ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.11, No.11, April 2015
E-mail: wltx@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于.NET的维哈柯多语种网上数据采集系统的设计与实现

再吐娜木·阿巴白克力[1],侯存义[1],米尔阿迪力江·麦麦提[2],张立新[1]

(1.乌鲁木齐职业大学 现代教育技术中心,新疆维吾尔自治区 乌鲁木齐 830002;2.新疆大学 信息科学与工程学院,新疆维吾尔自治区 乌鲁木齐 830046)

摘要:主要是为了给维吾尔语、哈萨克语及柯尔克孜语在自然语言处理、语音识别、语音合成、机器翻译、信息检索、维吾尔语智能信息监控以及维吾尔语舆情分析等研究领域提供语料作为目的。在软件的设计和实现过程中参考维吾尔语、哈萨克语和柯尔克孜语的语法规则以及语言特征,同时引入此三种语言的国际编码,除此根据该网页的特征来分析网页的结构进行判断文本而研发了从网上抓取维哈柯多语种纯文本的数据采集器。最后实现了为少数民族自然语言处理研究搭建语料库准备大规模语料。

关键词:多语种;自然语言处理;.NET;数据抓取;语言特征;语料库

中图分类号:TP311 文献标志码:A 文章编号:1009-3044(2015)11-0023-03
DOI:10.14004/j.cnki.ckt.2015.0593

Design and Implementation of Uyghur Kazak Kirghiz Multi-lingual Online Data Capturing System Based on .NET

Zaytuna Ababakri[1], HOU Cun-yi[1], Miradeljan Mamat[2], ZHANG Li-xin[1]
(1.Modern Education Technology Center, Urumqi Vocational University, Urumqi 830002, China; 2.Colleges of Information Science and Engineering, Xinjiang University, Urumqi 830046, China)

Abstract: Mainly as a purpose of in order to provide data for Uyghur Kazak Kirghiz languages in some research fields such as NLP, Speech recognition, Speech synthesis, Machine translation, Information retrieval, Uyghur Intelligent Monitoring as well as the Uyghur Public Opinion Analysis. In the process of design and implementation of software, referred to the syntax rules of Uyghur Kazak Kirghiz languages. Introducing these three languages International coding, In addition to according to current webpage's features to analyze structure of webpage and judging the text to development data collector Uyghur Kazak Kirghiz multilingual pure text from web. Finally achieved for minority NLP research to build corpora prepared a large corpus.

Key words: multilingual; NLP; .NET; data capture; language feature; Corpus

随着互联网的广泛普及和计算机技术的不断发展,社会也逐渐进入到了一个由物联网发展主导的智能化、数字化阶段。如何用一种自然、便捷的方式与计算机进行交流是目前信息化时代当中的迫切需求之一。近年来少数民族信息处理技术的快速发展,对新疆的政治、经济、文化、教育领域的发展发挥了一定的推动作用,但还是存在着一些问题和较大的差距。新疆是以维吾尔、哈萨克、柯尔克孜族为主的多民族聚集的少数民族自治区,截至2007年底,全疆总人口为2095.19万人,其中少数民族占60.68%。在新疆1271.78万少数民族人口中,维吾尔族965万人,哈萨克族148万人,柯尔克孜族17万人,三个民族总人口占全疆人口总数的一半以上[1]。尽管在维吾尔、哈萨克、柯尔克孜(以下简称"维哈柯")自然语言处理中搭建语料库、词干库、词缀库等作为前期的基础层工作,但在整个研究过程中处于核心的地位并且是必不可少的部分。然而为维哈柯词法分析、语法分析、句法分析、文本分析、词干提取、词性标注、维吾尔语机器翻译以及语音处理(维吾尔语语音识别和语音合成)准备语料是比较耗时间,成本高的工作。因此为了节省时间,降低成本,本文开发并设计了基于维哈柯的多语种网上文本采集处理器的开发迫在眉睫。下面给出自然语言处理技术范围如图1所示。

1 关键技术

维哈柯文都是以阿拉伯文为基础演化而来的拼音文字,语言文字相近,虽然手写方式很相似,但在一些细节上存在很大的区别。维吾尔语包含32个字母,其中8个元音,24个辅音。

收稿日期:2015-02-26
基金项目:乌鲁木齐职业大学校级课题"数字化校园资源整合与应用的研究"(No.2014XY005);"网络安全综合管理平台研究与开发"(No.2014XY007)
作者简介:再吐娜木·阿巴白克力(1981—),女,新疆伊犁人,硕士研究生,研究方向为计算机应用与网络技术;侯存义(1973—),男,山东人,副教授,研究方向为计算机网络;米尔阿迪力江·麦麦提(1989—),男,新疆喀什人,硕士研究生,CCF会员(NO.E200032951G),研究方向为嵌入式智能移动开发,语音处理,自然语言处理;张立新(1975—),男,四川人,高级实验师,研究方向为计算机网络及软件.

