第34卷 第24期 计 算 机 工 程 2008年12月
Vol.34 No.24 Computer Engineering December 2008
·软件技术与数据库· 文章编号：1000—3428(2008)24—0058—03 文献标识码：A 中图分类号：TP393

基于DOM修剪的藏文Web信息提取

珠 杰，欧 珠，格桑多吉
(西藏大学计算机科学与技术系，拉萨 850000)

摘 要：随着互联网的普及和藏文信息技术的不断发展，出现了大量的藏文网站。该文根据藏文“音节点”的特征识别藏文网页并进行抓取。在建立DOM树的基础上，分析网页的链接、非链接文本与主题信息块之间的相关度。通过语义修剪算法提取藏文主题信息。经测试证实，该算法在藏文网页识别和藏文主题信息提取中具有较好的适应性。

关键词：音节点；DOM树；藏文；Web信息提取

Tibetan Web Information Extraction Based on DOM Pruning

Zhu Jie, Ngodrup, GeSang Dorje
(Department of Computer Science and Technology, Tibetan University, Lhasa 850000)

【Abstract】With the widespread use of Internet and the development of Tibetan information technology, there are a lot of Websites of Tibetan information resource. This paper identifies Tibetan Web page and crawls it according to features of Tibetan syllable dot. Based on DOM, it analyzes relevance between linked and non-linked Web page text with topical information via pruning semantics algorithm to extract Tibetan topical information. Test result shows that the algorithm to identify and extract in the Tibetan Websites topical information has good adaptation.

【Key words】syllable dot; DOM tree; Tibetan; Web information extraction

随着信息技术的发展，藏文网站在国内得到了迅速发展。目前，国内的藏文网页有以下特点：大多以新闻、西藏文化和历史、风土人情、藏文论坛等为内容；网站规模小，数量不大，用户访问数量少；藏文网站中有汉英藏多种文字；编码多种多样。这些特点给藏文网页信息提取带来一定的难度。

1 藏文网页特征

1.1 音节点的统计分析

藏文是一种拼音文字，有30个辅音字母，4个元音字母和5个反体字。如果考虑梵音转写藏文，包括的字符更多。藏文音节是以一个或多个字母通过横向和纵向的形式组合而成，其中必须包含一个基字和一个元音，在不考虑梵音转写藏文时，藏文的一个音节不能超过7个字母[1]。

藏文的每个音节是通过一个点来进行划分的，称为音节点。根据藏文的构词方法，每少于7个藏文字符之后会出现一个音节点。如果考虑音节点出现最少的情况，即可以假设藏文句子中均是由7个藏文字符构成一个音节，那么每7个藏文字符之后会出现一个音节点，这样可以判断在8个藏文字符中，音节点出现的概率 12.5%，所以一篇藏文文章中音节点出现的概率不会少于为 12.5%。如果按照平均计算，即1~7 个字符各构成一个藏文音节，含音节点的字符数分别为2, 3, 4, 5, 6, 7和8，这样音节点(7个)平均出现的概率是20%。

在藏文字频统计文章中，不少专家对音节点出现频率做过研究。根据对《中华大藏经•丹珠尔》统计结果表明：以构件计算，音节点出现频率为24.23%；以字丁(预组合)计算，音节点出现频率为30.66%[2]。

总之，如果按照字丁(预组合)形式统计，音节点出现的频率更高，它是每少于5个字符之后会出现一个音节点；按构件(ASCII编码方案或藏文编码国际标准方案)统计，每少于7个字符之后会出现一个音节点，出现的概率在 20%左右。藏文词频统计的结果表明，音节点出现的频度是最高的。

1.2 语义分析

网页中的链接用来表示其他页面或站点的连接，在网页中链接除了文本的内容外，还有图片链接等内容。可以发现网页中链接的内容一般不是信息提取的主题内容。通过链接的语义属性和非链接的本文的内容可以表示某个节点的主题信息的相关程度。例如，在某个节点链接文本的比重占得越多，与主题信息的提取相关度越低。

