

书面藏语排序的数学模型及算法

江 荻¹⁾ 康才峻²⁾

¹⁾(中国社会科学院民族学与人类学研究所计算语言学重点实验室 北京 100081)

²⁾(北京理工大学自动控制系 北京 100081)

摘要 针对中国国家标准及 ISO 藏文编码字符集提出书面藏语字词的排序涉及藏字结构序、构造级和字符序概念,是不同于中文、英文序性而性质独特的一种排序。文章详尽分析了藏字字形、结构形态、传统字符顺序以及藏字字长和层高等特征,构建出藏语排序的数学模型。然后依据模型要求为每类藏文符号进行数字赋值,通过算法逐步确定字符位置并识别字符,最后按照抽取字符的对应数值组合排序,完成了藏语字词的排序。该模型现已在 Windows 平台上实现。

关键词 藏字; 结构序; 构造级; 字符序; 计算机排序

中图法分类号 TP18

The Sorting Mathematical Model and Algorithm of Written Tibetan Language

JIANG Di¹⁾ KANG Cai-Jun²⁾

¹⁾(Key Laboratory of Computational Linguistics, Institute of Ethnology & Anthropology,
Chinese Academy of Social Sciences, Beijing 100081)

²⁾(Department of Automation, Beijing Institute of Technology, Beijing 100081)

Abstract According to GB16959-1997 and ISO/IEC 10646-1:1993 of coded character set for Tibetan information processing, there is an engineering need for applying the set to all kinds of software and databases, in which sorting is an important technology. As Tibetan sorting involves construction order, classes of constitution and character sequence in the dictionary order, A Written Tibetan word has an inconceivably complex structure with multi-hierarchies. The paper makes an exhaustive analysis to the structures of words, the order of construction categories, and the sequence of characters in each structural position, as well as the length of words and the hierarchies of vertical composition stacks, and then establishes a sorting mathematical model. On the basis of the analysis, the paper assigns distinctive values to all existing characters with numerals in a word, then step by step identifies each character in the words with special algorithm and match it with character-numeral lists. At last, the paper combines all the values extracted from characters of words and compares different combination to make an ordered arrangement for any words in Tibetan language. This processing strategy has been accomplished in Windows 2000/NT Operating System.

Keywords written Tibetan; construction order; classes of constitution; character sequence; sorting by computer

1 引言

本文讨论书面藏语字词的计算机排序问题。

1998 年颁布的中国国家标准《信息交换用藏文编码字符集基本集》(GB16959-1997)^[1] 及相关国际标准^[2]只包括了 41 个藏文编码字符(含藏文和梵音藏文),加上其它组合用字符及篇章装饰或标点类符号

共计 168 个^①. 按照这个编码方案以及藏文构字的二维结构方式, 计算机处理藏文时必须采用编码字符的线性排列与纵向叠置的复合方式来构造藏语字词. 由于这两方面的因素, 计算机不能简单依据编码字库的字符顺序为藏语文本语料处理或其他应用项目提供传统的藏语词典排序, 也就是说, 涉及书面藏语字词排序时需要构建特定排序模型及设计相应的算法.

2 藏语的结构序与构造级

为了清晰地说明藏语的排序问题,有必要简略地讨论书面藏语字符构字的两种基本构造以及传统的藏文字符序列. 图 1(a)和图 1(b)是藏文音节字(简称藏字)的构造形状,可称为藏字的基本结构. 每

个字最多可以由 7 个字符构成,形成 7 个字符构字的结构位置. 其中 Ba 是基本辅音位, Pr 是前置辅音位, Up 是上置辅音位, Lw 是下置辅音位, Vo 是元音位, 包括上置的 e,i,o 三种形式和下置的 u 形式, Sx₁ 和 Sx₂ 分别是后置辅音位和重后置辅音位^②. 尽管藏字在拼写上按照 Pr—Up—Ba—Lw—Vo—Sx₁—Sx₂ 的顺序书写(参见图 1(c)), 在下文的处理中, 后置辅音与重后置辅音合并为一类, 即 Sx, 字的排序却要按照传统约定的词典顺序 Ba—Pr—Up—Lw—Vo—Sx₁—Sx₂. 这种按约定顺序、由构字字符的位置区分先后顺序所形成的排序结构称为藏字的构造级, 以拼写顺序来看, 基本辅音位置级别最高, 对应数值为 0, 其他结构位置是: 越后拼写的位置其构造级别越低, 数值越大. 在处理技术上, 以 0~5 表成相应的构造级.

