面向维吾尔跨文字搜索引擎的统一转换机制设计

依不拉音·乌斯曼[1] 王 悦[2]
(新疆财经大学计算机科学与工程学院 乌鲁木齐 830012)[1]
(中央财经大学信息学院计算机系 北京 100081)[2]

摘 要 随着近年互联网技术在新疆地区的发展和普及,大量维语网站如雨后春笋般涌现。由于历史原因,维文呈现老维文、新维文、拉丁维文、西里尔维文等多种字母体系共存的"一语多文"的特点。现有的维文搜索引擎仅支持老维文,然而,目前国际通行的主流维语交流字母体系以拉丁维文及西里尔维文居多。由此,如何设计支持维文"一语多文"特点的维文搜索引擎将是维文信息检索研究领域的重要挑战,其研制成果将对广大维族网民的日常互联网使用及国家的"一带一路"战略产生深远的影响。研究拉丁维文、西里尔维文和老维文之间的转换规则;提出 Unicode 字符编码体系和 Unicode 字符编码转换算法,实现在维语搜索引擎系统中通过拉丁维文和西里尔维文来直接检索老维文网页内容,弥补了当前维文搜索引擎系统的空白;通过翔实的实验,验证了所提的 LCCU 编码转换率达到 100%,拉丁维文和西里尔维文的检索效果与老维文完全一致。

关键词 维文信息检索,维文搜索引擎,跨文字转换机制
中图法分类号 TP391 **文献标识码** A

Uniform Converting Mechanism for Cross-characters Search Engine of Uyghur

Ibrayim·OSMAN[1] WANG Yue[2]
(School of Computer Science and Engineering, Xinjiang University of Finance & Economics, Urumqi 830012, China)[1]
(Department of Computer Science, School of Information, Central University of Finance and Economics, Beijing 100081, China)[2]

Abstract With the development of the web technologies in Xinjiang, more and more websites for Uyghur people are on line. Due to the historical reasons, the Uyghur language has many different forms of characters, such as Uyghur ErebYëziqi (UEY), Uyghur Latin Yëziqi (ULY), and Uyghur SirilYëziqi (USY). Current Uyghur search engines only support UEY, however, the most common used characters in international communication are the ULY and USY. Therefore, how to design a search engine to support the multi-characters of Uyghur will be a big challenge for the Uyghur information retrieval area. The related breakout may affect the "The Belt and Road Initiative" deeply. This paper stu-died the converting technologies between UEY, ULY and USY, and proposed the corresponding converting algorithms based on the Unicode coding system. This paper also implemented a uniform converting prototype system to retrieve the contents of UEY webpages through the ULY and USY. We verified our methods converting different characters of Uyghur precisely and smoothly in the experiments. The search results by using ULY or USY reach the same rank of UEY based search engines in our prototype system.

Keywords Uyghur information retrieval, Uyghur search engine, Cross-character converting mechanism

1 Internet 上的维文"一语多文"现象

在历史上维吾尔族使用过很多文字,如古代突厥文、察哈台文、拉丁维文、阿拉伯字母为基础的老维文等。1985 年以前使用拉丁字母为基础的维吾尔新文字,1985 年开始使用现在的阿拉伯字母为基础的老维文(本文中的维文),所以拉丁文在维吾尔社会上有一定的影响力。维文属于阿勒泰语系中的阿拉伯文系统,其文字书写方向为自右向左,和现代中英文的自左向右书写有很大区别。因此在计算机系统中要完成维文的书写及阅读必须在预装维文输入法的基础上,使用其特有的 32 个字母来完成。而由于维文的每个字母在词首、词中及词尾形式的不同,其可能的字形超过 120 多种,这使得老维文的输入法系统拥有复杂的键盘布置设定。

在此背景下,为降低维文书写的复杂度,2000 年新疆大学和新疆维吾尔自治区语言文字工作委员会一起公布了维文与拉丁维文的对应标准,把拉丁维文叫做维吾尔族的计算机网络文字 UCY (UyghurComputerYeziki)。目前拉丁维文已成为大部分 80 年代后出生的以及欧洲国家的维族人群在互

本文受新疆财经大学科研基金:维吾尔语言文字信息化进程研究(2014XYB006),国家自然科学基金(61503422),北京市社会科学基金(15JGC150)资助。
依不拉音·吾斯曼(1974—),男,硕士,讲师,主要研究方向为数据库应用与数据挖掘、自然语言处理,E-mail:1152390290@qq.com;王 悦(1981—),男,博士,副教授,主要研究方向为图数据挖掘、自然语言处理。

