对 DUCET 中部分藏文字母排序码的商榷

黄鹤鸣[1],契嘎·德熙嘉措[2]
HUANG He-ming[1],ZHAO Chen-xing[2]

1.青海师范大学 物理系,西宁 810008
2.青海藏文信息研究所,西宁 810008
1.Physics Department,Qinghai Normal University,Xining 810008,China
2.Qinghai Institute of Tibetan Information and Technology,Xining 810008,China

HUANG He-ming,ZHAO Chen-xing.Discussion on some Tibetan character's collation code in DUCET.Computer Engineering and Applications,2008,44(29):241-244.

Abstract: The collation code of some Tibetan character in DUCET is not suit for Tibetan syllable's collation.The reason behind that is there are two alphabets in Tibetan:modern Tibetan alphabet and Tibetan transliterating alphabet.These two alphabets share some common letters and some of these common letters have different order in different alphabet.To these letters,only one collation code is not enough.The paper revises the collation code of some Tibetan letters so that they Default Unicode collation Element Table be collated in the desired order.The letters that their collation code have been revised are:(1)Tibetan subjoined consonant ཌ;(2)Tibetan Transliterating consonants ཀ་ཁ་ཆ;(3)Tibetan Transliterating consonants ཁ and ཆ;and(4)spacing vowels ཨ་ཨཱ ཨི ཨཱི ཨུ ཨཱུ ཨྲི ཨཿ.

Key words: modern Tibetan;transliterating Tibetan;collation code;Default Unicode Coalltion Element Table(DUCET);collating

摘　要:DUCET 为每个藏文字符规定了排序码,但部分藏文字母的排序码并不符合藏文字典的排序要求,根据藏文字典的字母顺序对 DUCET 做了以下修订:修订梵音藏文字母ཀ、ཁ、ཆ、ཁ和ཆ的排序码;为占位元音ཨ ཨཱ ཨི ཨཱི ཨུ ཨཱུ ཨྲི ཨཿ赋予了适当的排序码;修订了下加辅音ཌ的排序码。通过这些修订 DUCET 能完全支持现代藏文音节、梵音藏文音节以及二者间的混合排序。

关键词:现代藏文;梵音藏文;排序码;基本排序元素表
DOI:10.3778/j.issn.1002-8331.2008.29.070　**文章编号**:1002-8331(2008)29-0241-04　**文献标识码**:A　**中图分类号**:TP391.1

在音节层面上,藏文音节串(例如:ཀ་ཤ་ཏ་ཡ)的排序规则和汉文字符串相同:一个汉文字符串的顺序由串中各个汉字依次决定,一个藏文音节串的顺序也由串中各个音节依次决定。二者不同的是:每个汉字的顺序直接由排序码确定,而每个藏文音节的顺序进一步由它的构成字母依次确定。因此,要实现藏文音节的排序需要为每个藏文字母赋予适当的排序码。

为藏文字母赋予适当的排序码比较繁琐,主要原因是藏文有两个字母表:现代藏文字母表和梵音藏文字母表,这两个字母表有一些共同字母,而部分共同字母在两个字母表中的顺序不一致。例如:现代藏文字母表中字母ཀ、ཁ、ཆ在字母ཏ、ཐ、ད之后,但在梵音藏文字母表中它们却在字母ཏ、ཐ、ད之前。如果这些字母只有一个排序码,则无论是以现代藏文字母表为准还是以梵音藏文字母表为准,它们之间定会产生逆序。由于 Unicode 的基本排序元素表(Default Unicode Collation Element Table,以下简称为 DUCET)对这些情况没有考虑全面,因此,不能完全支持藏文音节特别是梵音藏文音节的排序。

1 DUCET 简介

藏文字符排序需要解决以下几类字符串间的比较:(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、图形符号构成的串;(5)藏文数字串,指由 10 个藏文数字和 10 个藏文半值数字构成的串;(6)外文字母串,例如,英文字母串;(7)混合字符串,指以上所列 6 种字符串中任意几种的混合。

由于有可能与其他文字的字符进行比较,因此藏文字符排序码需要遵从国际标准。Unicode 的 DUCET 就是字符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。DUCET 正文中,每一行规定了一个字符的排序码,每行的内容依次为:字符的 Unicode 编码、排序码(在中括号内)以及注释(符号"#"后的内容)。以下是 DUCET 中几个字符的排序码:

