第22卷　第4期
2008年7月

中文信息学报
JOURNAL OF CHINESE INFORMATION PROCESSING

Vol. 22, No. 4
Jul., 2008

文章编号: 1003-0077(2008)04-0109-05

基于DUCET的藏文排序方法

黄鹤鸣[1], **契嘎·德熙嘉措**[2] **(赵晨星)**

(1. 青海师范大学 物理系,青海 西宁 810008; 2. 青海藏文信息研究所,青海 西宁 810008)

摘　要: DUCET为每个藏文字符规定了排序码,但藏文音节的拼写复杂性使得藏文排序不能直接应用这些排序码,提出了基于DUCET的藏文音节排序方法,主要思想是: 首先,将二维的藏文音节转化成一维的字母串;其次,从DUCET中查出每个字母的排序码,得到藏文音节对应的排序码串;最后,通过比较排序码串实现藏文音节间的排序。还讨论了藏文音节与一般藏文字母串以及藏文字符串与外文字符串间的比较规则。

关键词: 计算机应用;中文信息处理;藏文字符串;藏文音节;DUCET;排序

中图分类号: TP391.1　　**文献标识码**: A

A DUCET-based Tibetan Sorting Algorithm

HUANG He-ming[1], ZHAO Chen-xing[2]

(1. Physics Department, Qinghai Normal University, Xining, Qinghai 810008, China;
2. Qinghai Institute of Tibetan Information and Technology, Xining, Qinghai 810008, China)

Abstract: DUCET (Default Unicode Collation Element Table) is an international standard of character collation. This paper proposes a method of DUCET-based Tibetan sorting algorithm. It first expands two-dimensional Tibetan scripts into a one-dimension string of Tibetan letters. Then it locates the collation code of each Tibetan letter from DUCET. Finally, by comparing any two distinctive collation code strings, including Tibetan scripts and non-Tibetan scripts, a correct DUCET-based Tibetan order will be achieved.

Key words: computer application; Chinese information processing; Tibetan scripts; DUCET; Tibetan strings; collation

从音节层面上看,藏文音节串(例如: ཁ་བྱང་དང་མཆན་པ)的排序和英文、汉文字符串的排序相同: 一个音节串的顺序由串中各个音节依次决定。但和汉字或者英文字母不同的是,每个藏文音节的顺序还需由构成它的各个字母依次决定。实际上,藏文字符排序的关键就是将音节中各个字母对音节顺序的影响用简洁通用的算法表示出来,如果这个问题解决了,藏文字符排序问题就解决了,因为音节层面上的排序可以利用汉文或英文字符串的排序算法。

和汉文或者英文字符排序比起来,藏文排序较烦琐,因为: ①虽然对于结构完整的现代藏文音节而言,字典顺序是由基本辅音、前加辅音、上加辅音、下加辅音、元音、后加辅音和又后加辅音依次决定的,但对于一般现代藏文音节而言,不能直接应用这个规则来排序; ②和现代藏文音节比起来,梵音藏文音节在字母系统、拼写规则等方面差异较大,排序也相应较困难; ③藏文语法(例如: 格助词)会影响相关音节的拼写。

到目前为止,讨论藏文音节排序的文献主要有文献[1～3],其中文献[1]由于没有考虑像སྐུད这种有上加辅音而没有前加辅音的音节在字典顺序中的特殊性,排序结果不完全符合字典顺序;文献[2]、[3]通过对藏文字典顺序的分析,建立了藏文音节的

收稿日期: 2007-10-18　**定稿日期**: 2008-01-14

基金项目: 信息产业部电子信息产业发展基金资助项目(信部运[2002]393号)

作者简介: 黄鹤鸣(1969—),男,硕士,副教授,研究方向为藏文信息技术,模式识别;契嘎·德熙嘉措(1946—),男,教授,硕士生导师,研究方向为藏文信息技术。

排序模型,得到了符合字典顺序的排序算法,但由于排序码是自行定义的,无法实现藏文字符串和外文字符串的比较。

本文以国际通用的DUCET排序码为基础重点讨论了:①现代藏文音节间的比较规则;②梵音藏文音节间的比较规则;③梵音藏文和现代藏文音节间的比较规则;④藏文音节与不能构成藏文音节的字母串间的比较规则。

1 DUCET简介

藏文字符排序可能涉及的字符串类型有:①现代藏文音节串;②梵音藏文音节串;③一般藏文字母串,指不能构成音节的藏文字母串;④藏文符号串,指由藏文标点符号、图形符号等构成的串;⑤藏文数字串,指由藏文数字、藏文半值数字构成的串;⑥外文字符串,指非藏文字符构成的串;⑦混合字符串,指以上六种字符串中任意几种的混合。要彻底解决藏文字符的排序问题不仅要解决同类字符串间的比较,还需解决不同类字符串间的比较。

