文、维吾尔文等大量少数民族的文字，周边国家的片假名、谚文，还包括古汉语文字、西夏文、契丹文等。

中文信息处理的研究范畴很广，主要包括：

（1）基础研究——字频统计、词频统计、自动分词、句法属性研究、编码字符集、通用字样库、字属性字典、语料库等。

（2）输入技术——键盘输入、手写识别输入、语音识别输入等。

（3）输出技术——字模技术（字形库）、激光照排、语音合成等。

（4）存储技术——字库标准等。

（5）转换技术——编码转换等。

（6）信息处理——情报检索、文本校对、机器翻译、自然语言理解、人机界面等。

中文信息处理的相关学科较多，包括语言文字学、计算机科学、模式识别、人工智能、心理学、数学、控制论、神经计算、模型论、信息学、形式化理论、声学等。

1.3.1 汉文信息处理的发展历史

计算机在1946年被发明，当时主要用于数值计算。到1960年，商用计算机开始普及，计算机被用于处理大规模的数据，其一是图书馆的目录整理。在当时，美国国会图书馆及多家大学都有不少来自东亚的藏书，为了有效管理这批藏书，需要有一套有效处理东亚文字的系统。这套系统需要包括两个方面：其一是如何把东亚文字储存在计算机内；其二是如何在计算机内表示出东亚文字。

在过去，每一台计算机都有各自的数据表达方式，计算机之间不能沟通。直到1960年，

美国信息交换标准码（ASCII）的出现，使得计算机之间可以互相沟通。不过，ASCII 并不能有效地处理英文以外的文字。

最早可以处理中文的计算机，可以追溯到 1970 年。在当年举办的日本大阪万国博览会上，IBM 公司公开了部分汉字处理系统的技术资料，到 1971 年正式发表。当时公布的机种包括“IBM 2345 汉字印刷机”、“IBM029 汉字穿孔机”、“IBMSystem/360-System/370 OS/VS”及“DOS/VSE”等。其后，日本本土公司也争相开发，包括富士通的 JEF（Japanese processing Extended Facility）、NEC 的 JIPS（Japanese Information Processing System）及日立的 KEIS（Kanji processing Extended Information System）等。到 1979 年 5 月，NEC 出产了可使用汉字的个人计算机 PC-8000 系列；到 1982 年 10 月，开发出了 16 位的 PC-9801 个人计算机，使得能处理汉字的计算机在日本渐渐普及。

1973 年，新华社派考察团到日本，参观了日本共同社、日立、日本电气、松下及东芝等公司。他们看到共同社采用磁芯技术解决了 2 000 左右汉字和片假名的存储问题，并发现工作人员使用大键盘方式输入稿件。他们回国后，提出了采用计算机技术改善新华社收、发、编、印四方面的方案，并由中国四机部（民用机械、核工业和核武器、航空、电子工业）与北京市科技局邀请富士通等公司到中国进行技术座谈。在 1974 年 8 月，中国开始了“748”工程，其中包括用计算机来处理中文，到 1980 年发布了 GB 2312—80 汉字编码的国家标准，1985 年中国科学院推出了 Unix 中文版。

我国台湾方面，曾经与 IBM 公司合作研发中文计算机，斥资六千万、历时十年，研究计算机中文字的处理方法，结论是计算机不能处理中文。而当时美国的图书馆已经采用了计算机管理图书，涉及一批中文图书编目问题。另外，多家公司也开发了终端机式的中文系统，如王安、中华一号到中华三号、神通等。这些机器大多数采用大键盘的输入方式，有数十至数百个键。到 1976 年，朱邦复发明了一套形意检字法，并在 1978 年改进为仓颉输入法，以英文键盘实现了中文的输入。在 1979 年，朱邦复提出以图形功能及从显示器的英文字符产生器入手，研发中文计算机，并由宏碁公司出产。在原有英文操作系统上外挂中文系统的方法大行其道，多套中文软件相继推出，包括国乔、倚天及仲鼎等。

20 世纪 80 年代，汉字推出了多种输入方案，形成了“万码奔腾”的局面，也解决了计算机处理汉字的问题；再经过几年的发展低潮之后，到 90 年代末，中文信息处理的重点开始转向了语音识别、语音合成和语义处理等新的领域。

