第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

少数民族语言文字网站的自动识别和采集

兰义湧[1] 刘海峰[2] 杨媛媛[3]

(中央民族大学理学院 北京 100081)[1] (中央民族大学信息工程学院 北京 100081)[2]

(中央民族大学少数民族语言文学系 北京 100081)[3]

摘 要 分析了少数民族语言文字网站的特殊性,综合采用基于特殊字符、网页标签属性和 N-gram 的方法对传统蒙古文、藏文、阿拉伯字母体系的维吾尔文、哈萨克文和柯尔克孜文以及彝文、新傣文、朝鲜文、俄文和壮文等 10 种少数民族语言文字网站进行了自动识别研究。所提方法对 10 种少数民族语言文字网站的平均正确识别率达到 95% 以上,效果令人满意。

关键词 少数民族语言文字,网站,网页,自动识别,采集

中图法分类号 TP391 文献标识码 A

Minority Language Websites' Automatic Identification and Collection

LAN Yi-yong[1] LIU Hai-feng[2] YANG Yuan-yuan[3]

(College of Science, Minzu University of China, Beijing 100081, China)[1]

(School of Information Engineering, Minzu University of China, Beijing 100081, China)[2]

(Defartment of Minority Language and Literature, Minzu University of China, Beijing 100081, China)[3]

Abstract This paper presented features of Chinese minority script collection on websites, analysed the problems of webpage identification of Chinese minority script, and put forward an identification method. Based on this method, we designed a software to identify and collect Chinese minority language script such as: Mongolian, Tibetan, Uyghur, Kazak, Kirgiz, Yi script Tai Lue script, Korean, Russian, Zhuang script and so on. The average correct identification rate reaches above 95%.

Keywords Chinese minority language, Websites, Webpage, Automatic identification, Collection

我国是一个多民族、多语言、多文字的国家。建国伊始,党和国家领导人就高度重视少数民族语言文字的保护和传承,并成立语言学专家组对我国少数民族语言文字(以下简称民文)进行了系统的梳理、完善和创制等工作,极大地促进了民族地区的文化繁荣和社会进步。新形势下,为进一步推动少数民族地区的社会发展和促进非物质文化遗产的保护,在国家的大力支持下,民文在数字化、信息化等方面取得了长足发展。近些年来,计算机的普及和网络的覆盖,尤其是智能手机的使用,更进一步促进了民文信息化的发展,一些民文网站应运而生,这既方便了少数民族文化的传播和交流,也丰富了我国的网络文化信息资源。至此,民文不只是停留在纸面上的少数民族文化符号,更是当今信息化浪潮中的浪花。

民文网站不仅为少数民族文化的传播、展示和交流提供了一个现代化平台,同时也是我们了解少数民族地区社会和经济发展的一个重要窗口,更是进一步推动我国民文信息化处理技术向前发展的宝贵资源,具有重要的文化价值和资源价值。然而,受民文信息化处理技术相对落后、民文网站的受众有限以及民文信息网络资源生命周期短等因素的影响,一些民文网站常常处于故障频出、影响力小和难以为继的生存窘境。虽然近年来我国的民文网站取得了蓬勃发展,数量和质量也在逐年提升,但与诸如以汉字、英文等大文种的网站资源相比较,还是显得微不足道,要发现它们无异于大海捞针。因此,如何在互联网中及时发现宝贵的民文网站资源,并对其进行采集和储存并加以利用,是目前民文信息处理中一个亟待解决的问题。

1 相关研究

网站的自动识别研究属于文本文字识别领域。文本文字识别方法一般可分为 3 类:一类是基于规则的方法,即利用语言的构词和语法等规则实现文本文字的自动识别,如各种文字所具有的可区分符号和特殊字符[1,2]等;另一类则是从统计学角度建立统计模型实现文本文字的自动识别,主要统计模型有:N-gram[4]、马尔科夫模型[5]和熵等统计模型[6];还有一类采用规则与统计相结合的方法,该方法结合了以上两种方法的优点。

一般认为,1965 年 Mustonen 等[7]提出的根据语言间的特征差异识别不同的文字是文本文字识别的开始。最初的研究主要基于语言规则进行识别。随着计算机技术的发展,自