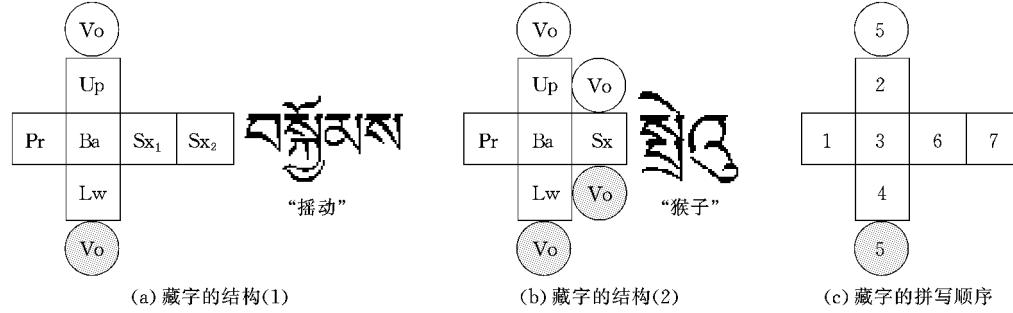


图 1

还有两个因素是藏语排序所需要考虑的, 第一, 藏字每一个结构位置上都存在数量不等的字符集合, 这些字符同样存在传统上约定的前后排列顺序, 即字符序. 如基本辅音位置上的 30 个字符, 一定按照传统藏文字母排列, 是儿童从小习得的藏文知识. 第二, 藏字的 6 个结构位置(技术处理上, 后置与

重后置辅音合一), 除基本辅音位置外, 其余位置都可能为空, 这是语言文字构造上的随机性质决定的. 因此, 空缺位置相同的字形组成一类集合, 每类字形称为一类结构形态, 不同结构形态之间也形成固定的排列顺序. 观察图 2 的几种结构形态有利于理解藏语字词的排列顺序.

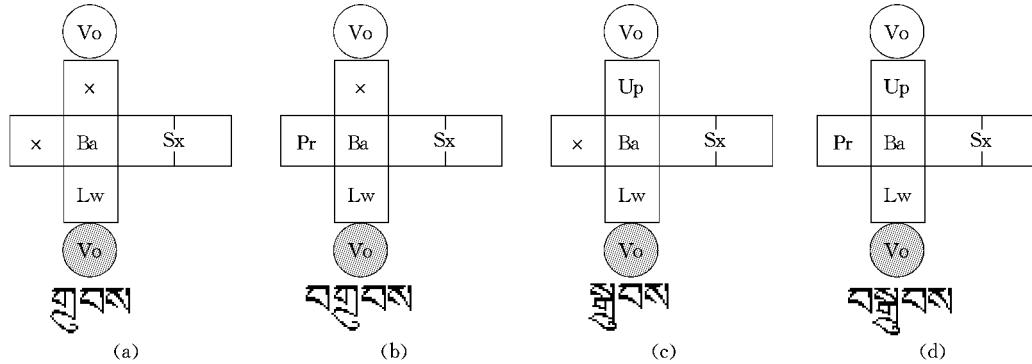


图 2

由于下置辅音、元音和后置辅音的缺省并不影响藏字的排列顺序, 技术上也易于处理, 所以, 通常只需要考虑前置辅音和上置辅音是否缺位的情况(缺位用×表示). 为此, 对藏字排序时, 实际只需区

^① 传统书面藏语只有 30 个辅音字母和 4 个元音符号, 并且不包括上置和下置辅音及变形符号, 而这些符号在计算机处理时字形及位置与基本辅音差异较大, 国标中另设组合类符号表示.

^② 元音符号在国标编码中设计为组合类符号, 图中阴影部分表示出现上置元音则无下置元音, 反之亦然.

分处理四种结构形态. I 型: 缺省前置辅音(Pr)和上置辅音(Up), 如图 2(a); II 型: 缺省上置辅音(Up), 如图 2(b); III 型: 缺省前置辅音(Pr), 如图 2(c); IV 型: 前置辅音和上置辅音都出现, 如图 2(d).

不同的结构形态按顺序排列形成藏字的结构序, 以上四种结构形态的结构序是: I 型, II 型, III 型, IV 型. 同时, 这种结构形态讨论还包括了图 1(b) 所反映的基本结构, 尽管这种结构中的 Sx 实际是一种准基本辅音, 但在排序问题上, 可以当作后置辅音处理.

