The Key Technologies of Collecting Tibetan Websites

http://dx.doi.org/10.3991/ijet.v8i3.2680

Wang Zhijuan
Minzu University of China, Beijing, China
National Language Resource Monitoring & Research Center, Beijing, China

***Abstract*—Tibetan websites are important for studying the Tibetan language and Tibetan culture. The domain names and character encodings of Tibetan websites are complex. Three key technologies of collecting Tibetan websites are discussed in this paper. Firstly, using high frequency syllables in different Tibetan character encoding to searcher the web pages may be in Tibetan as many as possible. Then, using Tibetan syllable dot and the properties of HTML to judge the language of web pages is in Tibetan or not. Thirdly, according the relationships between the URL of Tibetan web page and the URL of Tibetan website, the URLs of Tibetan websites are gotten. About 100 Tibetan websites in three Tibetan character encodings (Unicode, BZD and Tongyuan) are collected by these methods. These key technologies are also useful for collecting websites in other minority language.**

***Index Terms*—Tibetan websites, Language recognition, Tibetan character encoding**

I. INTRODUCTION

Internet is one of the most important mass medias in our life with its fast spreading rate and easy to access. Tibetan is an old language. It is used more than 1400 years. Along with developing of information construction of Tibetan, there are more and more Tibetan websites and web pages [1-3]. Theses websites and web pages make it possible for Tibetan all over the world to read information in Tibetan any time and anywhere. But how many Tibetan websites are there? What are the features of Tibetan websites that are different from Chinese or English websites? Which information is popular in Tibetan websites? And so on. In order to answer these questions, we must know first how much Tibetan websites are there!

There is little information on how much Tibetan websites are there and how to collect Tibetan websites. So the key technologies of collecting Tibetan websites should be studied first. And these technologies will be discussed in this paper. The main contents of this paper are:

- Firstly, two popular methods of collecting websites are introduced.
- Then, the features of Tibetan websites are introduced. And the definition of Tibetan websites is given.
- Thirdly, three key technologies of collecting Tibetan websites are introduced.
- Finally, the conclusion is given.

II. THE COMMON METHODS TO SEARCH WEBSITE

There are two commons methods to collect websites: get the websites from the management institutions of domain name, obtain the website information from navigation website.

A. Getting the websites from the management institutions of domain name

Every website has a unique domain name. The domain name should be registered from management institutions of domain name. In mainland, China Internet Network Information Center (CNNIC) is responsible for Chinese domain name registry. From CNNIC, all websites that are registered in CNNIC can be gotten.

The domains of Tibetan websites are complicated. The domains of some Tibetan websites are second-level. And these websites may be gotten from CNNIC. But there are some Tibetan websites that their domains are third-level (such as http://tb.tibet.cn/) or subdirectories of second-level (such as http://www.qhrch.com/zw/). These Tibetan websites can not be found in CNNIC. And this method needs the support of CNNIC and we cannot get websites information directly. So we cannot use this method to collect all Tibetan websites.

B. Obtaining the websites from navigation website

There are some websites that can provide a lot of links of other websites. Such as http://www.hao123.com/ (it provides many links of Chinese websites).

As shown in Figure 1, there are several navigation websites about Tibetan websites. The forth (http://www.xizang123.com/tibetan.html) provides 43 links of Tibetan websites (some websites have same second domain name and some websites can not be opened!). The last (http://www.884343.com/) provides a lot websites about Tibetan. But the number of Tibetan websites is less than 50.

Because the number of websites included in navigation website depends on manual work. The number of websites provided by navigation website is limited and cannot guarantee including all websites in some language. So we cannot use this method to collect Tibetan websites.

III. THE FEATURES OF TIBETAN WEBSITES

The first Tibetan-website appeared in 1999. After that, more and more Tibetan websites appeared. Previous studies found that Tibetan websites have following features.

Figure 1. The navigation websites that provides the links of Tibetan websites

A. *The domain names of Tibetan websites are complex*

Generally, a website should have a separate second-level domain. But for Tibetan websites, their domain names are more complex. There are three conditions.