由于可能会和外文字符串比较,藏文字符排序码需要遵从国际标准。Unicode的基本排序元素表(Default Unicode Collation Element Table,简称为DUCET)是排序码的一个国际标准,它为所有具有Unicode编码的字符定义了排序码。以下几行是DUCET为几个字符规定的排序码:

```
0F40 ; [.1C22.0020.0002.0F40] # TIBETAN LETTER KA ;
0F57 ; [.1C46.0020.0002.0F56][.1C65.0020.0002.0FB7] # TIBETAN LETTER BHA ; QQCN
0F73 ; [.1C6E.0020.0002.0F73] # TIBETAN VOWEL SIGN II
0F71 0F72 ; [.1C6E.0020.0002.0F73] # TIBETAN VOWEL SIGN II
0020 ; [*0209.0020.0002.0020] # SPACE
```

DUCET正文中,一行定义了一个排序码。每行的内容依次为:字符的Unicode编码、排序码以及注释。排序码一般由四个十六位的权重构成,这四个权重依次称为一级权重、二级权重……四级权重。

"空格"比较特殊,它有两个排序码:需要将它看作普通字符时用排序码[0209.0020.0002.0020],而需要将它忽略时则用排序码[0000.0000.000.0020]。

两点说明:

1) 一个排序码包含四个权重是为了实现一些特殊字符的精确排序,但本文讨论中,一级权重就能满足要求,因此排序码用一级权重来代表;

2) 两个现代藏文字母排序码的大小与它们在现代藏文字母表中的顺序一致,因此本文讨论中,除非必要,否则不列出排序码而是直接利用字母间的这种顺序关系来判断。

显示时,一个藏文组合字符(例如:སྒྲི)只占据一个字母的宽度,这个宽度由第一层辅音决定,这样的辅音称为占位辅音;其他字母只是叠加在占位辅音的上方或者下方而不单独占据宽度,称为不占位字母。例如,组合字符སྒྲི中ས是占位字母而ག、ྲ和ི是不占位字母。一个辅音可能占位也可能不占位,因此,每个辅音有两个Unicode编码,分别表示占位和不占位。例如,占位辅音ཀ的编码是U+0F40,而不占位辅音◌ྐ的编码是U+0F90。

DUCET为每个辅音的占位编码和不占位编码分别规定了排序码。这些排序码满足不等式:

ཀ的排序码<◌ྐ的排序码<ཁ的排序码<◌ྑ的排序码<…<ཨ的排序码<◌ྸ的排序码。

即一个不占位辅音的排序码大于相应占位辅音的排序码而小于下一个占位辅音的排序码。

2 现代藏文音节间的比较

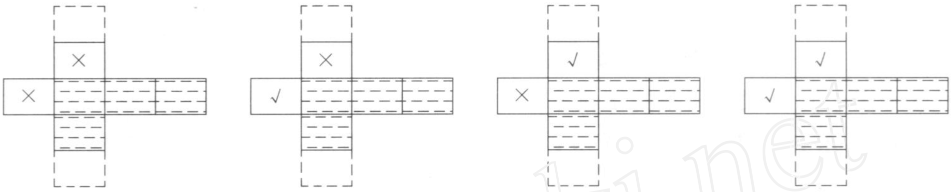
藏文音节往往是二维的而最有可能和它比较的外文字符串(主要是汉文字符串和英文字符串)以及藏文符号串、藏文数字串是一维的。为了实现它们之间的比较需要将二维的藏文音节展成一维字母串。

一个结构完整的现代藏文音节如图1.a所示,它的字典顺序由基本辅音、前加辅音、上加辅音、下加辅音、元音、后加辅音和又后加辅音依次决定(见图1.b)。因此,这样两个音节间的比较可以通过以下步骤实现:首先,将每个音节按图1.b所示的顺序展成一维字母串;其次,查出字母串中每个字母的排序码,得到每个音节对应的排序码串;最后,通过比较排序码串决定两个音节间的顺序关系。

一般的现代藏文音节根据是否包含前加辅音和上加辅音可以分为以下四种类型(见图2):①没有前加辅音和上加辅音,例如:ཀ;②有前加辅音但没有上加辅音,例如:དཀ;③有上加辅音但没有前加辅音,例如:རྐ;④同时有前加辅音和上加辅音,例如:བརྐ。