本文受中央民族大学 2014 年校级自主科研项目(2014MDLXYZY04)资助。

兰义湧(1980—),男,实验师,主要研究方向为自然语言处理、计算机网络;刘海峰(1988—),男,硕士生,主要研究方向为自然语言处理;杨媛媛(1986—),女,博士生,主要研究方向为计算语言学。

然语言文字的识别方法从基于语言规则分析向基于统计方法转变。1994 年 Cavnar[4] 等提出的基于 N-gram 的文本文字自动识别方法是基于统计分析的经典的方法，他们使用 N-gram 方法测试了含有 8 种语言的 3478 篇文档，实现了 99.8% 的正确识别率；同年，Dunning 在 N-gram 的基础上结合马尔科夫模型使识别的准确率提升了 0.1 个百分点，达到 99.9%。之后，有学者将相对熵[6] 和支持向量机[8] 等统计模型应用到文本文字识别中，并在这些算法的基础上使用数据平滑等技术使得识别率达到 99.998%[9]。随着互联网的快速发展，互联网上多种语言文字的网页共存现象日益加剧，这就吸引了越来越多学者对多文种文字识别技术[10-13] 的研究，文本文字的自动识别效果也随着研究的深入而不断加强。

目前，国内外对藏文[14,15]、蒙古文[16] 和维吾尔文[17,18] 等我国少数民族语言文字网页文本的自动识别也开展了一些研究工作，主要采取了规则和统计相结合的方法。公开发表的研究论文中，正确识别率分别达到了 100%、80% 和 97%，而其他少数民族文字网页文本自动识别研究却还鲜有涉及。

2 民文网站及其特殊性

1999 年 12 月，我国互联网上第一个民文网站——“同元藏文网站”正式开通，这标志着中国民文信息网络服务的开端。经过十几年的快速发展，目前在互联网上已经拥有了蒙文、藏文、维吾尔文、哈萨克文、柯尔克孜文、朝鲜文、俄文、傣文、彝文和壮文等民文网站，可谓是百花齐放，绚丽多姿。

通过对大量民文网站的研究发现，民文网站在互联网中主要有 3 种存在形式：(1) 拥有自己的独立域名（如中国藏语广播网：http://www.tibetcnr.com/，青海藏语网络广播电视台：http://www.qhtb.cn/，中国维吾尔语广播网：http://www.uycnr.com/ 等）；(2) 拥有二级域名，作为汉语网站的民文版（如人民网藏文版：http://tibet.people.com.cn/，新华网维文版：http://uyghur.news.cn/ 等）；(3) 作为汉语或其他语言网站的一个子目录（如天山云视频：http://www.ucatv.com.cn/weiyu/，奈曼旗人民政府网站蒙文版：http://www.nmqnw.cn/mgl/ 等）。本文所涉及的民文网站包含以上 3 种存在形式的所有类型网站，即“民文网页聚合而成的独立网站、二级域名的民文网站和挂靠某个网站子目录的民文网页”的统称。

与汉语、英文网站相比，民文网站具有其特殊性。这主要体现在以下几个方面：

由于历史，同种语言拥有多种文字。如：蒙古文有传统蒙古文、托忒蒙古文和新蒙文 3 种文字，维吾尔文有阿拉伯字母和拉丁字母之别，哈萨克文也有阿拉伯字母、拉丁字母、基里尔字母等文字方案，而傣文则有新傣文、老傣文之分等。

在民文数字化的过程中，由于技术滞后和标准不统一等原因，使得同一种文字采用了不同的编码方案。藏文和传统蒙古文尤为突出，如目前藏文就有 Unicode、方正、ASCII、华光、西藏大学、同元、班智达等众多编码方案[15]；而传统蒙古文则有 Unicode、蒙科立、塞因、方正、明安图、Oyuta（智能）和布日古德等编码方案[1,16]。这就造成了部分民族文字编码繁多、互不兼容等一系列问题。

另外，在数字化编码方面，同种文字不同编码方案还存在编码交叉重叠的问题。如在藏文的编码方案中（如表 1 所列），部分基于 GB2312 的编码方案之间存在交叉重叠区；在蒙古文编码方案中（如表 2 所列），部分基于 Unicode 的编码方案之间也存在交叉重叠区。这些问题都会给网页的正确识别带来不利的影响。