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

基金项目:信息产业部(No.信部运[2002]393 号)。
作者简介:黄鹤鸣(1969-),男,藏族,副教授,研究方向:藏文信息技术,模式识别;契嘎·德熙嘉措(1946-),男,藏族,教授,硕导,研究方向:藏文信息技术,计算机理论和软件。
收稿日期:2007-11-26　**修回日期**:2008-01-30

0F71 0F72;[.1C6E.0020.0002.0F73] # TIBETAN VOWEL SIGN II

0020;[*0209.0020.0002.0020] # SPACE

字符排序码由4个16位的权重构成，这4个权重分别称为一级权重、二级权重、……。一个排序码包含4个权重是为了实现特殊字符的精确排序，但为了简洁，下文讨论中有时仅用一级权重来代表一个排序码。大多数字符只有一个排序码，部分字符有两个排序码。例如，对于字符“空格”而言，排序时如果需要将它看作一个普通字符则利用排序码[0209.0020.0002.0020]，而如果需要将它忽略则利用排序码[0000.0000.000.0020]。

一个藏文音节是藏文字母的二维组合，几个字母的垂直组合形成藏文组合字符。显示时一个组合字符只占据一个字母的宽度并且这个宽度由第一层辅音确定，这个辅音称为占位字母；其它字母只是叠加在第一层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ཁ্ཁ中ཁ是占位字母而ྐ、ྑ和ྒ是不占位字母。每个辅音字母既可能占位也可能不占位，因此每个辅音字母有两个编码：一个表示占位，另一个表示不占位。辅音的占位编码和不占位编码对应的排序码不同，例如，占位ྐ（U+0F40）的排序码是[.1C22.0020.0002.0F40]而不占位ྐ（U+0F90）的排序码是[.1C23.0020.0002.0F90]。

下面是DUCET中所有藏文字母的排序码，除元音°和ः外，所有字母只列出了一级权重。其中没有变形的辅音只列出了占位编码对应的排序码。

ཀ[1C22] ཁ[1C24] ག[1C26] ང[1C26][1C65] ཚ[1C28] ཅ[1C2A] ཆ[1C2C] ཇ[1C2E]
ཉ[1C30] ཏ[1C32] ཐ[1C34] ད[1C36] ན[1C36][1C65] ཕ[1C38] བ[1C3A]　　ཙ[1C3C]
ཝ[1C3E] ཞ[1C3E][1C65] ཟ[1C40] ཡ[1C42]ཱ[1C44] ར[1C46] ལ[1C46][1C65]
ཤ[1C48] ཥ[1C4A] ས[1C4C]ཿ[1C4E]ཾ[1C4E][1C65]ཧ[1C50]ཱ[1C51]ི[1C52]
ུ[1C54] ཤ[1C56] ཥ[1C58] ཹ[1C59] ཻ[1C5A] ོ[1C5B] ཱ[1C5C] ཿ[1C5E]
ཀ[1C60] ཁ[1C62] ག[1C64] ང[1C66] ཀ[1C22][1C61]ཱ[1C6C] ྀ[1C6D] ཱ[1C6E]
ྀ[1C71] ཱ[1C72] ྀ[1C73]　　ཱ[1C74] ཱ[1C75] ཱ[1C76] ྀ[1C77] ྀ[1C78]
ྀ[1C79] ཿ[1C7A] °[.0000.015A.0002.0F7E] ः[.0000.015B.0002.0F7F]

这些排序码并不完全支持藏文字符的排序，原因如下：(1)4个下加辅音的字典顺序应该是ྱ、ྲ、ྵ、ླ，但根据DUCET中的排序码，它们之间的顺序却是ླ、ྱ、ྲ、ྵ；(2)字母ཅ、ཆ、和ཇ字母ཅ、ཆ、ཇ既是现代藏文字母又是梵音藏文字母，按照DUCET中规定的排序码，字母ཅ、ཆ、ཇ在字母ཀ、ཁ、ག之后，这只符合现代藏文字母表却不符合梵音藏文字母表；(3)梵音藏文排序中占位元音的排序码是必需的，但DUCET中占位元音没有排序码；(4)梵音藏文字母ཌ和ཎ的排序码不符合梵音藏文字典顺序。由于以上原因这些字母的排序码都需要修订。