(a) 结构完整的现代藏文音节 (b) 结构完整的现代藏文音节的比较顺序

图1 现代藏文字的结构和位置序号



①型音节 ②型音节 ③型音节 ④型音节

图2 现代藏文字类型

实际上,只需加一点限定条件,就可以将两个结构完整的现代藏文音节间的比较规则用到任意两个现代藏文音节间的比较中。对于一般的现代藏文音节而言某个位置上的字母可能会空缺,因此在将这样的音节按图1.b的顺序展成字母串时,空缺的位置用排序码较低的字符(例如:空格“□”)代替。

下面讨论这种方法的合理性。当基本辅音不同时,容易根据基本辅音判断出两个现代藏文音节间的顺序关系。因此,只讨论基本辅音相同的情况。

首先,对于类型①和②来说,类型②有前加辅音而类型①没有,将它们按图1.b的顺序展开时,类型①的前加辅音位置为空格“□”而类型②有前加辅音,这样类型①就排在了类型②的前面。例如,表1中音节ཀ在音节དཀ的前面。其次,对于类型③和④来说,它们二者间的区别也是有无前加辅音,因此,同样可以得出类型③在类型④前面的结论,例如,表1中音节སྐ在音节བསྐ的前面。最后,对于类型①、②和类型③、④来说,类型①、②中基本辅音是占位的而类型③、④中基本辅音是不占位的,由于占位辅音的排序码小于相应不占位辅音的排序码,因此,类型①、②在类型③、④的前面。总之,当基本辅音相同时这四类音节间的顺序为:①②③④,这显然符合字典顺序。

例如,不考虑下加辅音、元音、后加辅音和又后加辅音的变化时,基本辅音为ཀ的现代藏文音节有ཀ་དཀ་བཀ་རྐ་ལྐ་སྐ་བརྐ་བསྐ།,将它们按图1.b的顺序展成字母串后比较,比较结果符合字典顺序(见表1)。

表1 现代藏文音节排序结果举例

音节	类型	对应的字母串	排序结果	音节	类型	对应的字母串	排序结果
ཀ	①	ཀ□□□□□□	1	ལྐ	③	ཀ̊□ལ□□□□	5
དཀ	②	ཀད□□□□□	2	སྐ	③	ཀ̊□ས□□□□	6
བཀ	②	ཀབ□□□□□	3	བརྐ	④	ཀ̊བར□□□□	7
རྐ	③	ཀ̊□ར□□□□	4	བསྐ	④	ཀ̊བས□□□□	8

因此,任意两个现代藏文音节间比较时,也是将它们按图1.b所示的顺序展成字母串后比较,但要注意两点:1)将音节展成字母串时如果某个位置上的字母空缺则用排序码较低的字符(例如,空格“□”)代替;2)确定每个字母特别是基本辅音的排序码时需要区分它是否占位。

3 梵音藏文音节的比较

藏民族全民信教,在历史长河中翻译了汗牛充栋的佛教经典。这些经典中借用了大量的梵文词,为了能够准确地音译这些梵文词,创造了与一般藏

文在字母系统、拼写规则等方面差异较大的梵音藏文。

梵音藏文有34个辅音：ཀ་ཁ་ག་གྷ་ང། ཙ་ཚ་ཛ་ཛྷ་ཉ། ཊ་ཋ་ཌ་ཌྷ་ཎ། ཏ་ཐ་ད་དྷ་ན། པ་ཕ་བ་བྷ་མ། ཡ་ར་ལ་ཝ། ཤ་ཥ་ས་ཧ། ཀྵ།；

梵音藏文有16个元音：ཨ་ཨཱ། ཨི་ཨཱི། ཨུ་ཨཱུ། རྀ་རཱྀ། ལྀ་ལཱྀ། ཨེ་ཨཻ། ཨོ་ཨཽ། ཨཾ་ཨཿ。

现在梵音藏文多用在科技名词、人名、地名的音译中。

3.1 梵音藏文音节的词典顺序

文献[5]是首部以梵音藏文音序进行编排的词典，完成了两万五千多个梵音藏文词条的字典排序。本文主要依据该词典来讨论梵音藏文音节间的排序。

一个梵音藏文音节是辅音字母、元音字母以及偶尔用到的辅助读音符的叠合体，梵音藏文没有前加辅音、上加辅音、后加辅音和又后加辅音。由于没有前加辅音，因此音节中的第一个字母是基本辅音而不是前加辅音。例如：ཀེ། མཐུར། བཙན།应该分别属于ཀ部、མ部和བ部；梵音藏文音节也没有上加辅音，当几个辅音重叠时，最上层的辅音不能按上加辅音对待，例如，རྟི། ལྷག། སྟི།应分别属于ར部、ལ部和ས部；但梵音藏文中有下加辅音ྭ、ྱ和ྲ（例如：ཀྭ། ཀྱ། ཀྲ།），由它们构成的音节在排同一辅音和元音ི和ཱི构成的音节之后，例如：ཀིཀཱིཀྭཀྱཀྲ。可以看出一个梵音藏文音节最多有两个辅音。