联网环境中使用的实际文字标准。此外20世纪初期一部分维族人由于历史原因移民到部分中亚国家，截止目前，中亚3国（哈萨克斯坦、乌兹别克斯坦和吉尔吉斯坦）的维吾尔族人口有54万多，占国外维吾尔族人口的一半以上[11]。而这些中亚国家的维吾尔人采用西里尔字母拼写的西里尔维文作为维语主要的书写文字。在国家“一带一路”战略的推动下，随着新疆地区维族人与欧洲以及中亚国家的维族人间互联网交流的加深，当前维文的使用呈现了3种文字（即老维文、拉丁维文及西里尔维文）共存的特有现象。维吾尔族网络使用文字分布情况如图1所示。

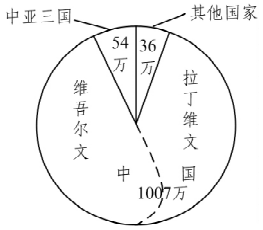

图1　维吾尔族网络使用文字分布图

现有的维文搜索引擎系统面向老维文的网页信息，不支持拉丁维文和西里尔维文，这为广大维族人群的互联网使用带来了极大的不便。由此，本文解决的主要问题为：在现有的维文搜索引擎基础上，设计一套维文多文转换机制，使现有的维文搜索引擎系统可通过拉丁维文和西里尔维文实现跨文字的维文信息检索。本文主要完成的工作如下：

（1）提出拉丁维文和老维文的转换规则，该转换规则依托Unicode编码以及拉丁维文与老维文的语法特征实现，完成拉丁维文及老维文使用时的无缝衔接，提出LTC（Latin-Traditional Conversion，LTC）算法完成实际的文字转换。在此基础上开发拉丁维文转换老维文的文字转换器模块（LTC），通过文字转换器来实现维文搜索引擎系统中通过输入拉丁维文来实现的维文网站的信息检索功能。

（2）研究西里尔文的编码特征，提出西里尔维文和老维文的编码对应法则（Cyrillic-Traditional Conversion，CTC），在此基础上设计出西里尔维文与老维文的文字转换器（CTC）来实现维文搜索引擎系统中使用西里尔维文的信息检索。

（3）通过翔实的实验，验证了本文所提的跨文字转换机制及算法可保证在维文信息检索过程中拉丁维文、西里尔维文与老维文文字编码转换的准确性；同时，原型系统还具有较高运行效率，具有相当的实用价值。

2　相关工作

2.1　维文搜索引擎的现状

随着近年互联网技术在新疆地区的发展和普及，新疆少数民族语言文字网站建设也进入了新的发展阶段，新疆少数民族语言文字网站数量和新疆少数民族网民数量日益剧增。据2014年的统计：新疆少数民族语言文字网站数量超过几万，新疆网民用户突破百万，维文网站在数量、内容、结构、技术等方面有了很大的发展并形成了一定的规模。维文网站共享维吾尔族的文化、历史、娱乐、贸易等海量的信息。目前在网络上正式使用的维文搜索引擎的代表有：新疆大学多语种信息技术重点实验室开发的维哈汉多语中搜索引擎系统[11]、izida[12]、维吾尔网址大全[13]等。

2.2　维文搜索引擎及维语多文编码转换的相关工作

近几年来维文网站的信息检索技术研究也有了长足的进步。文献[2]中介绍了与现代维吾尔文与斯拉夫维文的书写规则和对应关系，提出了现代维文与斯拉夫维文之间的转换规则；文献[4]中介绍了维文编码字符集和拉丁字符的对应关系，提出了维文与拉丁文之间的相互映射关系；文献[5]中研究了传统网络爬虫的缺点，提出了面向增量同生主题的维文爬虫技术；文献[6,7]中提出了基于本体的维文搜索引擎系统的设计策略、实现技术、网络爬虫技术和编码处理技术；文献[3]中提出了维吾尔库中自动获取单词的方法，及基于分段式思略和增量式策略的两种自适应组词算法。除此之外维汉——汉维双向电子字典技术的成熟和词汇量的增加给维文语料库和维文本体（Uwordnet）提供了很好的资源基础。

然而，尽管近年维文搜索引擎的技术取得了进步，现有的维文搜索引擎均是针对老维文而实现的（例如：www.ulinix.com，www.izda.com，www.xjlou d.com等）。当前拉丁维文以及西里尔维文的互联网用户日益增多，国内外网民的相关需求也与日俱增，然而迄今还没有相关工作基于拉丁维文和西里尔维文来实现维文搜索引擎，这是目前技术发展与实际需求的矛盾。虽然新疆大学的吾守尔教授和新疆语言文字工作委员会的亚森老师分别在文献[2,4]中提出了维吾尔文与拉丁维文、维吾尔文与西里尔维文之间的相互映射关系和转换规则，但现有工作均未给出实际可用的转换算法。由此，本文在接下来的章节中提出一套统一转换机制来尝试解决这一矛盾。

