

文章编号:0253-2328(2011)04-0349-04

在线夏汉电子字典的设计与实现

柳长青^{1,2}

(1. 宁夏大学 西夏学研究院,宁夏 银川 750021; 2. 宁夏大学 数学计算机学院,宁夏 银川 750021)

摘要:利用计算机技术实现了西夏文、汉文和英文之间的在线双向翻译. 首先讨论了在线夏汉电子字典应用程序基本结构,建立了夏汉电子字典关系型数据库. 然后对静态图像和动态水印 2 种方法显示西夏文字进行了研究,并利用模糊查询技术实现了基于汉文、西夏文、英文、四角号码及夏汉字典顺序号等 5 种关键字的自动检索功能. 最终实现了网络下的在线夏汉电子字典应用程序. 在此基础上,初步建立了西夏文献在线文本资源数据库,为今后建立西夏数字化资源库奠定基础. 同时也为少数民族古文字计算机数字化提供了一种可行的方法.

关键词:西夏文; 电子字典; 在线翻译; 图像处理; 信息处理

分类号:(中图)TP391

文献标志码:A

大夏国(1038—1227 年)是党项人李元昊建立的位于中国西北地区的一个古代少数民族政权,其语言文字也被称为西夏文、河西文、番文或唐古特文^[1]. 随着西夏政权的灭亡及党项族的迁移与融合,西夏王朝也逐渐尘封于历史长河之中. 20 世纪初期俄国探险家科兹洛夫在内蒙古额济纳旗的黑水城发掘了大量西夏文古籍文献,西夏文从此才逐渐被人们所广泛关注^[2-3]. 目前,西夏学者对西夏文的解读仍然停留在传统手工翻阅《夏汉字典》的阶段,研究工作因此耗时费力,异常辛苦. 如何把纸介质的《夏汉字典》转换为能够直接应用于网络下的“在线夏汉电子字典”成为亟待解决的问题.

当前,西夏文信息化主要有日本、中国和中国台湾地区学者进行过研究工作. 其中,日本国立亚非语言文化研究所 1996 年制作了西夏文字库,该所副教授荒川慎太郎与俄罗斯西夏学专家克恰诺夫合著了《西夏文字典》,1997 年,中国学者李范文教授和日本学者中岛干起利用该所计算机西夏文排版系统合著并出版了《电脑处理西夏文〈杂字〉研究》一书. 1999 年,由马希荣主编柳长青为主要完成人的国家自然科学基金项目“基于汉字字形的西夏文字研究”的成果“夏汉字处理及电子字典”软件由清华大学出版社正式出版,该成果是按照四角号码和顺序号检字法对西夏字进行排列、注音和释义的 Windows 单

机版应用软件^[4]. 它按照《夏汉字典》从西夏字的音、形、义等方面对每一个西夏字做出了汉、英双语的较为全面的解释,建立了 2 套西夏字库,成为当时在国内外第一个能够独立完整地在个人计算机上进行西夏文、中文和英文互译,并同屏混排、输入、输出的软件产品^[5-6].

上述软件成果在当时的西夏文数字化方面做出了一定的贡献,但随着信息技术的不断发展,这些基于单机版的软件成果已经不能适应越来越高的应用要求. 已有成果大多数出自上世纪末,其软件使用期均已超过 10 a. 由于当时西夏研究原始资料所限上述成果中还存在一定的错误,如:西夏文字形、字义及录入等错误. 因此,本文基于已有的单机版夏汉电子字典,重新建立一套基于互联网并集成最新西夏研究成果的在线夏汉电子字典,同时勘误已有单机版电子字典的错误^[7].

1 在线夏汉电子字典概要设计

1.1 数据库设计

在线夏汉电子字典数据库包含有 4 个基本表:

①西夏字基本表(表 1); ②汉文西夏文检索表(表 2); ③英文西夏文检索表(表 3); ④字典词条数据表(表 4). 其中,西夏字基本表是以《夏汉字典》中 6000 余西夏字为顺序编排.

收稿日期:2011-05-30

基金项目:国家自然科学基金资助项目(60803104);宁夏自然科学基金资助项目(NZ0836);宁夏高等学校科研资助项目(2009)

作者简介:柳长青(1976—),男,副教授,博士研究生,主要从事西夏文信息处理研究.