1.3.2 汉文信息处理的研究内容

1. 汉字信息处理

这是一项最关键的语言工程，字符如不能进入计算机，图书情报工作自动化、印刷出版现代化、办公室事务自动化都将化为空谈。近几十年来，汉字信息处理研究得到了很大的发展。曾设计了 400 多种汉字编码方案，其中计算机上通过试验或已被采用为输入方式的，也达数十种之多。研制了上百种汉字信息处理系统和设备，这些系统主要采用两种类型的键盘：一种是笔触式大键盘，另一种是小键盘。前一种除整体输入外，一般还有利用部件组合汉字的能力；后一种有的可兼容多种编码方案，有的还带有计算机引导的智能。

汉字信息处理除了在汉字编码方面进行研究外，还制成了若干种汉字输入、输出专用设

备，其中有各种类型的汉字输入键盘、汉字字库、汉字显示终端、汉字图形兼容终端、汉字印字机。1985 年 5 月，国家标准局公布了《信息交换用汉字 15×16 点阵字模集及数据集》和《信息交换用汉字 24×24 点阵字模集及数据集》两项标准，为各种设备的设计和推广提供了有利条件。

为了使各种系统之间的信息交换有共同性，也为了使各种输入、输出设备的设计有统一的根据，1981 年，国家标准局公布了《信息交换用汉字编码字符集基本集》（简称《汉字标准交换码》）。这个标准是根据汉字使用频度制定的，共分两级，一级 3 755 个字，二级 3 008 个字，共 6 763 个字。为了满足少数用字量超过基本集的用户和台湾、香港等地的需要，制定了《信息交换用汉字编码字符集辅助集》，辅助集将根据使用频度高低分作第一辅助集和第二辅助集，各收 8 000 余字。

2. 机器翻译

计算机和语言的最早结合开始于机器翻译。1956 年，机器翻译被列入中国科学工作的发展规划。1957 年，机器翻译研究工作正式开始，首先研究的是俄汉机器翻译，并于 1959 年成功地进行了试验，输出了汉字译文的代码。1958 年底至 1960 年初，又研制了一套英汉机器翻译规则系统。1966 年至 1975 年机器翻译工作处于停顿状态。近年来，先后在计算机上试验了英汉、俄汉、法汉、日汉和汉外（英、法、德、俄、日）机器翻译系统十余个。

3. 中文信息检索

我国计算机信息检索系统的研发始于 20 世纪 70 年代，由中国科技情报研究所和一些其他研究机构、情报机构及高等院校进行了研究，1983 年，交通部科技情报研究所成功研制了微机单机文献检索系统、微机非文献检索系统及缩微文献检索系统，实现了利用微机检索书目、事实和数据。

4. 汉语理解系统

最近几年，随着人工智能的进展，语言研究所、心理研究所、自动化研究所和一些大学开展了汉语理解系统（人机对话）的研究。上机试验结果表明，有的系统已有识别 30 多种句型的能力。

5. 计算机辅助语言教学

华东师范大学现代化教育技术研究所、哈尔滨工业大学、上海交通大学等单位已研制出了多种语言教学软件，推动了该领域的发展。

6. 语音识别和语音合成

语音打字的任务早在 1958 年便已提出。1964 年实现了“元音识别机”，1970 年前后又实现了 10 个汉语数字的识别机。1972 年，声学研究所利用语音图样匹配方法在一定范围内实现了语言的识别。近几年来，汉语语音的识别和合成取得了很好的发展。

中文信息处理还有很多研究领域，随着研究手段的改善和研究工作的深入，还将会开辟更多更新的研究领域。

1.4 藏文信息处理

1.4.1 藏文信息处理的概念

藏文信息处理就是用计算机对藏语的音、形、义等语言文字信息进行加工和操作，包括对字、词、短语、句、篇章的输入、输出、识别、转换、压缩、存储、检索、分析、理解和生成等各方面的处理技术。它是在语言文字学、计算机应用技术、人工智能、认知心理学和数学等相关学科的基础上形成的一门边缘学科。藏语语言学中的词法学、句法学、语义学和语用学给藏文信息处理的各个层面提供了可靠的理论依据，而人工智能的知识工程、机器学习、模式识别和神经计算，数学中的模型理论、形式化理论和数理统计等构成了藏文信息处理的方法基础。简单地说，藏文信息处理就是利用计算理论和计算技术处理藏文信息的一门学科，是计算机科学与藏语言文学的交叉学科。