3　维吾尔跨文字统一转换机制及算法设计

针对现有的维文搜索引擎缺点，本文设计了维吾尔跨文字统一转换机制（Latin Cyrillic Conversion to Uyghur，LCCU）。以元搜索引擎的方式，使用LCCU转换算法，在现有的维文搜索引擎系统上增加拉丁维文及西里尔维文的信息检索功能，让其为广大维族网民服务。其与现有维文搜索引擎的结合方式如图2所示。

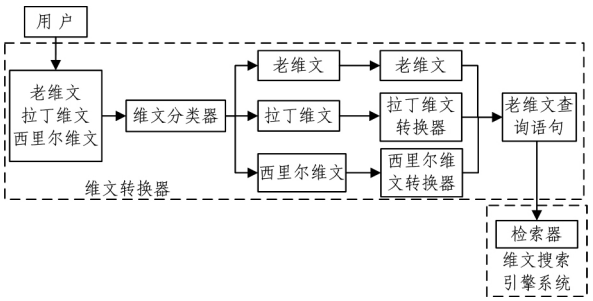

图2　维吾尔跨文字统一转换引擎结构图

3.1　维文分类模块设计

由于互联网上维文搜索引擎用户主要使用的文字分为3类：老维文、拉丁维文和西里尔维文。根据Unicode编码范围设计了维文分类模块，如表1所列。根据老维文、拉丁维文和西里尔维文的Unicode编码范围来分类不同编码的维吾尔语文本。通过原型系统运行测试，该分类模块可以准确识别3种不同的维文。在该分类器的基础上，本节余下部分设计了合理的维文转换机制。

表 1　维文 Unicode 分类范围

文字	Unicode 编码范围
老维文	第六区(0600-06FF)
拉丁维文	拉丁基本区,编码 00 开头
西里尔维文	第四区,编码 04 开头

3.2　LCCU-维文转换器

LCCU(Latin Cyrillec Conversion to Uyghur)维文转换器把用户输入的老维文、拉丁维文或西里尔维文查询语句统一转换成老维文查询语句,并传给搜索引擎检索器。它由 3 个子模块组成:LTC(Latin Traditional Conversion)拉丁维文转换老维文子模块,CTC(Cyrillic Traditional Conversion)西里尔维文转换老维文子模块,UCB(UyghurCharacterBase)维文字库。考虑到没有安装维文字库的计算机,LCCU 模块中添加了维文字库,这样,该模块在没有安装维文字库的计算机系统中也能正常运行,实现用户通过输入拉丁维文或西里尔维文查询语句来使用维文搜索引擎系统,LCCU 的编码转换流程图如图 3 所示。

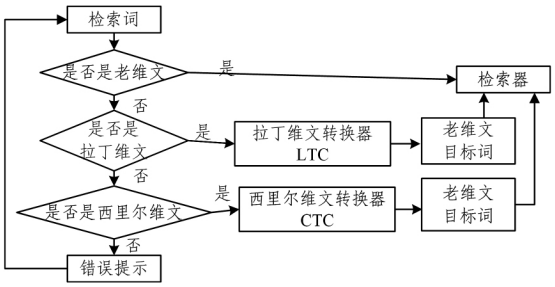

图 3　LCCU 编码转换流程图

3.3　LTC 拉丁维文/老维文 Unicode 编码转换算法设计

LTC 拉丁维文转换器的主要功能是把用户输入的拉丁维文查询语句转换成老维文查询语句转给检索器。维文有 32 个字母,8 个元音和 24 个辅音字母;拉丁维文是用键盘上的 26 个拉丁字母来表示维文中的 32 个字母的文字,也可以叫做维吾尔计算机文字(UCY),表 2 为老维文与拉丁维文的对照。

表 2　老维文与拉丁维文对照表

老维文	拉丁维文	老维文	拉丁维文	老维文	拉丁维文
ئا	Aa	ق	Qq	س	Ss
ب	Bb	ج	Jj	ش	SH sh
ى	Ii	ك	Kk	ۇ	Üü
د	Dd	ل	Ll	ۆ	Uu
ە	Ee	م	Mm	ۈ	Öö
ې	Ëë	ن	Nn	ۇ	Oo
ف	Ff	ڭ	NG ng	ي	Yy
گ	Gg	خ	Xx	ر	Rr
غ	GH gh	پ	Pp	ز	Zz
ھ	Hh	چ	CH ch	ژ	ZH zh
ت	Tt	ۋ	Ww		