结构序、构造级、字符序, 这三项内容是决定书面藏语排序的根本性因素, 而且相互交织出现在整个排序过程中, 因此藏语的排序具有其自身的独特性. 在下面的讨论中先分析结构序和构造级所产生的排序问题, 字符序问题放在下一节, 与字符的赋值一起讨论.

书面藏语的结构序和构造级综合起来的结构位置排序分别是:

I 型结构形态: 基本辅音位, 下置辅音位, 元音位, 后置辅音位(包括重后置辅音位);

II 型结构形态: 基本辅音位, 前置辅音位, 下置辅音位, 元音位, 后置辅音位;

III 型结构形态: 基本辅音位, 上置辅音位, 下置辅音位, 元音位, 后置辅音位;

IV 型结构形态: 基本辅音位, 前置辅音位, 上置辅音位, 下置辅音位, 元音位, 后置辅音位.

基本排序方法是, 从构造级最低的位置开始, 各个结构位置逐一顺序交替变换字符, 每当下一级位置的字符轮换完毕, 则进入上一级位置, 直到最终遍历每一级位置, 排序转入下一级结构形态, 并重复以上字符变换过程. 具体的排序过程是:

I 型的排序. 后置辅音位的字符全部循环穷尽后, 元音位字符才开始变换, 一旦元音字符变换, 则循环又从构造级最低的后置辅音开始循环, 如此反复, 直到元音位字符全部循环完毕后, 才开始下置辅音位的循环, 并且又从构造级最低的后置辅音位开始循环, 遍历下置辅音位和元音位的各个字符. 当 I 型后置辅音位、元音位、下置辅音位循环结束后, 除非该基本辅音构成的藏文字不再有 II、III、IV 型, 否则必须进入下一个结构形态(不排除某型空缺), 如 II/III/IV 型的排序. 只有当 I、II、III、IV 各型全部循环后, 基本辅音方可变换, 完成一个完整的循环, 排序进入下一个基本辅音的循环. 如果某个构造级空缺, 则后续级依次递补, 仍然遵照本规律.

II 型的排序. 按 I 型的排序循环结束以后, 再开始前置辅音位上的字符循环.

III 型的排序. 按 I 型的排序循环结束以后, 再开始上置辅音位上的字符循环.

IV 型的排序. 按 I 型的排序循环结束以后, 再开始前置辅音位和上置辅音位上的字符循环.

3 藏语排序的数学模型

上文已经指出, 藏字的 6 个结构位置都存在数量不等的字符集合, 为此可以设基本辅音集合为 B , 上置辅音集合为 U , 前置辅音集合为 P , 下置辅音集合为 L , 元音集合为 V , 后置辅音集合为 S . 考虑到藏字的结构形态除基本辅音外其它构造级均可为空(元音构造级缺省时的零形式实际为 a), 并假设各种组合都存在相应的藏字, 那么, 在理论上书面藏语所有可能字形可由以下笛卡尔乘积表示:

$$B \times U \times P \times L \times V \times S =$$

$$\{(b, u, p, l, v, s) \mid b \in B, u \in U, \\ p \in P, l \in L, v \in V, s \in S\}$$

其中, (b, u, p, l, v, s) 为各结构位置上的字符集合中元素构成的序组. 该笛卡尔乘积构成一个完整的封闭的全集合. 也就是说, 可以这样理解藏文的构造模型:

$$\{S(a, i), A(i), a \in A(i), i \in T\}$$

其中, T 为构造级的有序集合(构造级按升序排列); i 为某一具体构造级; $S(a, i)$ 为构造函数, 而 $A(i)$ 为 $S(a, i-1)$ 确定的前提下, 在构造级 i 上的可用字符的集合; a 为在可用字符集 $A(i)$ 中采用的字符.