2　藏文字母排序码的修订原则

DUCET已经为大部分藏文字符赋予了比较恰当的排序码而且排序码资源和字符的编码资源一样是有限的，如果为了部分藏文字符而重新制定所有藏文字符的排序码或者影响到其它文字的排序码是不可取的。因此，本文的修订原则是：①除非确实需要，一般不改变原有藏文字符的排序码；②对于确实需要修订字母，本文通过借用其它藏文字母排序码的方式将其值严格限制在藏文字符的排序码范围内，当然前提是不能产生排序码的冲突。

3　修订部分梵音藏文字母排序码

DUCET中的排序码能正确实现一般藏文字母串、藏文符号串、藏文数字串以及外文字符串间的排序。因此，藏文字符排序的关键是现代藏文音节串、梵音藏文音节串的比较。要正确实现这两类音节间的比较需要利用不同的字母表：

(1)利用现代藏文字母表实现两个现代藏文音节间的比较。现代藏文字母表中辅音间的字母顺序为：ཀ་ཁ་ག་ང། ཚ་ཆ་ཇ་ཉ། ཏ་ཐ་ད་ན། པ་ཕ་བ་མ། ཙ་ཚ་ཛ་ཝ། ཞ་ཟ་འ་ཡ། ར་ལ་ཤ་ས། ཧ་ཨ། 。元音间的顺序为：ཨི་ཨུ་ཨེ་ཨོ། 。

(2)利用梵音藏文字母表实现两个梵音藏文音节间的比较。梵音藏文是一个专门用来音译外来词的藏文文字系统，过去主要用于藏文经典中借用的梵文词的音译，现在主要用于人名、科技词汇（例如化学元素名称）等的音译。梵音藏文与现代藏文在拼读方式、排序规则等方面差异较大。梵音藏文字母表中，辅音间的字母顺序为：ཀ་ཁ་ག་གྷ། ཚ་ཆ་ཛ་ཛྷ། ཏ་ཐ་ད་དྷ། པ་ཕ་བ་བྷ། ཡ་ར་ལ་ཝ། ཤ་ཥ་ས་ཧ། ཀྵ། 。元音间的字母顺序为：ཨ་ཨཱ། ཨི་ཨཱི། ཨུ་ཨཱུ། ཨྲི་ཨཿ། ཨཱི་ཨ། ཨེ་ཨཻ། ཨོ་ཨཽ། ཨཾ་ཨཿ། 。

(3)利用混合字母表实现现代藏文音节和梵音藏文音节间的比较。现代藏文音节和梵音藏文音节比较时，需要将两个字母表中的字母混合形成混合字母表，从而给出所有藏文字母间的一个顺序关系。目前对混合字母表中部分辅音间的顺序关系尚有争议，例如，有些专家建议将反写的梵音藏文字母插ཊ、ཋ、ཌ、ཎ、ཏ到相应的现代藏文字母ཏ、ཐ、ད、ན、ཏ之间，从而形成ཏ、ཊ、ཐ、ཋ、ད、ཌ、ན、ཎ、ཏ的顺序关系等。但大多数专家基本统一于藏文字符集基本集中字母间的顺序关系，并且DUCET已经按这种顺序为藏文字母赋予了排序码，因此，本文建议以基本集中的顺序为准。因此，混合字母表中辅音间的字母顺序为ཀ་ཁ་ག་ཀྵ། ཚ་ཆ་ཛ་ཛྷ། ཏ་ཐ་ད་ཌ། ཏ་ཐ་ད་ཌ། ཡ་ར་ལ་ཝ། ཤ་ཥ་ས་ཧ། ཀ་ཁ་ག། 。元音间的顺序和梵音藏文中元音间的顺序相同：ཨ་ཨཱ། ཨི་ཨཱི། ཨུ་ཨཱུ། ཨྲི་ཨཿ། ཨཱི་ཨ། ཨེ་ཨཻ། ཨོ་ཨཽ། ཨཾ་ཨཿ། 。