表 1 部分基于 GB2312 的藏文编码方案

编码名称	首字节范围	尾字节范围	音节点编码
方正 DOS	0xC0-0xEE	0x21-0x7E	0xC032
方正 Windows	0xAA-0xAC，0xB0-0xDE	0xA0-0xFE	0xAAAC
华光 DOS	0xB0-0xFB	0x21-0x7E	0xE162
华光 Windows	0xB0-0xFB	0xA1-0xFE	0xE1E2
同元编码	0x81-0xEE，0xF5	0x21-0x7E，0x40-0xFE	0xA6E6
西藏大学编码	0xAA-0xAF，0xF8-0xFB	0xA1-0xFE	0xFABB8

表 2 部分基于 Unicode 的蒙古文编码方案

编码名称	编码范围
Oyuta（智能）	0xE250-0xE377
布日古德	0xE246-0xE29F
蒙科立	0xE264-0xE34F
塞因	0xE246-0xE355
明安图	0xE254-0xE33E

在民文网站发展的初期，很大一部分民文网站都基于民间技术，由于受技术力量薄弱和资金支持不足等因素的限制，部分网页源代码书写不够规范、质量良莠不齐。如一些民文网站的网页 Meta 字符集标识（“Charset”和“Encoding”）比较随意，这将给网页的正确解码和识别带来很大的不便。图 1 所示为对网址 http://www.nmqnw.cn/mgl/（奈曼旗人民政府网站蒙文版）解码后产生的乱码页面。

图 1 解码后产生乱码的页面

3 民文网站的识别和采集

一般情况下，对于某些特定网站，我们可以通过工信部或大型导航网站等渠道获得。但是，由于工信部备案的网站并没有报备网站的语言文字，因此我们无法从中筛选出民文网站，而且还有一部分民文网站没有备案过；同时，大部分民文网站由于受众面小、访问量少等因素，也很难被大型导航网站以及知名搜索引擎所收录。因此，民文网站的发现不能仅仅依靠一般的方法来实现，要想比较全面地收集民文网站，就必须采取特殊的识别办法。

在国内已经上线的民文网站有：蒙古文、藏文、维吾尔文、哈萨克文、柯尔克孜文、彝文、傣文、朝鲜文、俄文和壮文等。对于有多种文字的民族语言，本文选取其中使用比较广泛的一种文字方案作为研究对象，即蒙文选择传统蒙古文，维吾尔文、哈萨克文和柯尔克孜文均选择阿拉伯字母体系文字，而傣

文则选择新傣文等。

本文主要分为两部分进行民文网站的识别和采集：首先，识别民文网页文本；其次，采集民文网站。

3.1 识别民文网页

考虑到进行民文网页文本的识别会遇到诸如多编码和编码范围交叉重叠、网页源代码 Meta 标签属性不规范等问题，我们综合采用以下 3 种方法对民文网页文本进行识别，取得了比较理想的准确率。

(1)基于特征字符的识别方法

特征字符是指在其他语言文字中不出现或能够区分出某种文字的字符。例如，藏文中的音节点和下垂符基本上就不会出现在其他的语言文字中，同时这两个字符在藏文中出现频率较高，因此可以作为特征字符来识别藏文文本；另外，在 Unicode 编码方案中，具有独立码段的文字字符集也可以作为特征字符来处理，如基于 Unicode 编码的藏文、传统蒙古文、朝鲜文、彝文和傣文等都可以通过判断字符集所在的 Unicode 编码范围来识别所属文字。