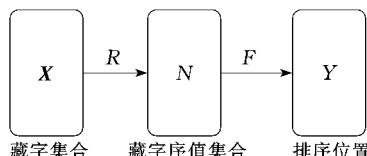
本文认为, 该构造模型类似于数学中递归函数的调用. 当 i 为基本辅音级时, $A(i)$ 为所有基本辅音字符的集合, 当选定 $a \in A(i)$ 后, $S(a, i)$ 被唯一地确定, 此时, $S(a, i)$ 为仅具有基本辅音级的字符结构; 再次重复构造模型, 此时 i 为上置辅音级, $A(i)$ 为在上一级选取的基本辅音的基础上, 所有可用上置辅音字符的集合. 值得强调的一点是, 此时的上置辅音集合不一定就是所有上置辅音的集合, 比如, 如果相对于上一步选取的具体的基本辅音, 只有两个上置辅音能构成有效藏字, 则上置辅音集合仅包括这两个可用的上置辅音字符, 以下各级构造模型均与此相同. 当选定 $a \in A(i)$ 后, $S(a, i)$ 再次被唯一地确定, 此时, $S(a, i)$ 为具有基本辅音级与上置辅音级的字符结构, 并继续以上过程, 直至 i 遍历整个构造级集合 T 后, 被唯一确定的 $S(a, i)$ 才是完整的藏字. 这样的构造模型就能满足实际存在字符的情况. 该构

造模型的示意如图 3 所示。



图 3 排序模型

藏字在结构上具有 6 个结构位置,在进行单个字的排序时,可将藏字实际存在的字形视为多维空间上离散的点,每一个字形都唯一对应一个相应的 6 维矢量 $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5, x_6)^T$. 其中, x_1 代表基本辅音位上的字符序值, x_2 代表前置辅音位上的字符序值, x_3 代表上置辅音位上的字符序值, x_4 代表下置辅音位上的字符序值, x_5 代表元音位上的字符序值, 而 x_6 代表后置辅音位上的字符序值. 当矢量 \mathbf{X} 中的所有分量都被确定的时候,其所对应的藏字也就确定了. 若将藏字在排序表中的位置视为 Y , 排序规则视为 F , 序值视为 N , 序值求取规则视为 R , 则矢量 \mathbf{X} 、序值 N 与排序位置 Y 存在图 4 所示的关系.



冬

即 $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5, x_6)$, $N = R(\mathbf{X})$, $Y = F(N)$.

由图4可见,对藏字进行排序,最重要的是对其对应矢量各分量依次进行序值的定位,这是藏文排序的计算机算法的关键所在。

在矢量 \mathbf{X} 中, 分量 x_1 代表的是基本辅音位的序值. 因为基本辅音是藏字的核心, 在构造上不会出现空缺的情况, 所以 x_1 是最重要的分量, 本文以对 x_1 的序值的确定作为藏字排序的开始.

4 字符序及其赋值

藏语的排序还需要考虑梵文转写为藏文(梵音藏文)的情况,虽然其中存在多种复杂的特殊格式,

但仍可以按照藏文方式构造和排序,同时,梵音藏文字符的顺序也可以根据传统约定顺序嵌入藏文一并处理^[3].

由于计算机藏文字符本身不具备编码顺序,为此有必要给每个具体字符赋值. 赋值的方法是采用数值作为字符的代码, 其中既要考虑字符的结构位置, 还要考虑字符自身的顺序(字符序)和字符的数量. 各结构位置上, 凡字符数量超过两位数的, 用两位数值表示, 否则只用一位数值表示. 字符的顺序按数值升序排列.

(1) 基本辅音序。按照传统约定藏文字符次序排列,从梵文转写来的叠置字符作为单一字符处理(符合国家标准),其顺序排在相关的藏文字符后面。请观察表 1。

表 1 基本辅音排序及赋值表

字符	赋值	字符	赋值	字符	赋值	字符	赋值
𠂇	01	𠂔	11	𠂅	21	𠂆	31
𠂇	02	𠂄	12	𠂈	22	𠂉	32
𠂇	03	𠂊	13	𠂋	23	𠂌	33
𠂇	04	𠂎	14	𠂎	24	𠂏	34
𠂇	05	𠂐	15	𠂑	25	𠂒	35
𠂇	06	𠂔	16	𠂔	26	𠂔	36
𠂇	07	𠂔	17	𠂔	27	𠂔	37
𠂇	08	𠂔	18	𠂔	28	𠂔	38
𠂇	09	𠂔	19	𠂔	29	𠂔	39
𠂇	10	𠂔	20	𠂔	30	𠂔	40

(2)前置辅音序. 藏文有5个前置辅音,梵文转写无前置辅音. 因为前置辅音位可能为空,故需设置数值00表示空位. 另外,上置辅音、下置辅音和后置辅音都存在相同的情况,同样设置00表示. 这一类还包括带辨识符标志赋值的情况,可参见第5节的讨论.