(1)转换规则

从表 2 可知,维吾尔语 32 个字母中的 ئا،ب،ى،پ،ت،ج،ش،د،ر،ز،س،ف،ق،ك،گ،ل،م، ، ي،خ،ۇ،ۆ،ئو،ن،، 等 24 个字母直接对应拉丁文基本区中的 A,E,F,S,Z,R,D,X,J,T,P,B,Q,K,G,L,M,N,H,O,U,W,I,Y 等 24 个字母,维吾尔字母中的ش،ڭ،غ،چ، ژ 4 个字母用拉丁文中的 ch,gh,ng,sh,zh 等字母组合表示,剩下的 3 个维文字母用拉丁文扩展区中的 3 个字母来对应。老维文和拉丁文是完全不同的两种文字,所以书写和转换时必须遵守如下规则:

1)输入拉丁维文时缩略语必须在用大写字母的同时中间加一空字符来写,否则转换时出现错误。如 BDT:بدت(错),JKP:جكپ(错)。

2)输入的拉丁词中先后出现的两个字母不是字母组合而是单字母时,为了避免转换程序将其误认为字母组合,两个拉丁字母中间加音节符“`”。例如,不加音节符的转换:

ügengin(学习)—ئۈگەڭگىن(错误)。

yüzhatireqilmak(给面子)—يۈزاتىرەقىلماق(错误)。

加音节符后的转换:

ügen`gin(学习)—— ئۈگەنگىن(学习)。

yüz`hatire qilmak(给面子)——يۈزخاتىرە قىلماق。

3)A,E,O,U,I,Ö,Ü,Ë等 8 个元音字符在词首出现时,维文元音字符的词首字形“ ئا،ئە،ئو،ئۆ،ئۇ،ئۈ،ئى،ئې”来特殊对应处理。

4)标点符号和数字采用一对一关系来转换。

5)维文虽然有 32 个字符,但每个字符在词中出现的位置不同,有 120 多个字形,所以转换后必须进行自动选型。

(2)维文 Unicode 字符编码体系

ISO 没有为维文专门开辟代码区并把维文字符不连续的分配到阿拉伯基本区,阿拉伯扩展区 A,阿拉伯扩展区 B。如图 4 所示,图中箭头指向的两个黑色模块是阿拉伯编码区,维文单字符基本在 0627—06AD 范围,维文不同字形在 FA8C—FEEE 范围。维文的这种分配法对维吾尔字母的自动选型、排序、数据库操作非常不利。

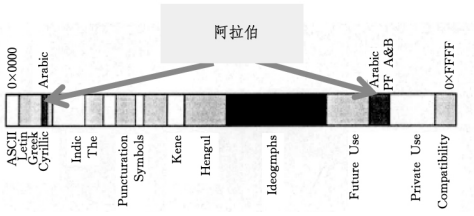

图 4　Unicode 编码布置图

(3)LTC(Latin Traditional Conversion)转换算法

由于拉丁文编码是按顺序排列的,而维文 Unicode 在阿拉伯编码区是分散分布的,所以拉丁维文与维文 Unicode 编码间有不连续的现象,因此通过查询表的方法可以提高查询速度和查询的正确率。预先编制出每一个拉丁维文编码对应的 Unicode 编码,再将其按照拉丁维文编码的排列循序排列,存储于一个二维数组变量 *LatinUyghurChar* 中,把它叫做查询表;再创建一个维文字库表 *UyghurCharacterBase*,字库表中保存着维文 32 个字母的 125 种字形。维文 Unicode 编码表可以从 Unicode 组织的 FTP 上下载,网络上也有很多其它下载源,查询表通过文本的方式描述了每一个拉丁维文所对应的老维文 Unicode 编码,查询表的第一字节是拉丁字母编

码，第二字节是维文字母的 Unicode 编码，如图 5 所示。

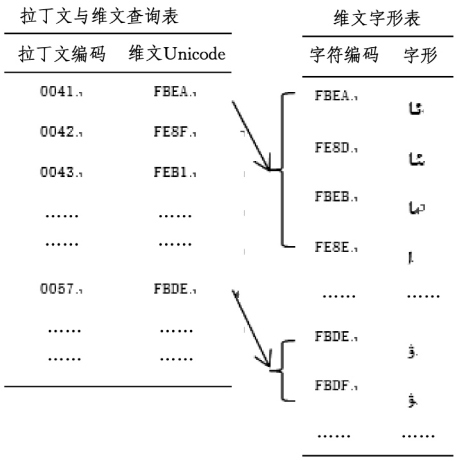

图 5 LTC 转换算法中的查询表关系图

因为维吾尔字母在词中出现的位置不同，有 2～4 种字形，所以还要建一个维文字形库，先从查询表中找到拉丁维文对应的维文 Unicode 编码后，自动选型模块通过该拉丁维文在词中的位置参数和对应的维文 Unicode 编码从维文字形库中找到相应的维文字形（如：拉丁字母 A 对应的维文字母 ئا،ا）中的一个。