(2)基于网页标签属性的识别方法

一般来说，根据 HTML 网页源代码〈Meta〉标记中的“Encoding”、“Charset”和“Font-family”属性即可判断该网页的所属文字。对于采用非 Unicode 编码标准且具有不同编码方案的少数民族文字，尤其是蒙古文和藏文，其网页源代码中的“Font-family”属性基本上可以表明该网页的所属文字。目前藏文网页常见的“Font-family”字体类型有：BZDBT、BZDMT、BZDHT、TIBETBT、TIBETFG、TIBETCT、TIBETZT 和 TIBETHT 等，蒙古文网页常见的“Font-family”字体类型有：SYMN2008、Sy2008、symn2008f、HudeUI-Saiyin、Saiyinweb-cagantig、Menksoft2012、Menksoft2007、MenksoftQagan、menksoft2013regular、MENKSOF0、MenksoftQagan—mirror、Huritai 和 MGT-MHWT-OT 等，有些蒙古文网页含有〈Meta Name=“generator” Content=“MenkCms Portal-http://www.menksoft.com”〉等信息也可以表明该网页文字为蒙古文。因此，充分利用网页源码中能表明该网页文字身份的信息可以快速辅助识别该网页的所属文字。

(3)基于 N-gram 模型的识别方法

N-gram 模型是一种统计模型，通过对连续序列的概率计算来判断最优结果。该模型基于这样一种假设：连续序列中第 N 个元素的出现只与前面 $N-1$ 个元素相关，而与其它任何元素都不相关，整个序列的概率就是各个元素出现概率的乘积，这些概率可以通过从训练数据中直接统计得出。若将该模型运用于语言问题上，元素则可以是音素、音节、字母、字、词等语言单位，常用的是二元的 Bi-Gram 和三元的 Tri-Gram。N-gram 是文本文字自动识别的经典方法，对于那些不能通过特殊字符识别方法和网页标签属性判断的语言文本文字使用效果较好。例如，基于阿拉伯字母体系的维吾尔文、哈萨克文和柯尔克孜文之间所含的字符绝大部分都相似，特殊字符就不能很好地识别区分它们，而 N-gram 模型则可以做得很好。

综合以上 3 种识别方法的特点，本文采取了如下步骤对民文网页开展识别(流程如图 2 所示)。

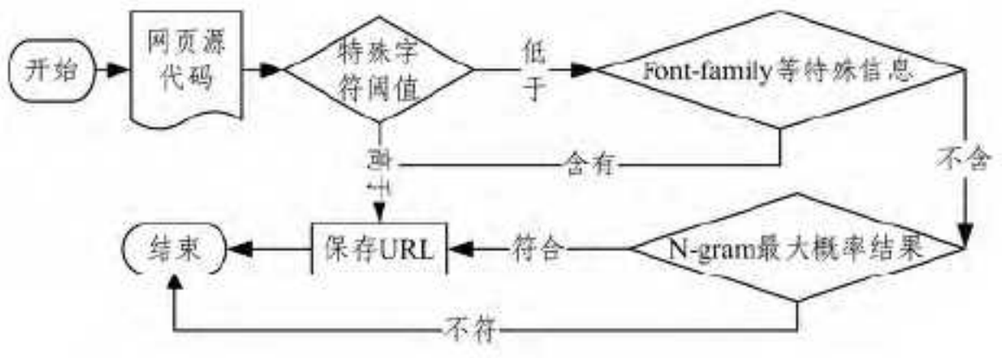

图 2 民文网页识别步骤

1. 统计网页特殊字符频率，若频率达到某种文字一定的阈值，则认为此网页为该文字，保存该网页的 URL；否则进入步骤 2。

2. 检测网页源代码的 META 等信息，若找到能表明某种文字的“Font-family”等信息，则认为此网页为该文字，保存其 URL；否则进入步骤 3。

3. 通过 Bi-Gram 方法识别，若识别出可信度最大的结果为该文字，则认为此网页识别成功，保存其 URL，并结束网页文字识别过程。

3.2 采集民文网站

民文网站文本的采集建立在民文网页识别的基础上，根据已经正确识别的民文网页 URL 地址追溯识别该网页所在的网站，最后对已知的网站使用广度遍历的方法采集更多的网站。

URL 地址追溯：从已经识别出的网页 URL 中抽取出子目录、三级域名和二级域名的信息，再依次对二级域名、三级域名和子目录下的站内网页链接的文本文字进行识别，判断该网站是否为该文字网站。在追溯过程中，如果能够确定该网站使用的是指定的语言文字，则停止追溯，并保存结果到数据库中；否则继续追溯，如图 3 所示。