哈萨克语包含33个字母，其中9个元音，24个辅音。而柯尔克孜语是包含30个字母和一个合体字母。本系统主要是对页面的语法进行分析从而消除网页噪声以及维哈柯文的编码标准两种问题[2]。

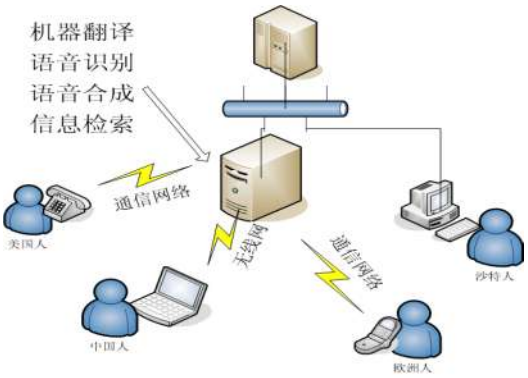

图1　自然语言技术应用范围

1.1 去除网页噪声

目前大部分网站所使用的开发语言与其所对应的脚本都不相同。不管是维哈柯文网站，还是英文、汉文网站都有相同的网页布局特征，消除网页中的各种噪声是系统需要解决的关键问题之一。若用ASP.net来开发本网站的脚本是c#，而由VB.net开发的网站脚本语言是VB或者是VBScript。由于每个脚本语言都有各种脚本标记[3]，因此为了实现抓取当前网页的纯文本信息，首先必须消除那些PHP, JavaScript, HTML, CSS等标记语言中的多余的标记符号。所以本文所开发的系统主要是根据以上各个脚本文件的特征，包括常见特征标记以及根据标记文法结构来消除网页中的噪声[4]。

1.2 判断维哈柯文

将各种脚本的标记符号消除而得到纯文本后，判断得到的文本是否是维哈柯文子是一个关键的问题[5-6]。为了解决此问题本文调用了维哈柯文国家标准编码范围，也就是说根据每种语言文字所具有的国家统一的标准Unicode编码来消除非维哈柯文字。因为个别维哈柯文网站首页最下面有一些汉文网站会作为友情链接而被列出来，所以仅通过消除网页中的噪声而直接获取文本时，也会同时抓取汉文、英文或者是数字等内容。因此系统还要对所采集出来的文本再进行非维哈柯文信息的过滤，从而只保留文本中的维哈柯文字信息[7]。

2 系统设计与实现

2.1 系统设计过程

本文系统的主要设计流程如下，系统工作流程如图2所示。

首先判断数据域（数据采集结果显示区域）是否有网址输入（如图3所示），如果没有网址输入则会弹出提示“输入网址”等信息，若已输入网址，系统会对所输入网址的格式进行判断，这一功能主要是根据正则表达式来对网址格式的正确与否进行检查。若格式有误则提示“输入正确网址”直到输入网址的格式正确为止。此后，若输入网址正确，系统会根据网站特征和当前网页中所显示的文字（维哈克文）特征来搜集当前网页上的文本内容如图4所示。此处所说的网页特征是指当前网站的开发语言的特征，也就是网页的脚本标记特征；而语言特征并不是指语法规则或者是词法、文法、句法上的规则，而是在国际标准Unicode编码上的特征。最后系统将采集出来的结果按*.txt或者是*.doc格式导出保存。

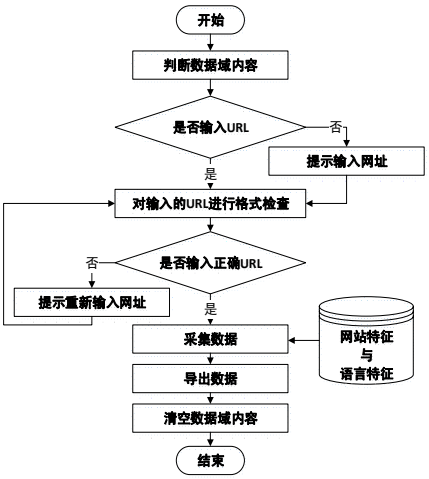

图2　多语种数据采集系统流程图

图3　判断输入正确网址

图4　数据采集

2.2 实现过程

因为所采集到页面的语法分析是基于HTML(Hypertext Markup Language)协议的，所以在消除网页噪声以获取网页正文文本内容之前，必须事先了解HTML的语法结构。HTML标准就是利用SGML定义了一些标记，主要用于描述文本的显示方式[8]。HTML的语法中主要包括四部分内容：实体(Entity)，元素(Element)，属性(Attribute)，以及注释(Comment)。由于网页是基于HTML的超文本文档，因此它包括纯文本和标记等。纯文本是不包含在标记中的字符串，它通过标记的定义体现出不同的字体、字型、颜色等因素，同时标记定义了网页的显示属性[9]。本文的系统根据少数民族语言（维哈柯语）的特征，并通过HTML有些常见特征标记（如表1所示）以及根据标记文法构造对网上的多语种数据进行采集。