LTC 算法首先把老维文与拉丁维文对照表（见表 2）转换成相应的查询表（LatinUighurchar），按照维文字符的不同字形的分类顺序创建一个维文字形表（Uighur CharacterBase）以供自动选型模块使用，LTC 算法的具体操作步骤如表 3 所列。

表 3 Latin Traditional Conversion 算法步骤

LTC 算法步骤
1. Public LatinUighurchar[2,36]，UighurCharBase[3,36] as string/＊定义拉丁文与维文对照数组和维文字库＊/
2. 根据表 2 输入内容，创建拉丁文与维文对照表。
3. 由不同维文字形创建维文字库 UighurCharacterBase。
4. 用户输入的拉丁维文字符串中读取第一个拉丁字符的编码。
5. 判断读取的字符是拉丁维文、空格或音节符等，如果是空格，初始化位置参数返回第四步，如果是音节符或拉丁文，继续下一步。
6. 使用二分法在 LatinUighurchar 中查找对应的行数。
7. 读取第二字段的维文编码和位置参数一起转给自动选型模块。
8. 自动选型模块根据字符编码和位置参数从维文字库中读取相应的维文字形来生成维文目标查询语句。
9. 检查循环是否结束（循环变量是否大于拉丁维文字符串长度），如果是，则生成的维文查询语句传给搜索引擎检索器，否则返回第 4 步骤继续循环。

图 6 是 LTC 转换模块的操作流程图，图中 n 是用户输入的拉丁维文查询词的字长。

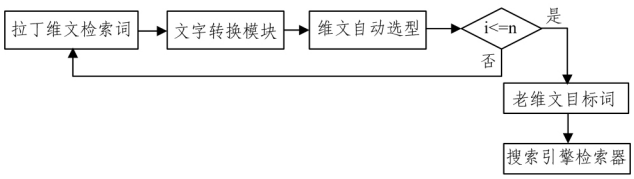

图 6 LTC 算法流程

3.4 CTC 西里尔维文/维文 Unicode 编码转换算法设计

西里尔文有 40 个字符，也有大小写、自从左向右读写的拼音字母，维文中的 24 个辅音字母和 6 个元音字母一对一地对应西里尔文中的 29 个字母，剩下的 1 个辅音字母和 2 个元音字母在西里尔文中各对应两个西里尔文字母，如表 4 所列。

表 4 西里尔维文与老维文字母对照表

西里尔	维文	西里尔	维文	西里尔	维文
Б	ب	Ж	ژ	М	م
П	پ	С	س	Н	ن
Т	ت	Г	غ	һ	ھ
Ж	ج	Қ	ق	В	ۋ
Ч	چ	Ф	ف	Я	يا
Х	خ	К	ك	А	ئا
Д	د	П	گ	Э е	ئە
Р	ر	Ц	ڭ	йы	ئى
З	ز	Л	ل	Шщ	ش

(1)转换规则

1)维文字母中的 23 个辅音字母和 6 个元音字母的字符编码应直接转换为一对一的西里尔维文字母的编码，例如：كال（牛）应转换为 капа，كال（手）应转换为ш。

2)维文字母“ش”与两个西里尔维文字母“ш，щ”对应。

3)维文中的两个元音字母“ئە、ئى”分别对应两个西里尔维文字母“э，а”和“и，Ы”，所以用一对多的对应关系来转换。

4)西里尔维文单词中的“Я，ю”字母转换为维文中的双字母“يا，يۇ”来处理。

5)西里尔维文分大小写字母，句首字母和缩略语中的字母是大写字母，例如：“АЛП”（阿里木），“ШУАР”（新疆维吾尔自治区），转换时使用维文的词首字母中间加一空格，例如：ش ئۇ ئا ر（新疆维吾尔自治区）。

(2)CTC(Cyrillic Traditional Conversion)转换算法

西里尔维文有 40 个字母，而维文有 32 个字母，所以有些字母之间存在一对多的关系，如维文字母中的“ش،ئى،ئە”等字母在西里尔维文中对应两个字母，查询表中通过重复输入维文字符编码的方式可以描述一对多的关系，查询表的第一字节是按顺序排列的西里尔维文字符编码，第二字节是对应的维文字符的词首编码，当然还有在上面创建的维文字形表，CTC 算法中的查询表和字形表之间的关系如图 7 所示。