表 1 西夏字基本表

Field Name	Type	Size	Key
XxzID	Alpha	4	*
Xxz	Alpha	6	

表 2 汉文西夏文检索表

Field Name	Type	Size	Key
Chinese	Alpha	50	
XxzID	Memo	50	

表 3 英文西夏文检索表

Field Name	Type	Size	Key
English	Alpha	50	
XxzID	Memo	50	

表 4 字典词条数据表

Field Name	Type	Size	Key
XxzID	Alpha	4	*
Sjhm	Alpha	6	
En_mean	Memo	50	
Ch_mean	Memo	50	
Jp_mean	Memo	50	
Rus_mean	Memo	50	
Img	Graphic		

表 1~4 中: XxzID 表示西夏字的顺序号从 1~6073; Xxz 表示每个西夏字的 Unicode 编码^[8], 其类型为字符型; Chinese 表示中文检索关键词; English 表示英文检索关键词; Sjhm 表示西夏字的四角号码; En_mean 表示西夏字英文解释, 其类型为备注型; Ch_mean 表示西夏字中文解释; Jp_mean 为预留属性代表西夏字日文解释; Rus_mean 为预留属性代表西夏字俄文解释; Img 表示该西夏字的原始古籍西夏文献切割图. Web 后台数据库系统采用微软 SQL Server 2005 数据库实现, Web Server 采用 IIS4.0 for Windows Server 2003.

1.2 基本结构设计

在线电子字典为用户提供在线查询、在线输入等功能. 西夏文在线电子字典与一般的中英文电子字典的主要区别在于其显示西夏字的特殊性, 在实现一般意义上的电子字典功能时, 还需要考虑到西夏文字符在用户端显示与输入的问题.

图 1 为在线夏汉电子字典的功能模块流程图. 图 1 中用户可以通过模糊查询功能输入查询关键字, 关键字分拣器能够自动分拣出用户所输入的字串是哪一类关键字. 关键字分类主要为: 中文、英文、西夏文、四角号码及西夏文顺序号五大类. 为方便今后扩充, 还特别为俄文、日文等非中英文语言预留了查询接口. 例如: 西夏字“𐵑”, 其四角号码为 212222, 其在夏汉字典中的唯一序号为 2027, 中文解释为“穷、尽、绝、无”, 英文解释为“limit; end”. 按

照图 1 的功能表述, 用户可以通过输入“𐵑”、“212222”、“2027”、“穷、尽、绝、无”和“limit; end”中的任意一个内容作为关键字进行查询, 且所有关键字均可反向检索. 其中, 除按西夏文顺序号一对一检索之外, 其他关键字检索均可以出现一个或多个的检索结果. 图 2 为按照西夏文顺序号 2318 为关键字所得检索结果示例图, 其中黑色粗体为西夏文字符的图像转换图, 西夏文细体小字为西夏文的 Unicode 字符显示, 该 Unicode 字符可直接在常用文本编辑软件中复制、粘贴并与中英文混合排版编辑.

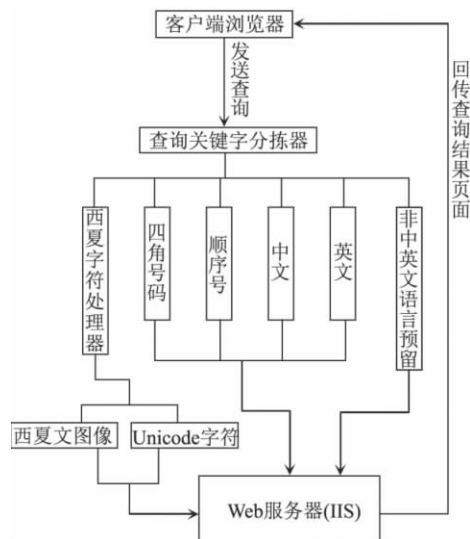


图 1 功能模块流程图



图 2 按顺序号检索结果示例图

2 在线显示西夏文

在线显示西夏文主要采用 2 种方法: ① 图像法, 即将西夏文字库中的每一个 TrueType 西夏字输出为图像并按顺序号命名; ② 直接西夏文本字符法, 该方法将在客户端安装西夏文字库. 以上两种方法亦可结合使用.

图像法主要是通过将已有西夏文 TTF 字库中的曲线字形用图像方式输出并保存和使用. 图像法又分为静态法和动态法 2 大类. 使用静态图像法, 在线夏汉电子字典可以通过西夏文顺序号方式调用服务器端保存的对应西夏字图像来实现客户端的西夏文显示.