表 2 前置辅音排序及赋值表

字符	赋值	带辨识标志赋值
𠂇	00	10
𠂊	01	11
𠂉	02	12
𠂔	03	13
𠂎	04	14
𠂏	05	15

(3)上置辅音序. 藏文只有3个上置辅音,但梵音藏文中充当上置辅音的字符甚多

表 3 上置辅音排序及赋值表

字符	赋值	字符	赋值	字符	赋值	字符	赋值	
00	𠂇	07	𠂅	14	𠂆	21	𠂈	28
01	𠂉	08	𠂊	15	𠂋	22	𠂌	29
02	𠂄	09	𠂅	16	𠂃	23	𠂅	30
03	𠂆	10	𠂇	17	𠂅	24	𠂇	31
04	𠂉	11	𠂊	18	𠂁	25	𠂉	32
05	𠂉	12	𠂊	19	𠂆	26	𠂉	
06	𠂉	13	𠂊	20	𠂅	27	𠂉	

(4)下置辅音序与元音字符序.无论单一的下置辅音还是叠置的双下置辅音,本文处理时均作为一个排序单位.元音符号中也包括了梵音藏文.参见表 4 和表 5.

表 4 下置辅音排序及赋值表

字符	赋值	字符	赋值	字符	赋值	字符	赋值	
0	𠂉	2	𠂊	4	𠂁	6	𠂉	8
1	𠂉	3	𠂊	5	𠂁	7	𠂉	

表 5 元音字符排序及赋值表

字符	赋值	字符	赋值	字符	赋值	字符	赋值	
01	𠂉	03	𠂊	05	𠂁	7	𠂉	9
02	𠂉	04	𠂊	06	𠂁	8	𠂉	10

(6)后置辅音序.后置辅音包括双后置辅音形式、梵音藏文形式和一些特殊形式(如随音点“ \circ ”、涅盘点“ \circ ”),以及图 1(b)中的粘着性叠置元音形式.参见表 6.

表 6 后置辅音排序及赋值表

字符	赋值	字符	赋值	字符	赋值	字符	赋值	
00	𠂉	11	𠂊	22	𠂁	33	𠂉	44
01	𠂉	12	𠂊	23	𠂁	34	𠂉	45
02	𠂉	13	𠂊	24	𠂁	35	𠂉	46
03	𠂉	14	𠂊	25	𠂁	36	𠂉	47
04	𠂉	15	𠂊	26	𠂁	37	𠂉	48
05	𠂉	16	𠂊	27	𠂁	38	𠂉	49
06	𠂉	17	𠂊	28	𠂁	39	𠂉	50
07	𠂉	18	𠂊	29	𠂁	40	𠂉	51
08	𠂉	19	𠂊	30	𠂁	41	𠂉	52
09	𠂉	20	𠂊	31	𠂁	42	𠂉	53
10	𠂉	21	𠂊	32	𠂁	43	𠂉	54

区分构造级并为藏语字符赋值的目的是,产生一个按照构造级顺序以及字符顺序的组合编码.以上赋值对每个藏字都构成一个 11 位定长编码,其中基本辅音占两位,前置辅音占两位,上置辅音占两

位,下置辅音占一位,元音占两位,后置辅音占两位,这是一个以构造级降序从左至右依次排列的构造级序值组合.

5 藏语排序的算法

依据以上讨论,藏语排序的计算机算法可分为以下几步进行.

第一步:确定藏字的位长及基本辅音位.所谓位长,是指藏文字符及叠加组合线性排列构成藏字的字符长度,每个藏字的最大位长数为 4 位.分别是前置辅音、基本辅音或基本辅音位的叠置组合、后置辅音(含重后置辅音).除了基本辅音位置不能为空外,其它位置都可以是空位.因此,藏字的位长最短为一位,最长为四位^[3].

第二步:确定基本辅音位的层高.所谓层高,是指在基本辅音位置上下存在上置辅音、下置辅音和元音的叠置,最高层高为 4 层.又分两种情况,不计元音时(元音有些上置有些下置),最上层是上置辅音,第二层为基本辅音,第三层为下置辅音;第二种情况是,第一层为基本辅音,第二、第三层为下置辅音(叠置).除基本辅音位置不能为空外,其他位置都可以是空位.一旦确定了基本辅音的层高也就确定了基本辅音的位置,随后也就能确定叠置在基本辅音上下的其他辅音的位置,并最终对各位置上的字符赋值.注意,图 1(b)结构中处理为后置辅音的黏着音节也有两个层次,但这种结构永远出现在基本辅音位之后,并不会造成对基本辅音位判断的干扰.关于藏字位长与层高的详细说明以及识别各类辅音位置的流程详情,请参见文献[3],本文不再赘言.