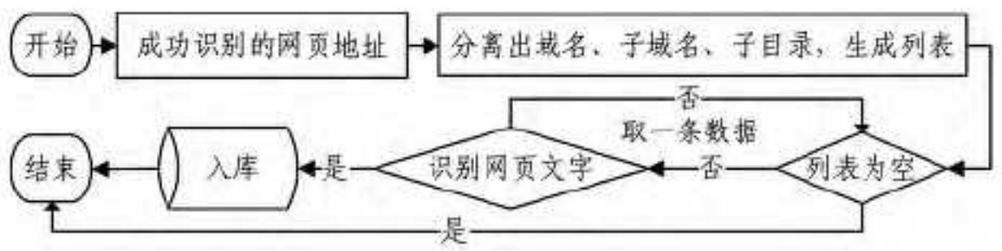

图 3 民文网站识别

网站广度遍历：最初的网页链接来源于知名搜索引擎(谷歌、Bing、百度、360、雅虎等搜索引擎)返回的结果，然而由于知名搜索引擎对民文网站的收录率不高，能够直接通过搜索引擎返回的民文网站数量很有限，因此，还需要通过已正确识别的民文网站上提供的外部链接来加以补充，即对入库的网站通过广度遍历获取同种语言文字的其他网站，如图 4 所示。

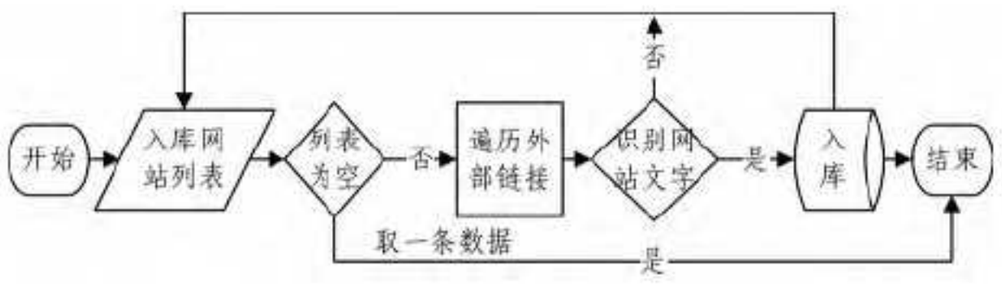

图 4 民文网站广度遍历

4 识别和采集软件的实现

4.1 软件框架

民文网站的识别和采集软件采用 Python 编程语言实现，根据功能不同划分为以下几个模块：用户界面模块、系统设置模块、网络爬虫模块、文本文字识别模块、网站识别模块、网站遍历模块和数据存储模块(见图 5)。各模块之间协调工作，

其中，用户界面模块提供可视化操作界面（见图6），系统设置模块设置采集的语言和关键字、爬虫的线程数、识别的阈值、爬虫爬行深度等参数。最终的网站采集数据结果如图7所示。

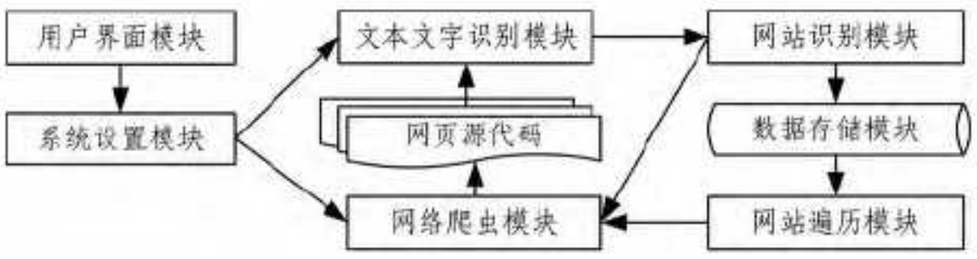

图5 采集软件系统总体框架

图6 软件界面

图7 数据结果

4.2 采集结果

使用上述识别软件对传统蒙古文、藏文、阿拉伯字母体系的维吾尔文、哈萨克文和柯尔克孜文、彝文、新傣文、朝鲜文、俄文和壮文等10种民文网站进行识别采集，得到表3所列的结果（其中网站的详细信息从站长工具获取，网站的服务器地址通过淘宝网IP地址服务器获取）。软件对这10种民文网页的平均正确识别率达到了99%以上，对相应的网站的平均正确识别率达到95%以上。数据显示：(1)俄文和朝鲜文作为跨境文字，其网站主要分布在国外；(2)未备案的网站数量比重较大；(3)在备案的网站中，维吾尔文、哈萨克文和柯尔克孜文的个人网站数量占的比例很大。