3.1　修订梵音藏文字母ཅ、ཆ、ཇ的排序码

现代藏文中字母ཅ、ཆ、ཇ在字母ཀ、ཁ、ག之后，而在梵音藏文中它们间的顺序刚好相反。DUCET中字母ཅ、ཆ、ཇ（一级权重分别为[1C4A]、[1C4C]和[1C4E]）和ཀ、ཁ、ག（一级权重分别为[1C3A]、[1C3C]和[1C3E]）的排序码只符合现代藏文字母表的顺序要求。为了使它们间的顺序关系既满足现代藏文字母表又满足梵音藏文字母表，为字母ཅ、ཆ和ཇ分别赋予两个排序码：(1)现代藏文排序中，仍然使用DUCET中规定的排序码；(2)梵音藏文排序中，为使它们排在字母ཀ、ཁ、ག之前，字母ཅ、ཆ和ཇ分别借用现代藏文字母ཅ、ཆ和ཇ的一级权重[1C2A]、[1C2C]、[1C2E]。由于字母ཅ、ཆ、ཇ不会出现在梵音藏文中，因此这样做并不会引起排序码的冲突。

3.2　修订梵音藏文字母ཌ的排序码

DUCET中字母ཌ的排序码为[.1C51.0020.0004.0FBA][.0000.019A.0004.0FBA]而字母ཎ的排序码为[.1C5C.0020.0002.0F63]，这符合现代藏文中字母ཌ在ཎ之前的要求。但梵音藏文字母表却要求字母ཌ在ཎ之后，为此在梵音藏文字符的排序中，为字母ཌ赋予新的排序码：当字母ཌ占位时，使用排序码[1C5C.0020.0002.0F63][1C51.0020.0002.0FAD]；当字母ཌ不占位时，使用排序码[1C5D.0020.0002.0F63] [1C51.0020.0002.0FAD]。这里将字母ཌ看作了字母ཎ和ད的垂直叠加，这不符合藏文语义，只是利用字母ཎ和ད的排序码为字母ཌ排序而已。

这样ཌ有两个排序码：现代藏文排序中利用[.1C51.0020.0004.0FBA][.0000.019A.0004.0FBA]，而在梵音藏文排序中利用[1C5C.

0020.0002.0F63][1C51.0020.0002.0FAD]（占位时）或者[1C5D.0020.0002.0F63] [1C51.0020.0002.0FAD]（不占位时）。

3.3　修订梵音藏文字母ཁ的排序码

对于梵音藏文字母ཁ、ག、ཁ和ཁ，DUCET 将它们分别看作两个辅音的垂直叠加。例如，将字母ཁ看作了字母ཀ和ཁ的叠加，得到字母ཁ的排序码为[1C26][1C65]（[1C26]和[1C65]分别是字母ཀ和ཁ的排序码），这能使字母ཁ介于字母ཀ和ཏ之间，满足了梵音藏文字母表的要求，但这种方式并不适合字母ཁ。DUCET 中字母ཁ的排序码为[1C22][1C61]，显然是将ཁ看作了字母ཀ和ཁ的垂直叠加。按照这个排序码字母ཁ只能排在字母ཀ[1C22]和ཁ[1C24]之间，这与梵音藏文字母表中字母ཁ在字母ཁ（排序码为[.1C64.0020.0002.0F67]）之后的实际要求相差太大。为了实现正确排序，字母ཁ的占位和不占位编码都借用不占位字母ཁ（[.1C65.0020.0002.0FB7]）时的一级权重，并且二级权重分别取值为 0021 和0022（为了和字母ཁ相区别），从而构成新的排序码[.1C65.0021.0002.0F69]和[.1C65.0022.0002.0FB9]，这样就能使排ཁ在ཁ之后。

3.4　为占位元音字母赋予适当的排序码

梵音藏文字典共有 50 部：34 个辅音和 16 个元音各为一部，也就是说 50 个梵音藏文字母分别按照下列顺序各自构成一部：ཀ་ཁ་ག་ང་ ཚ་ཛ་ཝ་ཞ་ཟ་ ཏ་ཐ་ད་ན་ ཏ་ཐ་ད་ན་ ཕ་བ་མ་ཙ་ཨ་ ཡ་ར་ལ་ཤ་ ས་ཧ་ཨ་ཀ་ ཀྵ་ཡ་ཡ་ ཨ་ཡ་ ཨ་ཡ་ ཨ

SIGN AA。通过这些修订,DUCET 能更好地支持现代藏文音节、梵音藏文音节以及二者间的混合排序。

参考文献:

[1] Davis M,Whistler K.Unicode collation algorithm[EB/OL].http://www.unicode.org/reports/.

[2] 安世兴.梵藏汉对照词典[M].北京:民族出版社,1991.