随意点“ཾ”和涅槃点“ཿ”既可以单独作元音也可以和其他单元音组合成复合元音，因此一个梵音藏文音节最多有两个元音，例如：音节ཧིཿ中的元音为ཨི和ཿ。由于一个梵音藏文音节最多可以有两个辅音和两个元音，因此，可以统一用“辅音＋下加辅音＋元音1＋元音2”来表示一个梵音藏文音节。

梵音藏文的音序排列中，共有50部（34个辅音和16个元音各成一部），这50部按照梵音藏文字母表的顺序排列：ཀ部ཁ部……ཨཾ部ཨཿ部。

其中，第一音节的排列规则是：

ཀཀཾཀཿ ཀཱཀཱཾཀཱཿ ཀིཀིཾཀིཿ ཀཱིཀཱིཾཀཱིཿ ཀུཀུཾཀུཿ ཀཱུཀཱུཾཀཱུཿ ཀྲྀཀྲྀཾཀྲྀཿ ཀྲཱྀཀྲཱྀཾཀྲཱྀཿ ཀླྀཀླྀཾཀླྀཿ ཀླཱྀཀླཱྀཾཀླཱྀཿ ཀེཀེཾཀེཿ ཀཻཀཻཾཀཻཿ ཀོཀོཾཀོཿ ཀཽཀཽཾཀཽཿ（ཀྭ、ཀྱ和ཀྲ构成的音节按照此规则排列）（ཁ至ཀྵ的音节按照ཀ的规则排列）ཨ་ཨཱ། ཨི་ཨཱི། ཨུ་ཨཱུ། རྀ་རཱྀ། ལྀ་ལཱྀ། ཨེ་ཨཻ། ཨོ་ཨཽ། ཨཾ་ཨཿ；

以后各个音节按照此规则进行排列。

需要注意的是占位元音རྀ、རཱྀ、ལྀ和ལཱྀ单独为一个音节时，需要保持顺序关系རྀ、རཱྀ、ལྀ、ལཱྀ并要排在元音ཨཱུ和ཨེ之间，为了实现这点需要修改它们的排序码，此处略过。

3.2 梵音藏文音节间的比较

梵音藏文中，被隔音符分割的可能是一个组合字符也可能是多个组合字符，例如：ཀཀྵ、ཛྙཀྭ、ཤེར་ཕྱིན་ཧྲཱིཿ等，每个组合字符可能是多个辅音和元音的垂直叠加。为了完成梵音藏文的排序，先将组合字符展成音节串，再将音节串展成字母串后才能比较。

首先，按下面的规则将一个组合字符展成音节串：①如果最下层辅音是ྭ、ྱ或者ྲ，则它们和倒数第二层辅音以及元音构成一个音节，其余各层辅音各自构成一个音节，例如：དརྡྲེཿ应展成三个音节ད、ར和དྲེཿ；②如果最下层辅音不是ྭ、ྱ或者ྲ，则最下层辅音和元音构成一个音节，其余各层辅音各自构成一个音节，例如：组合字符རྻྻཿ应展成三个音节ར、ཡ和ཡཿ。

其次，将每个音节按照“辅音＋下加辅音＋元音1＋元音2”的顺序展成字母串，空缺的位置用空格代替。例如，音节“མྱོ”展成字母串为“མ ྱ ོ □”而音节串“མྱོཧྱུ”展成字母串为“མ ྱ ོ □ཧ ྱ ུ □”。

最后，得到排序码串，通过比较排序码串得出音节“མྱོ”在前、音节串“མྱོཧྱུ”在后的顺序关系。

3.3 梵音藏文音节与现代藏文音节之间的比较

梵音藏文音节与现代藏文音节比较时，将它们按各自的规则展成字母串，然后按以下规则比较：1）现代藏文音节中的前加辅音、上加辅音不参与比较；2）现代藏文音节中的基本辅音、下加辅音和元音与梵音藏文的第一个音节比较；3）现代藏文音节中的后加辅音、又后加辅音（如果有的话）各自看作一个音节，依次和梵音藏文的第二、第三个音节比较。