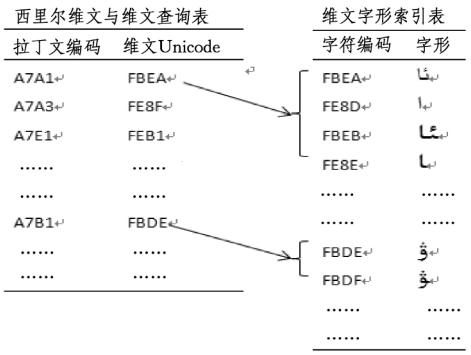

图 7 CTC 算法中的查询表关系图

先读取第一个西里尔维文字符的编码，在查询表中找到对应的维文字符编码，然后在维文字形表中读取对应的维文字形来转换第一个西里尔维文字母，CTC 算法的流程如图 8 所示。

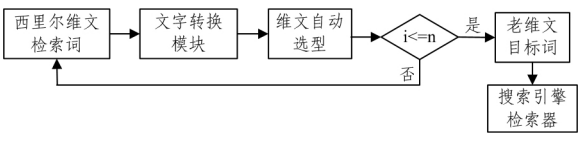

图 8 CTC 算法流程

5 实验与实验结果分析

5.1 实验设定

本节对维吾尔跨文字统一转换机制维文转换器 LCCU (Latin Cyrillic Conversion to Uyghur)进行了实验研究与测试,并对测试结果进行了分析。实验分两个步骤进行:第一步是维文转换器 LCCU 转换功能的测试;第二步是新疆大学多语种信息技术重点实验室开发的维哈汉多语中搜索引擎系统平台下 LCCU 转换器转换的拉丁维文和西里尔维文进行的信息检索,检索结果与老维文的信息检索结果进行了比较。

(1)实验数据

收集新疆发展、新疆地名、新疆旅游景区、维吾尔人的生活、文化等各方面的 100 个关键词(例如:"一带一路"、"新疆经济发展"、"喀什"、"吐鲁番"、"那拉提"、"新疆足球"、"新疆水果"、"新疆舞蹈"等)作为实验数据。

本文所有的实验均在一台处理器为 Intel(R)Pentium(R) CPU B940@2.00GHz,内存为 4GB 的笔记本 PC 上完成。

(2)评价标准

1)采用人工方式和 UyghurSoft 公司开发的 ALKORIC-TOR(检查维文文档文字的拼写和语法)软件来测试 LCCU 编码转换器的效率。

2)对于西里尔维文和拉丁维文在搜索引擎系统中的检索效率,针对同一词语,先找出它们的查准率和查全率(见式(1)和式(2)),通过其与老维文的查准率和查全率比较的方式来进行评价。

$$R(\text{查全率})=\frac{\text{检出的相关文献数量}}{\text{检索系统中相关文献总量}}\times 100\% \quad (1)$$

$$P(\text{查准率})=\frac{\text{检出的相关文献数量}}{\text{检出的文献总量}}\times 100\% \quad (2)$$

5.2 文字转换模块的实验

(1)拉丁维文转换器 LTC 的转换效率测试

通过 100 个维吾尔拉丁文关键词来进行测试,测试当中需特别注意转换器的双字母和单字母的分别能力,维吾尔字母的词首、词中、词尾的自动选型能力,音节符和各种标点符号的识别能力。实验结果表明,拉丁文转换器的转换正确率已达到 100%,如表 5 所列。

表 5 LTC 模块试验结果

序号	拉丁文	转换结果	原维吾尔文	中文含义
1	Birbelbaghbir yol	بىر بەلباغ بىر يول	بىر بەلباغ بىر يول	一带一路
2	Turpan üzümbayrimi	تۇرپان ئۈزۈم بـ ايرىمى	تۇرپان ئۈزۈم بايرىمى	吐鲁番葡萄节
3	Shinjangpotbol	شىنجاڭ پوتبولى	شىنجاڭ پوتبولى	新疆足球
…	…	…	…	…
…	…	…	…	…
100	Uyghur usuli	ئۇيغۇر ئۇسۇلى	ئۇيغۇر ئۇسۇلى	维吾尔舞蹈

(2)西里尔维文转换器 CTC 的转换效率测试

以同样的方式输入 100 个不同的关键词来测试西里尔文转换器的转换效果,转换结果表明西里尔文转换器的准确率达到 96%,转换结果如表 6 所列。

表 6 CTC 模块实验结果

序号	西里尔文	转换结果	维吾尔文	中文含义
1	Шўтжацйнтйш	شوتجاڭ ىخى	شوتجاڭ ئىخى	新疆经
	арй	سلى	سلى	济
2	ШрнжацротБол	شىنجاڭ پو	شىنجاڭ	新疆足
		تبول	پوتبول	球
3	Тўрпгайймеауг	تۇرپلۇزۇمبلۇى	تۇرپلۇزۇم	吐鲁番
	ь	مى	بلۇىمى	葡萄节
…	…	…	…	…
100	қаШқар	قەشقەر	قەشقەر	喀什