结合非链接节点的文本，网页中通过藏文本身的特征来分析语义的特点。根据藏文中音节点出现的频率，通过音节点的数量和文本的长度可以区分主题信息所在的位置和相关程度。

1.3 藏文的编码特征

不同的藏文网站采用了不同的藏文编码，在几十种藏文编码中，国内的藏文网站大都采用基于 GB2312 和基于 Unicode 标准的藏文编码，国外的藏文网站大都采用基于ASCII的藏文编码。由于藏文 Microsotf Himalaya Software 的发布，有些网站开始采用 ISO/IEC 国际标准的藏文编码。

目前藏文网页编码的复杂局面，不利于藏文网页信息的提取，也无法通过 html 网页中的“encoding”和“charset”来识别网页，并且需要特殊的转换程序把藏文转换成统一的编码。

根据音节点出现频度很高的特点，分析音节点编码特征。在藏文不同的编码中，音节点的编码如表1~表3所示[3]。

基金项目：国家自然科学基金资助项目(60763010/F0206)
作者简介：珠 杰(1973－)，男，讲师、硕士，主研方向：数据挖掘；欧 珠，教授；格桑多吉，讲师
收稿日期：2008-08-04 **E-mail：**roky_tibet@eyou.com

表 1　基于 ASCII 的藏文编码

编码名称	码点范围	音节点编码
Ltibetan	0x21~0xFE	0x2D
TCRC	0x21~0xFE	0x2D, 0x2E
Old Sambhota	0x21~0xFE	0x2D
New Sanbhota	0x21~0x7E	0x2D
Tibetan Machine	0x21~0xFE	0xCD
TMW	0x21~0x7E	0x2D
Tibword	0x21~0xFE	0x2D
TibKey	0x21~0xFE	0x2D
Tsamkey	0x21~0x7E	0x2E
SUZTIB	0x21~0xFE	0x2D
UCHAN	0x21~0x7E	0x2D

表 2　基于 GB2312 的藏文编码

编码名称	首尾字节范围	音节点编码
方正 DOS	0xC0~0xEE	0x21~0xFE, 0x2D
方正 WIN	0x21~0xFE	0x2D, 0x2E
华光 DOS	0x21~0xFE	0x2D
华光 WIN	0x21~0x7E	0x2D
同元编码	0x21~0xFE	0xCD
西藏大学编码	0x21~0x7E	0x2D

表 3　基于 Unicode 的藏文字符集

编码名称	首尾字节范围	音节点编码
Unicod	标准	U+0F00~U+0FCF U+0F0B
扩充集 A	U+F300~U+F8FF	U+0F0B
直贡藏文	U+E000~U+F3A6	U+E0DF

2　藏文网页信息提取算法

本文根据中文、英文网页信息提取算法，考虑藏文本身的特点，提取藏文网页的内容。

2.1　藏文网页信息提取系统

藏文网页的信息提取系统采用了 Heritrix+Lucene 的方法。首先，经过对 Heritrix 的改造，识别出藏文网页，并抓取到镜像目录中；其次，对抓取到的藏文网页采用 Jtidy 转换器进行 HTML 到 XML 转换，并将相关的文本通过编码转换器转换成统一的编码格式；最后，对产生的 XML 文档利用 XML 解析器建立 DOM 树，通过语义分析器分析，使用剪枝器剪枝与主题无关的节点，并提取藏文网页内容。藏文网页的信息提取系统如图 1 所示。

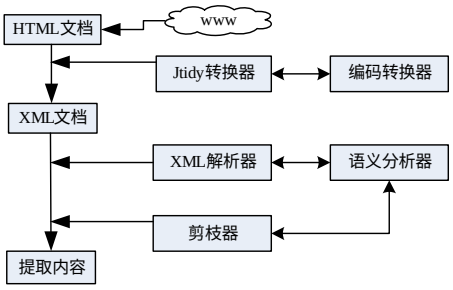

图 1　藏文网页信息提取系统

2.2　藏文网页的识别算法

在藏文网页中，国内藏文网站的字符集设定大都采用 charset=gb2312，国外的藏文网站或国内的有些藏文网站字符集设定采用 charset=UTF-8，所以用 charset 或 encoding 来识别藏文网页是不可能的。

在藏文网页中，字体采用或字体样式采用{font-family: BZDBT;}等特征，但是目前藏文编码的复杂性和网页中动态技术的采用，使得该特征无法保证藏文网页识别的准确性。