藏文信息处理可划分为藏文字符信息处理和藏语语言信息处理两个层次。藏文字符信息处理层面包括操作系统以及信息技术编码字符集、办公软件、文字识别技术、输入技术、字形描述与生成、存储、编辑、排版、字频统计和藏字属性库等方面的研究；藏语语言信息处理层面包括机器翻译、自动分词、语音识别、信息检索、信息提取、文本校对、文本生成、文本分类、自动摘要以及藏文文字识别和语音识别处理等的研究。两者之间也有关系，藏语语言信息处理要以藏文字符信息处理的实现为基础，要提高藏文字符信息处理的智能水平，又要借助藏语语言信息处理的成果。

1.4.2 藏文信息处理的主要研究对象

藏文信息处理的研究对象主要有藏文操作系统、信息技术藏文字符编码、藏文键盘输入技术、藏文输出技术、藏文字形识别技术、藏语语音识别技术、文本分类、信息检索等。

1. 藏文操作系统

在计算机操作系统方面，需要解决藏文信息在计算机内部的表示、藏文信息的输入和输出、藏文信息在 Internet 上的传输、网页上藏文的显示以及数据库中藏文的存放等问题。

2. 信息技术藏文字符编码

为了使藏文软件和电子信息实现相互交换与共享，在藏文编码方面，一定要用国际、国家标准，只有建立了最基本的标准，才能统一现有软件和数据，从而规范藏文软件。在应用方面，需要解决包括编码转换、计算机翻译等一系列与语言信息处理相关的问题。

3. 藏文键盘输入技术

藏文信息处理的必要前提是要把藏文信息输入计算机中。根据是否使用键盘，可以把输入方法分为键盘输入法和非键盘输入法。藏字键盘输入法就是通过键盘把藏字输入计算机中的技术。键盘输入法从单字输入、词输入、常用短语输入一直发展到句输入，输入方法在智能化程度上越来越高。非键盘输入法是指通过藏文字符的图形扫描自动识别和藏语语音的自动识别等方法输入藏文字符的技术。

藏文的编码输入就是用西文键盘上的字符、数字、符号等对藏字进行编码。用户通过输入藏文编码，通过计算机内部的藏文键盘输入程序把它转换成机器内部代码，从而达到输入藏文字符的目的。

4. 藏文输出技术

藏文的输出是指把存储在计算机内的藏文字形信息转换成符合显示或打印需要的形式，并送到输出设备输出。因此，藏文字符的输出技术往往与藏文字模（字形）技术紧密结合在一起。藏文信息的存储和输出过程与西文字符有很大的不同。在藏文字库信息的描述方面，有点阵描述法，用点阵描述法构成的藏文字库就称为“点阵字库”（例如，GB/T 16960.1—1997就是典型的24点阵字库）。对字形的描述还可以采用矢量的方法，每个藏文字形用一组矢量进行描述，这种形式构成的藏文字库就称为“矢量字库”。若藏文字符的字形线段采用数学函数（二次函数、*B*-样条、Besier 函数等）描述，则构成的藏文字库就称为曲线字库。有了藏文字库，藏文输出程序就能对藏文机内码对应的字形信息进行处理（包括还原、放大、缩小等），然后再根据具体的输出设备进行输出。