5.3 网页信息检索模块的实验与实验结果分析

(1)网页信息检索模块的实验

搜索引擎系统的性能主要由它的查全率和查准率来评价,在新疆大学多语种信息技术重点实验室开发的多语种搜素引擎系统 www.xjuloud.com 下使用拉丁维文、西里尔维文和老维文等不同的输入方式对关键词"喀什"进行信息检索搜索,搜索结果如表 7 所列。实验结果表明输入拉丁维文和西里尔维文来检索的相关文档数量、查找结果数量等各参数指标跟输入老维文来检索的各种参数基本一样,实验结果分析如图 9 所示。

表 7 查询实验结果

查询词	相关文档	查找结果数	结果相关数	查全率(%)	查准率(%)
قەشقەر	130	156	128	98	87
Qashqar	130	156	128	98	87
ҚаШқар	130	150	125	96	85

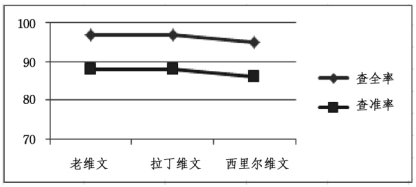

图 9 查询结果分析图

(2)实验结果分析

从图 9 中可以看到 3 种文字在搜索引擎上的查询结果基本上在一条直线上,本文已经实现了拉丁维文和西里尔维文在维吾尔语网站上的信息检索功能,检索效率非常接近输入老维文的检索效率。

用本文中的文字转换器在百度上做实验,不使用文字转换器而直接输入拉丁维文和西里尔维文关键词,百度搜到的结果基本上不是我们想要的结果,通过文字转换器转换后的检索结果中一部分是我们想要的结果。实验结果表明百度检索到的文献数量远远大于维文搜索引擎系统检索到的文献数量,但是查准率非常低,不如维文搜索引擎系统的查准率。原因是百度的分词、网络爬虫等技术比较成熟,网页资源也非常丰富,而维文搜索引擎技术还在初步成长阶段,网页资源也很少,维文网站比较分散,互相链接的比较少,这造成网络爬虫抓取的网页数量大大减少,也是百度检索到的文献数量大的原因,但是百度的分词和网络爬虫等技术不符合维文,所以它的查准率比维文搜索引擎低。

结束语 维文搜索引擎的研制对新疆政治、经济、文化、

社会稳定及教育等多个方面的发展具有非常重要的意义。本文所设计的维吾尔多文统一转换引擎实现了多种维语文字向统一的老维文的准确转换。这使国内外维族网民可以便捷地在不改变自身维文使用习惯的基础上使用现有的维文搜索引擎。由此,本文的研究成果对现有的维文搜索引擎及国内外维族网民的日常互联网使用具有重要的实际意义。未来工作方向包括:进一步提高原型系统的稳定性及处理效率,完成接口通用化;设计一套完整的维吾尔多文统一元搜索引擎系统,结合现有的维文、通用互联网搜索引擎为广大维族网民服务。

参 考 文 献

[1] Turditohti,akbar,askarhamdulla. Adaptive word grouping algorithm based on mutual information in Uyghur language[J]. 计算机应用研究,2013,30(2):82-85