本文通过藏文本身的特点和字体的相应名称来判断藏文网页。具体算法如下：

(1)查看 font face, font-family 和 charset 的值，如果存在小字符集字体名称(ASCII 和国际标准)转到(2)，否则转到(3)。

(2)计算网页中音节点的个数，并计算音节点之间的字符的距离。判断距离是否在 1~7 之间，如果具有这种特征的数量达到一定的阈值，判断为藏文网页，否则不是藏文网页，并转到(6)。

(3)如果具有大字符集字体名称和基于 GB2312 编码标识，转到(4)，否则转到(5)。

(4)计算网页中音节点的个数，并计算音节点之间字符的距离，判断距离是否在 1~5 之间，具有这种特征的数量如果达到一定的阈值，判断为藏文网页，否则不是藏文网页，并转到(6)。

(5)计算网页中音节点的个数，并计算音节点之间的字符的距离，判断距离是否在 1~7 之间，如果具有这种特征的数量达到一定的阈值，判断为藏文网页，否则不是藏文网页，并转到(6)。

(6)如果是藏文网页，保存到镜像目录中，否则放弃存储。

根据测试，藏文网页识别中阈值的设置结合音节点的数量与满足藏文识别算法的音节数量来确定，如果网页中藏文音节的数量达到音节点数量的 1/3 以上，认为该网页为藏文网页。

2.3　DOM 树的建立

Document Object Model(DOM)，即文档对象模型，是 W3C 制定的标准接口规范。将 HTML 或 XML 文档被解析后，转化为 DOM 树，DOM 树的每个节点是一个对象。本文中 Jtidy 把 HTML 文档转换成 XML 文档，并解析成 DOM 树。按照 DOM 树提供的接口，遍历整个 DOM 树，利用语义修剪算法，剪枝与藏文主题无关的节点，提取藏文文本内容。DOM 树的结构如图 2 所示。

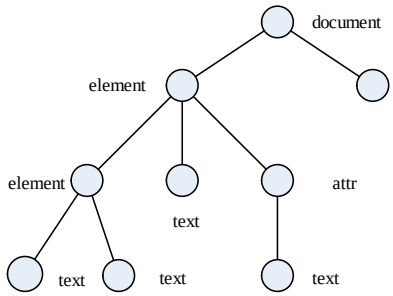

图 2　XML 文档的 DOM 树结构

2.4　藏文网页语义修剪算法

根据Jtidy创建完DOM树后，剪枝器开始递归地遍历DOM树[4]，调用语义分析器，即语义修剪算法。如果达到一定的阈值，保留该节点；否则认为是无关节点，删除该节点。

通过如下的语义算法设定阈值：

(1)网页的语义属性是由某节点(某个节点可能包含了多个子节点)链接上的文本数量*LinkContentLength*和该节点上的文本数量*ContentLength*作为语义属性进行分析。由于网页

链接上的文本既不能表示主体信息块[5]，又不一定都包含了文本内容，因此通过某节点链接上的文本数量和某节点上文本数量进行比较，可以确定主题信息所在的相关程度，可以由下式来表示：

$$Relativity(n_i)=\frac{LinkContentLength(n_i)}{ContentLength(n_i)} \quad (1)$$

当某节点n_i的*Relativity*值越接近于 1 时，需要提取的主题信息的相关程度越低；该值越接近于 0 时，主题信息的相关程度越高。根据测试，当*Relativity*阈值小于等于 0.5 时，认为与主题信息的相关度较好 ;如果阈值小于 0.8，保留该节点；如果阈值大于 0.8，认为是与主题信息无关的，删除该节点。除文本较少的信息块外，基本能够提取主题信息。

(2)用某节点上非链接藏文文本数量 NodeContentLength 以及与藏文音节点的数量 *SyllableDotCount* 比来确定主题信息所在的位置和信息提取的准确性，由下式来表示：

$$Veracity(n_i)=\frac{SyllableDotCount(n_i)}{NodeContentLength(n_i)} \quad (2)$$

根据 1.1 节中的讨论，某个节点n_i的*Veracity*的值至少应该大于 0.125。另外为了避免均是音节点的情况，可以限制节点n_i的*Veracity*值的上限。从音节点统计的情况上限不会超过 0.4，但是在网页中为了处理断字问题和书写美观，往往在行末多加几个音节点来保证藏文的正常断行。根据测试，*Veracity*的上限调宽为 0.6。这样能够保证主题信息提取的准确性。