[3] 国家技术监督局. 信息技术 信息交换用藏文编码字符集 基本集[S].北京:中国标准出版社,1998.

[4] 黄鹤鸣,赵晨星.藏文字符集基本集的修订方案[J].计算机工程与应用,2007,43(20):156-159.

[5] 林河水.一种符合 ISO14651 语义的藏文排序实现方法[J].中文信息学报,18(5):36-41.

[6] 扎西次仁.藏文的排序规则及其排序的计算机自动实现[J].中国藏学,1999(4):128-134.

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

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

[9] Jiang Di.The current status of sorting order of Tibetan dictionaries and standardization[C]//The 20th Pacific Asia Conference on Language,Information and Computation:Proceedings of the Conference.Beijing:Tsinghua University Press,2006:231-236.

(上接 213 页)

纹理折射等几方面进一步加强。

进行图像的三维创建时,应注意以下几点:(1)要选择合适的摄像机标定方法;(2)基元匹配要做到准确,这一步直接影响到最后的创建效果;(3)进行图像投影时,选择好角度;(4)此过程用 Vc++编写(也可借助 Opengl 实现)。

参考文献:

[1] 贾云得.机器视觉[M].北京:北京科学出版社,2000:1-85.

[2] 何斌.Visual C++数字图像处理[M].北京:清华大学出版社,2003:40-97.

[3] 章疏晋.图象理解与计算机视觉[M].北京:清华大学出版社,2002:102-133.

[4] Ma S D.A self-calibration technique for active vision system[J].IEEE Transactions on Robot Automation,1996,12(1):114-120.

[5] Semple J G,kneebone G T.Algebraic Projective Geometry[M].Oxford:Clarendon Press,1952.

[6] Faugeras O,Luong Q T,Maybank S.Camera self-calibration:theory and experiments[C]//Proceedings of the 2nd European Conference on Computer Vision,Italy,1992:321-334.

[7] Luong Q T.Matrix fondamentale et calibration visuelle surl'en vironnement[D].Centre D'Orsay:Universite de Paris Sud,1992.

[8] Maybank S,Faugeras O.A theory of self-calibration of a moving camera[J].Internationl Journal of Computer Vision,1992,8(2):123-151.

[9] Wampler C,Morgan A,Sommese A.Numerical continuation methods for solving polynomial systems arising in kinematics,Technical Report GMR-6372[R].General Motors Research Labs,1988.

(上接238 页)

参考文献:

[1] Pawlak Z.Rough sets[J].International Journal of Mation and Computer Science,1982,11:341-356.

[2] 安利平,仝凌云.粗糙集理论中一种属性离散化算法[J].河北工业大学学报,2002,31(3):39-43.

[3] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.

[4] Han Jiawei,Kamber M.数据挖掘概念与技术[M].北京:机械工业出版社,2001:223-236.

[5] Lingras P,Wset C.Interval set clustering of Web users with rough K-means[J].Journal of Intelligent Information Systems,2004,23(1):5-16.

[6] 张化光,徐悦,孙秋野.基于模糊粗糙集的系统连续变量离散化方法[J].东北大学学报:自然科学版,2008(1).

[7] 叶斌.基于粗糙集理论和聚类分析的全局离散化方法[J].机械与电子,2007(12).

[8] Han J W,Kamber M.Data mining:concept and techniques[M].San Francisco:Morgan Kaufmann Publishers,2001:112-130.

(上接 240 页)

[3] Xu Y.Transmitting tone and intonation simultaneously-the parallel encoding and target approximation(PENTA) Model[C]//Proceedings of International Symposium on tonal Aspects of Languages:with Emphasis on Tone Languages,Beijing,2004:215-220.

[4] 邓丹,石峰,吕士楠.普通话双音节韵律词时长特性研究[R].第七届中国语音学学术会议暨语音学前沿问题国际论坛,2006.

[5] 张保轩.计算机语音处理技术[M].济南:山东科学技术出版社,1996.

[6] Xie Xiaohua,Xie Lingyun.A new approach for tone recognition of isolated mandarin syllables[Z].ISCIT 2006:955-959.

[7] 赵知劲,吴杰.小波变换与自相关结合法在基音周期检测中的应用[C]//中国电子学会第五届青年学术年会文集,1999:119-125.

[8] 杜达.模式分类[M].北京:机械工业出版社,2004.