[2] 图尔妮萨塞麦提,吾守尔斯拉木. 现代维吾尔文与斯拉夫维吾尔文转换规则研究[J]. 标准化研究,2013,9:56-59

[3] 吐尔地托合提,维尼拉木沙江,艾斯卡尔艾木都拉. 基于词间关联度度量的维吾尔文自动切分方法[J]. 北京大学学报,2016,52(1):155-162

[4] 亚森依明. 基于国际标准编码系统的维文拉丁文转写规则研究[J]. 标准化研究,2011,6:49-51

[5] 赵永霄,哈力旦阿布不都热依木. 面向增量同生主题的维吾尔文爬虫的研究[J]. 计算机应用,2014,31(11):3269-3272

[6] 李连倍. 基于跨语本体重用的维文本体构造方法研究[D]. 乌鲁木齐:新疆大学,2014

[7] 沙吾提江亚森. 基于本体的维文语义搜索引擎的研究与实现[D]. 成都:电子科技大学,2015

[8] 瓦依提阿布力孜,依不拉音吾斯曼,阿依佐克拉. 提高维吾尔搜索引擎质量的一些关键技术[J]. 数学的实践与认识,2013,43(3):119-122

[9] 艾孜尔古丽,齐向卫,玉素甫. 基于网站用词调查的现代维吾尔语词干提出和应用研究[J]. 计算机应用与软件,2012,29(3):32-34

[10] 刘丽杰. 基于量子行为进化算法的聚集爬虫搜索策略[J]. 计算机应用研究,2012,29(11):4281-4283

[11] 王新青,池中华. 丝绸之路经济带中亚5国语言状况考察与思考[J]. 云南师范大学学报(社会科学版),2015,47(5):14-20

[12] 陈国华,汤庸. 基于学术社区的学术搜索引擎设计[J]. 计算机科学,2011,38(8):171-175

[13] 岑荣伟. 基于用户行为分析的搜索引擎评价研究[D]. 北京:清华大学,2010

[14] 徐戈,王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报,2011,34(8):1424-1433

[15] 江腾蛟,万常选,刘德喜. 基于语义分析的评价对象-情感词对抽取[J]. 计算机学报,2016,39:1-15

[16] 付剑生,徐林龙. 分布式全网职位搜索引擎的研究与实现[J]. 计算机技术与发展,2015,25(5):6-9

[17] http://www.xjuloud.com

[18] http://www.izda.com

[19] http://www.ulinix.com

(上接第76页)

[2] 唐慧霞,李胜利. 超声估测胎儿体重的研究进展[J]. 中华医学超声杂志(电子版),2014,11(5):9-14

[3] Merz E, Lieser H, Schicketanz K H, et al. Intrauterine fetal weight assessment using ultrasound. A comparison of several weight assessment methods and development of a new formula for the determination of fetal weight[J]. Ultraschall in der Medizin (Stuttgart,Germany:1980),1988,9(1):15-24

[4] Schild R L,Sachs C,Fimmers R,et al. Sex-specific fetal weight prediction by ultrasound[J]. Ultrasound in Obstetrics & Gynecology,2004,23(1):30-35

[5] Farmer R M,Medearis A L,Hirata G I,et al. The use of a neural network for the ultrasonographic estimation of fetal weight in the macrosomicfetus[J]. American Journal of Obstetrics and Gynecology,1992,166(5):1467-1472

[6] Cheng Y C,Hsia C C,Chang F M,et al. Cluster-Based Artificial Neural Network on Ultrasonographic Parameters for Fetal Weight Estimation[C]// 6th World Congress of Biomechanics (WCB 2010). 2010 Singapore. Springer Berlin Heidelberg,2010:1514-1517

[7] Cheng Y C,Chiu Y H,Wang H C,et al. Using Akaike information criterion and minimum mean square error mode in compensating for ultrasonographic errors for estimation of fetal weight by new operators[J]. Taiwanese Journal of Obstetrics and Gynecology,2013,52(1):46-52

[8] Mohammadi H,Nemati M,Allahmoradi Z,et al. Ultrasound estimation of fetal weight in twins by artificial neural network[J]. Journal of Biomedical Science and Engineering,2011,4(1):46

[9] Hinton G E,Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science,2006,313(5786):504-507

[10] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition:The shared views of four research groups[J]. IEEE Signal Processing Magazine,2012,29(6):82-97

[11] 王坚,张媛媛. 基于深度神经网络的汉语语音合成的研究[J]. 计算机科学,2015,42(6A):75-78

[12] Krizhevsky A,Sutskever I,Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012,25(2):1097-1105

[13] 王莹,樊鑫,李豪杰,等. 基于深度网络的多形态人脸识别[J]. 计算机科学,2015,42(9):61-65

[14] 李海朋,李晶皎,闫爱云,等. 人脸识别中的遗传神经网络并行实现[J]. 计算机科学,2015,42(6A):168-174

[15] Lipton Z C,Kale D C,Elkan C,et al. Learning to Diagnose with LSTM Recurrent Neural Networks[J]. Computer Science,2015

[16] 孙志远,鲁成祥,史忠植,马刚. 深度学习研究与进展[J]. 计算机科学,2016,43(2):1-8

[17] Uzuner Ö,Luo Y,Szolovits P. Evaluating the state-of-the-art in automatic de-identification[J]. Journal of the American Medical Informatics Association,2007,14(5):550-563

[18] Stubbs A,Uzuner Ö. Annotating longitudinal clinical narratives for de-identification:The 2014 i2b2/UTHealthcorpus[J]. Journal of Biomedical Informatics,2015,58:S20-S29

[19] Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1[J]. Journal of Biomedical Informatics,2015,58:S11-S19

[20] Srivastava N,Hinton G,Krizhevsky A,et al. Dropout:A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research,2014,15(1):1929-1958