表1　HTML 文件的特点

序号	特点
1	所有 HTML 语句都是<>…</>结构，<>表示开始标记，</>表示结束标记。
2	有的语句只有开始标记<>，没有结束标记</>，如<P>。
3	所有语句的标记名称不分大小写。
4	注释语句是<!-- … -->结构。
5	转义字符的形式为“&#nnnn”或“&xxx”。
6	所有语句都可以循环嵌套

在维吾尔语语音识别，语音合成以及维汉双向统计机器翻译中所使用的生语料和平行语料都是由此系统而得到的。系统对比较热门的维吾尔文门户网站、访问量较多的论坛以及累计各种文学资料的网站列表（如表2所示）进行数据采集。此过程当中有些网站的各个网页上的文字在后台中或者是在前台评论方式进行输入时，各种不同的网站有可能用各种不规范的输入法来输入文本的话，通过该软件来采集当前网页上的数据时就将整个文本采集下来，而不会判断是否是标准的字体。

采用上述方法，当采用本系统来抓取维哈柯文本信息时，消除网页噪声以及根据编码范围从相当大的容器，也就是整个纯文本中选择维哈柯文，而不会把中文或者是英文或者是数字等文本信息也采集出来。如图5所示，有些维哈柯门户网站中也会存在多种语言文字混合在一起出现的情况也有，也就某块儿文本，某段文本或者是某行中维文、汉文、英文、数字同时出现，在此情况下就只能根据维哈柯文的特征来抓取维哈柯文字，并且过滤其他种语言文本信息。

表2　实验数据采集参考网站列表

序号	类型	网址	题目数	评论数	成员数
1	论坛	http://bbs.misranim.com/	52227	1445998	127508
2		http://bbs.alkin.cn/	18057	217375	12183
3		http://bbs.alkuyi.com/forum.php	6472	130997	95632
4		http://bbs.anatuprak.com/	18082	130789	46096
5		http://tawpek.com/bbs/	9157	63124	1579
6		http://www.7qiz.com/forum.php	3817	42996	5917
7		http://www.yarp.cn/forum.php	1870	18444	4765
8		http://qirah.com/bbs/	2901	15894	2434
9		http://bbs.eldawa.com/forum.php	2395	8960	2296
10		http://hakaniye.com/forum.php	1111	8752	1355
11		http://baghrax.com/bbs/forum.php	235	6606	2714
12		http://bbs.apandim.com/forum.php	3	5928	2641
13		http://bbs.xjtv2.cn/forum.php	914	3610	3850

图5　过滤其他文字光采集维（哈/柯）纯文本

除了某块儿文本中的汉文、英文、数字、特殊符号以及图片等信息外，纯汉文或者是纯英文网页，系统不抓取任何信息。虽然在此网页中能够抓取纯文本，不过根据编码范围来过滤时系统不返回任何文本信息。这是因为汉文和英文跟维哈柯文的国家标准 Unicode 编码范围不同，因此系统很容易识别到非维哈柯文文本。

3 结束语

本文主要根据网页结构以及对页面进行语法分析来消除当前页面的网页噪声，然后对采集好的纯文本进行筛选操作，过滤非维哈柯文内容而获取维哈柯纯文本为维吾尔语的语音识别，语音合成以及维汉双向统计机器翻译准备语料库时提供所需要的生语料。但是此系统只能采集当前网页的文本，若将系统进一步优化，使其能够采集网页子链接下的文本信息，那么会更加节省工作量以及时间。除此之外，如果系统对藏文和蒙文等其他少数民族语言的文本信息也能采集的话，将会给更多的少数民族研究者、学者提供一个较好的平台。

参考文献：

[1] 陈英. 维哈柯语言文字软件开发及产业化专项介绍[J]. 信息技术与标准化, 2011(6): 4-6.
[2] 纪希禹. 数据挖掘技术的应用实例[M]. 北京: 机械工业出版社,2009.
[3] 明日科技. C# 技术大全 [M]. 北京: 人民邮电出版社, 2011: 650-652.
[4] 谢丹夏. WEB上的数据挖掘技术和工具设计[J]. 计算机工程与应用, 2001(6): 85-87.
[5] 吴俊森. 维哈柯多语种搜索引擎倒排索引模块的实现[D]. 乌鲁木齐: 新疆大学, 2007.
[6] 吐尔洪·吾司曼,维尼拉·木沙江. 维哈柯多语种搜索引擎中索引器的研究[J]. 新疆大学学报: 自然科学版, 20112(28): 132-135.
[7] 吐尔地·托合提, 维尼拉·木沙江, 艾斯卡尔·艾木都拉. 维哈柯多文中全文搜索引擎的设计与实现[J]. 计算机应用与软件, 2009, 6(26): 96-98.
[8] 于静, 李森. 基于WEB信息抽取的主动服务技术研究[J]. 计算机系统应用, 2008(1): 54-60.
[9] 袁园, 王永平. WEB数据挖掘技术综述[J]. 科技信息, 2007(27): 65-67.