2.1 静态图像显示法

TrueType 文件中的“GLYPH”表存放的是 True-

Type 字模数据,存放所有的轮廓描述信息,包括数据信息和指令信息.此表中存放的是一系列的在线和轮廓点的坐标以及作用于字模的指令信息,每个字模的轮廓描述信息都有一个头结构,如下所示:

```
USHORT numberOfContours //轮廓数目;
Word xMin //轮廓点的最小 x 坐标;
Word yMin //轮廓点的最小 y 坐标;
Word xMax //轮廓点的最大 x 坐标;
Word yMax //轮廓点的最大 y 坐标.
```

读取操作主要通过调用 Windows API 来完成,以下是读取字体轮廓线信息的 API 函数:

```
DWORD GetGlyphOutline(HDC hdc, UINT uChar,
UINT uFormat, LPGLYPHMETRICS lpgm, DWORD
cbBuffer, LPVOID lpvBuffer, CONST MAT2 *lpmat2).
该函数的功能是获取指定设备的 TrueType 字体的字符轮廓或位图信息.利用该函数提取轮廓信息可以还原显示 TrueType 字体.其中参数如下:
```

hdc:设备环境句柄;
uChar:指定被返回其数据的字符;
uFormat:指定函数取得的数据的格式.
可用下列值之一,各值含义分别为

GGO_BITMAP:函数获得字形位图.要得到关于内存分配的信息,参见后面备注部分.

GGO_NATIVE:函数获得光栅器(rasterizer)的本地格式的曲线数据点,并使用字体的设计单位,当指定了此值,由 lpMatrix 指定的任何变换都被忽略.

GGO_METRICS:函数只获得由 lpgm 指定的 GLYPHMETRICS 结构,其余缓冲区被忽略.此值影响函数失败时返回值的含义,参见后面的返回值部分.

GGO_GRAY2_BITMAP:函数获得含 5 级灰色的字形位图.

GGO_GRAY4_BITMAP:函数获得含 17 级灰色的字形位图.

GGO_GRAY8_BITMAP:函数获得含 65 级灰色的字形位图.

对 GGO_GRAYnBITMAP 值,函数获得 $n \times n + 1$ 级灰色的字形位图.

lpgm:指向结构 GLYPHMETRICS 的指针,用于描述字表在字符单元的放置.

cbBuffer:指向缓冲区的大小,该缓冲区用于复制轮廓字符的信息.如果此值为 0,函数返回需要的缓冲区大小.

lpvBuffer:指向缓冲区的指针,该缓冲区用于复制轮廓字符的信息,如果此值为 NULL,函数返回需要的缓冲区大小.

lpmat2:指向 MAT2 结构的指针,该结构为字符信息转换矩阵.

返回值:如果指定了 GGO_BITMAP, GGO_GRAY2_BITMAP, GGO_GRAY4_BITMAP, GGO_GRAY8_BITMAP, 或 GGO_NATIVE 值且函数调用成功,返回值将大于 0,否则,返回值为 GDI_ERROR.如果指定了上述值之一,但缓冲区或地址是 0,则返回需要的缓冲区的字节数.

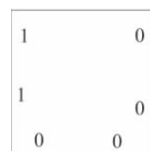
通过该方法显示出的西夏文字属于静态网页显示,如果西夏字库中的西夏字有修正时,则需重新输出修正后的西夏字图像.因此,该方法在更新数据时较为不便.

2.2 动态水印图像显示法

目前,在网络应用中有很多图像验证码都采取动态水印图像的方式来随机产生验证信息,这种方法也可应用在西夏字实时图像显示中.实现文本图像生成功能的组件有很多,其中 AspJpeg 和 XY_Watermark 是常用的 2 种. AspJpeg 是基于微软 IIS 环境的图片处理组件,它可以在 Web 服务器端动态创建高质量的文本图像,完成缩略图片生成、水印图片生成、图片合并、图片切割、数据库支持及安全码技术等功能.它支持目前常用的图像格式. XY_Watermark 组件是一款 IIS 使用的增强 COM 组件.主要用于对 BMP, JPEG 图片生成文字或图片水印,同时具有缩略图功能.本文采用 AspJpeg 组件实现了西夏文的动态实时显示.

首先,定义写入函数 xy(input, i, font, x, y, fontsize, opentime). 其中 input 表示用户输入的西夏字; i 表示目标图片名称; font 表示字体; x, y 表示写入的坐标; fontsize 表示字的大小; opentime 表示图片打开次数.