4 藏文音节串与其他类型字母串间的比较

当藏文音节和不能构成音节的一般藏文字母串比较时，首先将藏文音节展成字母串，然后比较两个字母串。例如：音节S1＝“དགེ”和藏文字母串S2＝“ཀཁགང”比较时，先将现代藏文音节“དགེ”展成字母串“ག□ད□ེ□□”后和S2比较，比较结果是音节S1应该在字母串S2的后面。

多文种混合字符串（例如：“ab མྱོཧྱུ”和“ab དགེ་ཅན”）间比较时，同样需要将其中的藏文音节转

换成字母串。

5　黏着型词缀的排序

黏着型词缀在藏文文本中大量存在，它们大多为动词、名词、形容词的后缀。黏着型词缀只有语法意义没有词汇意义因而不能单用。含黏着型词缀的词条有པའི བའི ཕོའི བོའི མའི མོའི ངའི ཁོའི དེའི ངས ཁོས དེས པས བས等。排序时将词缀ས和འི当作后加辅音处理比较方便并且符合字典顺序。

6　藏文字符串的比较过程

综上所述，要实现藏文字符串间的比较，需要经过以下几个步骤：

1. 判断待比较字符串中是否有藏文音节，如果有进一步判断该音节是现代藏文音节还是梵音藏文音节；

2. 将现代藏文音节、梵音藏文音节按各自的规则展成字母串；

3. 从DUCET表中查出每个字母的排序码，得到排序码串；

4. 比较排序码串：梵音藏文音节和现代藏文音节间的比较按第3.3节的方式比较排序码串；其他各类字符串间的比较，顺次比较排序码串。

7　结果分析及结论

本文选取了有代表性的六类字符串来测试本算法的有效性，这六类字符串是：①英文字母串“collation”；②藏文数字串“༡༢༣”和符号串“࿕࿖”；③一般藏文字母串“ཁགང”；④梵音藏文音节串“ཀྲྀ”和“ཀམྃ”；⑤现代藏文音节串“ཀྱུངས”、“དཀོན”、“སྐྲས”和“བསྐུག”以及⑥混合字符串“AB བསྐུག”，排序结果见表2。

表2　本文排序结果与藏文字典顺序对比

比较对象	一维字母串	排序码串（只列一级权重）	本文排序结果	字典顺序
collation	collation	0FFE 113B 10E5 10E5 0FD0 11CF 10A5 113B 111B	࿕࿖	
༡༢༣	༡༢༣	0FC7 0FC8 0FC9	༡༢༣	
࿕࿖	࿕࿖	03DA 03DB	AB བསྐུག	
ཁགང	ཁགང	1C24 1C26 1C28	collation	
ཀྲྀ	ཀ□□□ར□□□ ◌ྲ□□□◌ྲྀ□□	1C22 0000 0000 0000 1C5A 0000 0000 0000 1C3B 0000 0000 0000 1C3B 1C59 0000 0000	ཀམྃ	ཀམྃ
ཀམྃ	ཀ□□□མ□□ྃ	1C22 0000 0000 0000 1C48 0000 0000 1C79	ཀྲྀ	ཀྲྀ
ཀྱུངས	ཀ□□◌ྱུངས	1C22 0000 0000 1C5D 1C71 1C28 1C62	ཀྱུངས	ཀྱུངས
དཀོན	ཀད□□ོན□	1C22 1C3E 0000 0000 1C79 1C40 0000	དཀོན	དཀོན
སྐྲས	◌ྐ□སྲ□□□	1C23 0000 1C62 1C59 0000 0000 0000	སྐྲས	སྐྲས
བསྐུག	◌ྐབས□ུག□	1C23 1C46 1C62 0000 1C71 1C26 0000	བསྐུག	བསྐུག
AB བསྐུག	AB◌ྐབས□ུག□	0FD0 0FE6 1C23 1C46 1C62 0000 1C71 1C26 0000	ཁགང	ཁགང

这个结果说明：本算法能实现和藏文相关的各类字符串间的比较。

感谢：中国社科院的江荻博士在百忙之中阅读了本文的初稿，并提出了一些修改建议，特此感谢！

参考文献：

[1] 林河水，等. 一种符合ISO14651语义的藏文排序实现方法[J]. 中文信息学报，2006，20(2)：94-100.

[2] 江荻，等. 书面藏语排序的数学模型及算法[J]. 计算机学报，2004，4.

[3] 江荻，等. 论藏文的序性及排序方法[J]. 中文信息学报，2000，14(1)：56-64.

[4] Mark Davis and Ken Whistler: Unicode Collation Algorithm [EB/OL]. http://www.unicode.org/reports/；

[5] 安世兴.《梵藏汉对照词典》[M]. 北京：民族出版社，1991.