1. **The URL of Tibetan website is second level domain (2LD).** For example, http://www.sorig.net/ is a Tibetan website and its URL is 2LD.
2. **The URL of Tibetan website is third level domain (3LD).** For example, Chinese Tibetan Information Network has four language-versions: Chinese, Tibetan, English and Traditional Chinese. Its Chinese version is 2LD (http://www.tibetinfor.com/). Its Tibetan version is 3LD (http://tb.tibet.cn/).
3. **The URL of Tibetan website is a subdirectory of 2LD.** For example, the domain name of Qinghai Red Cross Hospital website is http://www.qhrch.com/zw/. The domain name of its Tibetan version is http://www.qhrch.com/zw/. It is a subdirectory of 2LD.

In order to collect Tibetan websites as more as possible, we must concern this feature of Tibetan websites.

B. *The Tibetan character encoding of Tibetan websites are complex[4-7]*

As shown in table I, Microsoft Himalaya is based on Unicode, BZD, Tongyuan and so on that are used in mainland are based on GB2312, Sambhota and so on that are often used outside of mainland are based on ASCII. The diversification of Tibetan character encoding leads the diversification of Tibetan character encoding of Tibetan websites.

TABLE I.
TYPE SIZES FOR CAMERA-READY PAPERS

The types of character encodings	**The name of character encoding**
The Tibetan character encodings based on Unicode	Microsoft Himalaya
The Tibetan character encodings based on GB2312	BZD Tongyuan Fangzheng Huaguang
The Tibetan character encodings based on ASCII	LTibetan TCRC Old Sambhota Tibetan Machine(TM) New Sambhota TibetanMachine Web(TMW)

From above all, we can see that Tibetan websites are not similar to Chinese or English websites. A definition of Tibetan websites should be given. If a URL has more than two subdirectory and all of web pages of these two subdirectories are in Tibetan. It is the URL of Tibetan website.

In view of the above futures of Tibetan websites, the definition of Tibetan websites is following:

1. It has two and two subdirectories that webpage language is Tibetan.
2. It should include all Tibetan websites in all character encodings of Tibetan.

C. *The method to collect Tibetan websites*

According to the definition of Tibetan websites, a method to collect Tibetan websites is designed as shown in Figure 2.

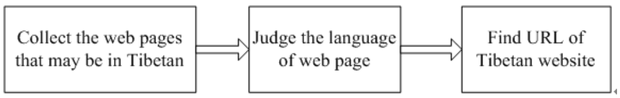

Figure 2. The structure frame of collecting Tibetan websites

Firstly, collect web pages that may be in Tibetan. Secondly, judge the language of web page is in Tibetan or not. Finally, use the URL of Tibetan web page to find the URL of Tibetan website. According to the method of collecting Tibetan website, the key technologies of collecting Tibetan websites will be discussed.

IV. KEY TECHNOLOGY 1: COLLECT WEB PAGES THAT MAY BE IN TIBETAN

Although collecting web pages is introduced in many papers [8-13], there is no introduction on how to collect Tibetan web pages.

In order to collect the web pages that may be in Tibetan, the Tibetan characteristic should be discussed firstly. As shown in Figure 3, a Tibetan word is consists of one syllable or several syllables. A Tibetan syllable includes root letter, prefix, head letter, vowel, suffix and post suffix. Tibetan syllable is divided with syllable spot.

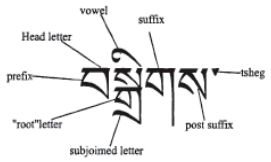

Figure 3. The component of Tibetan word

All Tibetan web pages contain Tibetan high-frequency words or syllables. So we can use Tibetan high frequency words or syllables to collect web pages that may be Tibetan. Table II is the Tibetan high frequency words that are gotten by three experts using different corpus. [14-16]

TABLE II.
TIBETAN HIGH-FREQUENCY WORDS

Number	Zhaxiciren	Word frequency %	Wang weilan	Word frequency %	Jiang Di	Word frequency %
1	ད	4.79	ད	3.88	བཅད་	5.75
2	ཀྱི	2.61	དུ་	2.58	ད	5.40
3	དུ་	2.48	བའི	2.24	ཤ	5.20
4	ད	2.37	ཨ	2.19	མ	3.35
5	ཨ	2.23	ད	1.84	ཞིང་	2.93
6	བའི	2.02	ལ	1.37	ཁ	2.33
7	ལ	1.82	ཀྱི	1.27	གཏོང	2.09
8	དར	1.62	དར	1.24	ད	2.01
9	ཞེས	1.54	མ	1.23	ཚུ	1.39
10	ལ	1.50	ཅ	1.13	ན	1.38

Because of multiple character encodings of Tibetan, it is necessary to collect high frequency word in different character encoding in Search engine. The search results are shown in Figure 4.