(3)在算法使用时，对某个节点n_i先进行相关度的判断，再进行准确度的分析。

在 DOM 树中，某个节点下面会有很多个子节点(子树)，这些子节点和该节点之间的整体相关度如何，可以通过下式来进行解释：

$$Relativity=\sum_i Relativity(n_i) \quad (3)$$

某个节点的相关度是该节点下子节点相关度的并集。如果该节点下，子节点有很多个链接，链接的文本数量也很多，但是其中某个子节点中还是包含了主题信息，这时整体相关度表现为很低。显然，这时不能保证信息提取的准确性。

为了能够提取到该节点下某个子节点中的主题信息，在相关度判断的基础上，利用整体准确性，即下式来保证信息的提取的准确性：

$$Veracity=\sum_i Veracity(n_i) \quad (4)$$

3 实验

3.1 测试环境

为正确提取藏文网页的信息，测试环境使用了 Heritrix+Lucene 的方法。通过对 Heritrix 抓取的网页进行 Jtidy 的转换、XML 的 DOM 解析，利用语义修剪算法，经过遍历完成藏文信息的提取。

3.2 测试结果

测试网页的数据来源于 http://zw.tibet.cn, http://www.tibettl.com/, http://www.hl88.com 等网站，图 3 是其中的一个实例，可以看到该网页既包含了多个文本的链接，也包含了主题信息块，具有较好的代表性。初步的实验表明，通过音节点提取藏文信息具有较好的适应性。在提取完成的文本文件中，由于藏文编码的不统一，显示成乱码形式，经过手工处理能够得到如图 4 所示的实例。

图 3　藏文信息提取前的网页实例

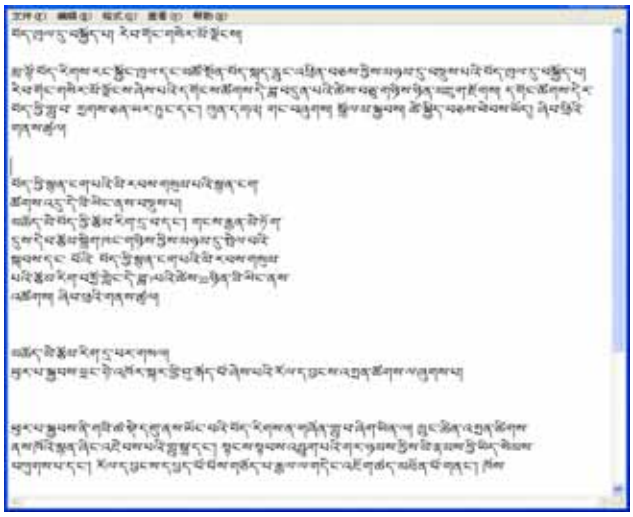

图 4　藏文信息提取后的内容实例

4 结束语

在藏文网页信息提取中，利用音节点的语法特征来识别藏文网页并提取藏文网页中的主体信息，具有较好的适应性。在藏文网页信息提取研究过程中，发现还有诸如藏文分词算法、文档自动分类、编码转换、词库的建设、数据库中藏文的支持等多种基础性的工作，有待于进一步的研究和完善。

参考文献

[1] 土弥三菩扎. 藏文文法四种合编[M]. 北京: 民族出版社, 2005.

[2] 扎西次仁. 《中华大藏经•丹珠尔》藏文对勘本字频统计分析[J]. 中国藏学, 1997, (2): 122-133.

[3] 刘汇丹, 芮建武, 吴　建. 藏文网页的编码识别与转换[M]//民族语言文字信息技术研究. 北京: 西苑出版社, 2007.

[4] 张惠颍, 曲著伟. 基于子树匹配的交互式 Web 信息抽取方法[J]. 计算机工程, 2006, 32(9): 78-80.

[5] 王　琦. 基于 DOM 的网页主题信息自动提取[J]. 计算机研究与发展, 2004, 41(10): 1786-1792.