表3 10种民文网站采集结果统计

	备案	未备案	个人	企业/团体	事业/政府	国内	国外	总数
蒙古文	88	51	22	17	48	119	20	139
藏文	80	91	38	19	23	93	78	171
维文	869	398	679	163	27	1073	194	1267
哈萨克文	88	47	56	13	19	115	20	135
柯尔克孜文	22	9	18	1	3	25	6	31
朝鲜文	15	286	2	5	8	16	285	301
彝文	2	0	0	2	0	2	0	2
壮文	1	0	0	1	0	1	0	1
傣文	4	0	0	3	1	4	0	4
俄文	2	531	1	0	1	2	531	533

结束语 本文分析了民文网站的特殊性，综合采取基于特殊字符、网页标签属性和N-gram的方法对传统蒙古文、藏文、阿拉伯字母的维吾尔文、哈萨克文和柯尔克孜文以及彝文、新傣文、朝鲜文、俄文和壮文等10种少数民族语言文字网站进行了自动识别，无需人工干预，便可实现较为理想的识别效果。目前对这10种语言文字网站的平均正确识别率已达到95%以上，得到了比较满意的正确率。在以后的研究中，我们将继续完善系统功能，提供更多其他语言文字的识别接口，为尚未进入互联网的其他民文做好技术准备，并努力提高民文网站的查全率。

参 考 文 献

[1] 金良，散旦玛，玉英. 传统蒙古文编码及其应用现状分析[J]. 语文学刊，2012(7)：16-17

[2] Newman P. Foreign language identification：First step in the translation process[R]. Sandia National Labs.，Albuquerque，NM(USA)，1987

[3] Ziegler D. The automatic identification of languages using linguistic recognition signals[D]. State University of New York at Buffalo，Buffalo，NY，USA，1992

[4] Cavnar W B，Trenkle J M. N-gram-based text categorization[J]. Ann Arbor MI，1994，48113(2)：161-175

[5] Dunning T. Statistical identification of language[M]. Computing Research Laboratory，New Mexico State University，1994

[6] Sibun P，Reynar J C. Language identification：Examining the issues[Z]. 1996

[7] Mustonen S. Multiple discriminant analysis in linguistic problems[J]. Statistical Methods in Linguistics，1965，4：37-44

[8] Kruengkrai C，Srichaivattana P，Sornlertlamvanich V，et al. Language identification based on string kernels[C]// Proceedings of the 5th International Symposium on Communications and Information Technologies. IEEE，2005

[9] Brown R D. Finding and identifying text in 900+ languages[J]. Digital Investigation，2012，9：34-43

[10] Yamaguchi H，Tanaka-Ishii K. Text segmentation by language using minimum description length[C]// Association for Computational Linguistics. 2012

[11] Chew Y C，Mikami Y，Nagano R L. Language Identification of Web Pages Based on Improved N-gram Algorithm[J]. International Journal of Computer Science Issues(IJCSI)，2011，8(3)：47-58

[12] King B，Abney S. Labeling the languages of words in mixed-language documents using weakly supervised methods[Z]. 2013

[13] Lui M，Lau J H，Baldwin T. Automatic Detection and Language Identification of Multilingual Documents[Z]. 2014

[14] 藏文网页及其编码的识别方法[Z]. Google Patents，2007

[15] 王思丽. 藏文网页自动发现与采集技术研究[D]. 兰州：西北民族大学，2010

[16] 王睿. 蒙古文网页抓取及编码识别转换研究[D]. 呼和浩特：内蒙古大学，2008

[17] 自动识别网页中维吾尔文的方法及其系统[Z]. Google Patents，2010

[18] 买日旦．吾守尔，维尼拉．木沙江. 电子词典软件系统中对维、哈、柯文进行自动判别技术的研究[J]. 新疆大学学报：自然科学版，2011(01)：88-92