第三步:确定单个藏字的序值.以上第一、二步确定了各结构位置上的具体字符后,调用排序赋值表确定每个字符的序值,然后依据排序模型将字符序值排定为整个藏字的序值.这里还需说明,处理 II 型和 III 型时,简单根据各结构位置上字符序值组合对藏字进行排序将可能发生 III 型出现在前置辅音不为空的 II 型之前的情况,违反了 II 型在前 III 型在后的约定顺序.为此,可以在序值中的前置辅音位增设一位辨识位,当结构形态为 I、II 型时,该辨识位为 0;否则,该辨识位为 1,即上置辅音为空时,其值为 0×,否则,其值为 1×(其中×可为 0~5 中的任意数值).

第四步:在每个藏字的序值确定后,将其序值存入数据表.如果是多字(音节)词,则分别对后续字进

行以上操作,并对不同词的相对应字进行序值组合的比较。因为采用前几步算法计算出的单个藏字的序值通过数值大小体现了藏字结构序、构造级、字符序的全部特征,因此可以经过比较和判断该序值的大小,将对象移到排序表中合适的位置。

本项研究依据以上算法实现了对藏语字词的排序,排序结果符合现行藏语词典的排序,是一种结构简单而又切实可行的排序算法。试举数例如下^[4]:

结构序 传统词典序 数值编码序值(升序)

༄༅	I	1	01000080415
དྲ	II	2	01020000926
西藏	III	3	01102720149
蜀	III	4	01103120453
汉	IV	5	01133100411
藏	I	6	02000000926

6 结语

本文认为,虽然书面藏语系统的传统序性包含了糅合结构序、构造级、字符序等诸多因素在内的复杂关系,但却反映出藏文自身的逻辑性,不仅科学有效,且已为广大使用者所接受。因此,本文设计的计算机排序算法力求符合传统藏文序的规律性。通过



JIANG Di, born in 1954, Ph. D., professor of Linguistics. His research interests include modern Tibetan grammar, computational linguistics, Sino-Tibetan languages.

Background

This paper is a part of the project “The Study of a Dictionary-Based and Rule-based Tibetan Automatic-Segmentation System”, which is supported by National Natural Science Foundation of China (No.: 60173024). The aim of the project is to auto-analyze (automatic-segmentation system) the Tibetan texts with dictionaries. Yet, in processing the dictionary and texts in Tibetan, the authors need to rearrange the words in order from different dictionaries and words ex-

计算机编程检验,该算法已完满达成了上述要求。

另一方面,按照笛卡尔乘积方式描述藏语排序规则,获得的是藏字所有可能的排序结构。然而,语言文字的发展总是具有很大程度随机性的,语言词汇的丰富程度也与具体语言群体的社会发展相关联,因此,现实存在的藏字结构集合必然只是一定量离散元素所构成的集合,仅仅是笛卡尔乘积方式描述藏语排序结构的一个子集。因此,根据现有藏字构造的排序表必定存在着编码数值上的不连续性,这也正反映了语言文字随机发展中所导致的离散特性。

参考文献

- 1 National Standard of PRC. Information Technology, Tibeyan Coded Character Sets for Information Interchange, Basic Set(GB 16959-1997). Beijing: Standards Press of China, 1998(in Chinese)
(中华人民共和国国家标准. 信息技术、信息交换用藏文编码字符集·基本集(GB16959-1997). 北京:中国标准出版社,1998)
- 2 ISO/IEC 10646-1: 1993: Information Technology — Universal Multiple-Octet Coded Character (UCS)
- 3 Jiang Di, Zhou Ji-Wen. On the sequence of Tibetan words and the method of making sequence. Journal of Chinese Information Processing, 2000,14(1):56~64(in Chinese)
(江 荻,周季文. 论藏文的序性及排序方法. 中文信息学报, 2000,14(1):56~64)

KANG Cai-Jun, born in 1980, graduate student. His research interests include pattern recognition.

tracted from texts. So it is necessary to design a computer program to process this problem. To resolve the problem early in 1999, the research group has proposed a Tibetan sorting concept and announced a flowchart of technological processing (reference[3]). This paper has fulfilled the sorting concept with the mathematical model and algorithm. On the whole, the paper builds up a foundation to the project.