其次,在模板图像上绘制西夏字.具体实现过程如下:在图片左上角画出对应西夏字的 6 位四角号码中的第 1 位. ASP 脚本实现代码: <% call xy(Cstr(a(0)), array1, "宋体", 3, 1, 8, 1) %> 使用同样方法依次绘制出 4 个角和图 3 四角号码底部的 2 个附号共 6 个号码,见图 3. 绘制图



最后,输出正中间的西夏字水印图(图 4),其 ASP 脚本代码实现为 <% call xy(Cstr(西), array1, "西夏字体", 3, 1, 8, 1) %>.至此,一个完整的西夏字显示图像就可输出.采用水印组件方法,可以有效避免静态法中的西夏字库更新后需要重新输出图像的问题.采用动态水印法,如需更新西夏字则只需更



图 4 完整西夏字显示图

新 Web 服务器端的西夏文字库即可。

3 模糊查询功能

在线夏汉电子字典提供了一个文本输入框来实现查询关键字的输入操作。程序能够根据输入的内容自动判断执行的查询方式,即图 1 所示的“查询关键字分拣器”模块的功能。查询主要分 2 步:①检查用户输入的合法性;②判断通过合法性检查的文本类别,执行相应的查询语句。用户合法性检查主要是检测用户输入的字符是否为字典程序可接受的字符。通过合法性检查后,对用户关键字进行分类,即判断用户输入的关键字是中文、英文、西夏文还是四角号码或顺序号。分类算法见图 5。

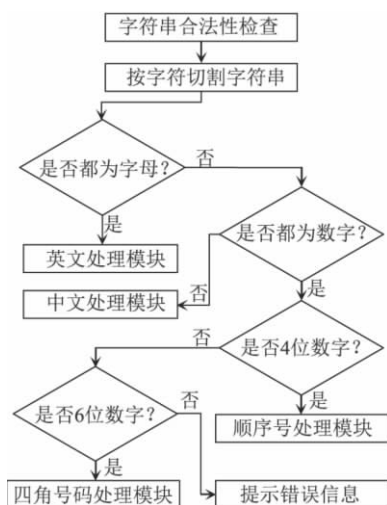


图 5 分类算法流程图

4 结语

西夏文在线电子字典是学习和研究西夏文十分重要的计算机工具软件。“在线夏汉电子字典”不同

于一般的网络字典,其特殊性在于西夏文字本身是一种极不常见的古代少数民族文字。而如何将其准确、稳定地显示于客户端浏览器中,是主要解决的问题之一。本文采用静态与动态水印方法有效地实现了西夏文字的显示。

西夏文字属于表意文字,而作为电子字典其检索方式则较汉字检索方式复杂,主要表现在检索关键字繁多,因此,在线夏汉电子字典的主要功能是如何实现较为方便的检索与查找。通过采用分类算法能够较好实现电子字典的模糊查询与检索。由于西夏文字的局限性,“在线夏汉电子字典”目前还存在与操作系统IME输入法不直接兼容等问题,今后研究的主要方向是如何建立与之相适应的在线西夏文输入法。

参考文献:

- [1] 陈育宁. 宁夏通史[M]. 银川:宁夏人民出版社,2008.
- [2] 史金波,可恰诺夫. 俄藏黑水城文献[M]. 上海:上海古籍文献出版社,1997.
- [3] 史金波,陈育宁. 中国藏西夏文献[M]. 兰州:敦煌文艺出版社,2005.
- [4] 马希荣. 夏汉字处理及电子字典[M/CD]. 北京:清华大学出版社,1999.
- [5] 张青,黄鹤鸣,章登义. 基于 ISO/IEC 10646 标准的藏文编码转换的设计与实现[J]. 中文信息学报,2009,23(4):118-123.
- [6] 柳长青,马希荣. 西夏字与汉字共存方案的实现[J]. 宁夏大学学报:自然科学版,2001,22(1):45-47.
- [7] 柳长青. 网络下的西夏文及西夏文献处理研究[J]. 宁夏社会科学,2008(5):113-115.
- [8] The Tangut UCS Encoding Project. 西夏文和统一码[J/OL]. 2006-07-12 [2007-09-01]. <http://unicode.org/~rscook/Xixia/>.

Design and Implementation of Online Xixia-Chinese Electronic Dictionary

Liu Changqing

(1. School of Xixia Study, Ningxia University, Yinchuan 750021, China;

2. School of Mathematics and Computer Science, Ningxia University, Yinchuan 750021, China)

Abstract: The paper discusses the translation of Xixia, Chinese and English on line. Firstly, it discusses the basic structure of online XiXia-Chinese Electronic Dictionary program, and sets up the relationship database of online XiXia-Chinese Electronic Dictionary. Secondly, it makes a comparison of Xixia character between static image and dynamic watermark, and carries out the automatic index of key words in Chinese, Xixia, English, four corner system and Xixia-Chinese Dictionary order number by using fuzzy query, and finally implements the application program of online XiXia-Chinese Electronic Dictionary. Based on that, a text resources database is created. The study provides a new approach of computer-aided Xixia studies, and lays foundations for the creation of digitalization of Xixia resource in future.

Key words: Xixia characters; electronic dictionary; online translation; image processing; information processing

(责任编辑、校对 张 刚)