5. 软件藏文本地化技术

软件藏文本地化是把中、西文软件直接改造成藏文软件的一种技术，分为系统层和应用层两个方面。系统层的藏文本地化通常是指西文操作系统和汉字操作系统藏化成藏文操作系统，使得系统能够处理藏文信息。应用层的藏文本地化通常是指使西文和汉文的应用软件经过藏化后能够具备处理藏文的能力。操作系统的藏文本地化又可以分为内核藏文本地化和外挂藏文本地化两种。内核藏化就是直接修改操作系统的底层内核模块，使得操作系统完全支持藏文信息的处理。这种方法一般要先取得操作系统的源代码，静态地修改操作系统的源码，通过编译、连接后，重新生成一个操作系统。因此，内核藏文本地化比较方便，也比较彻底，如 Linux Tibetan Desktop。外挂藏文本地化通常是在无法获得操作系统的源代码的情况下，在操作系统启动后，通过藏文补丁程序，动态地修改操作系统中有关信息处理部分的代码，如 Tibetan Star（藏文之星）等。

应用程序的藏文本地化也分为两个方面：一是用户界面的藏文本地化。用户界面的藏化较为简单，只要把相应的西文和汉字界面信息翻译成藏文即可；二是应用程序中藏文的通行（有时也称“程序藏文本地化”）。由于在大多数西文应用程序中包含了对诸如非法字符的检测、过滤等，系统层提供的藏文信息会被这些程序检测为非法字符，从而被过滤，导致藏文信息无法通行。这方面的藏化需要专门的技术才能完成。

6. 藏文字形识别技术

藏文字形识别是根据模式识别原理，通过藏字的字形信息识别，产生藏字的内码，实现藏字的识别输入。字形识别输入的原理是通过抽取代表未知藏字模式本质的表达形式与预先存储在计算机中的标准藏字的模式表达形式的集合逐一进行匹配，用一定的准则进行判别，找出最接近输入藏字的那个标准藏字。

7. 藏语语音识别技术

藏语语音识别技术利用产生声音的物理模型，通过语音分析手段，预先将语音特征提取

出来，并存储在处理系统中。当语音信号输入时，处理系统根据对该信号所提取的特征参数和所存储的参考特征进行比较，通过逻辑判断方法和距离测量法等对语音进行识别。

8. 激光照排技术

所谓激光照排，实际上是电子排版系统的大众化简称。激光照排是将文字通过计算机分解为点阵，然后激光照排技术控制激光在感光底片上扫描，用曝光点的点阵组成文字和图像。

由于藏字的自身特点，藏字信息的电子排版系统在20世纪90年代以前是十分落后的。1994年，北大方正推出激光排版系统的藏文部分，字库与华光藏文字库相似，字体有正体、普黑和美术体，输入方法分藏文字形输入法、梵藏文字母顺序输入法。中国藏学研究中心、青海民族印刷厂、山东潍坊华光等合作开发出了“华光藏文照排系统”，这个系统中字库的字符达4 000多个，字体美观，但字形仅有正体和黑体两种；1994年，藏文照排系统“华光V型”已投入书刊的排印工作。

9. 文本分类

文本分类是一种确定文章所属类别的情报分析方法。文本自动分类（automatic text categorization）就是利用计算机对文本集（或其他实体或对象）按照一定的分类体系或标准进行自动分类，属于同一类别的文本被标上相同的类别标记，为文本信息的检索提供系统化的解决方案。随着藏文文本的丰富，藏语文本自动分类技术将会成为藏文信息处理领域一个很好的研究方向。

10. 信息检索

文本信息检索包括了文本信息的存储、组织、表现、查询及存取等各个方面，其核心为文本信息的索引和检索。藏文信息的检索也将成为一个重要的研究领域。

1.5 藏文信息处理的发展历史

从20世纪80年代起，北京、上海、西藏、甘肃、青海等地的一些院校及科研机构纷纷开始了藏文信息处理的研究，研制开发了许多藏文信息处理系统，推动了藏文信息处理技术的发展。藏文信息处理技术得到了党和国家领导人的高度重视，取得了较好的成绩。以下从藏文字符处理、藏语自然语言处理、藏文软件本地化以及藏文信息处理在应用领域的研究等方面对藏文信息处理的发展历史进行回顾。

1.5.1 藏文字符的处理

字的处理技术是信息处理技术发展的前提，藏文字符处理技术的研究是最早的，取得的成绩也是最好的，在字符属性统计、键盘布局、编码、输入等方面都取得了很好的成绩。

1. 字符属性

藏文基本属性的研究是藏文信息处理技术的基础，其涉及面较广、统计难度大，但其