Figure (a) is the search results of དང་ in UTF8. Figure (b) is the search results of དང་ in BZD. Figure (c) is the search results of དང་ in Tongyuan. It is clear that, for the same high-frequency word, different character encoding is used, different search result is gotten.

Because the target is to collect Tibetan websites, the web pages have similar URL will be discarded. Finally, web pages with unique URLs are gotten.

As shown in Figure 5, some web pages that only contain Tibetan high frequency words may be gotten. So it is necessary to judge the language of web pages.

V. KEY TECHNOLOGY 2: LANGUAGE RECOGNITION OF WEB PAGES

There are several methods which are used to recognize the language of web pages.

(a).UTF8

(b). BZD

(c). Tongyuan

Figure 4. The search results of དང་ using different character encoding

Figure 5. The web pages contains part Tibetan

A. *Using the property of HTML lang to recognize the language of web pages [17]*

The "lang" property of HTML declares the language used in web pages. It makes search engine and Browser to read the content of web pages correctly. According to W3C standard, every web page should declare language with "lang" property of HTML. For example:

<html lang="en"> ... </html>	<html xmlns="http://www.w3.org/1999/xhtml" lang="en" xml:lang="en"> ... </html>

The abbreviations of different languages are defined in ISO 639. It is shown in Table 3.

TABLE III.
ABBREVIATIONS OF DIFFERENT LANGUAGES USED IN LANG

Language	639-1 standard	639-2 standard
English	en	eng
Chinese	zh	chi/zho
Japanese	ja	jpn
Tibetan	bo	tib/bod

Because many Tibetan web pages did not contain the HTML lang. We can't use this method to recognize the language of web pages.

B. *Using the "Charest" and "Font-family" of HTML to recognize the language of web pages*

The content-type of HTML provides the character encoding of web page. For example,

<meta http-equiv="content-type" content="text/html; charset=gb2312" />

There are several character encodings of Tibetan such as Unicode, BZD, Tongyuan and so on. The "charset" of Tibetan web pages is GB2312 or UTF8. The language of web page can't be judged using "charset ".

Different character encodings mean different character fonts will be used. So the language of web page can be judged by "font family" of HTML. If the "charset" is GB2312, the "font family" may be BZDBT, TIBETBT and so on. If the "charset" is UTF8, the "font family" is Microsoft Himalaya. But some Tibetan web pages did not contain the "font family". Therefore, we can't judge a web pages is in Tibetan or not only using "charset" and "Font-family".

C. *Using some language detection tools to recognize the language of web pages [18]*

There are some language detection tools such as "Google Language Detection", "What Language is This", "Polyglot 3000" and so on. They can work well on English, Chinese, Japanese and so on. But these tools do not work well in identifying Tibetan.

D. *Using the frequency of Tibetan syllable dot to recognize the language of web pages*

The Tibetan syllable dot is used frequently in every Tibetan text. So we can use the frequency of Tibetan syllable dot to recognize the language of web pages.

Here, about 1000 Tibetan web pages are selected to calculate the frequency of Tibetan syllable dot. The frequency of Tibetan syllable dot is about 30%. So we can use it to judge a web page is in Tibetan or not.

E. *Using high frequency Tibetan words in Internet*

In reference [19], three Tibetan websites are selected and about 1,600 web pages are downloaded. The total number of Tibetan words is about 480,000. Table 4 is the high frequency Tibetan words in Internet. The first five Tibetan words have high frequency of Tibetan words and coverage of web pages. So we can use them to judge a web page is in Tibetan or not.

TABLE IV.
HIGH FREQUENCY TIBETAN WORDS IN INTERNET

number	Tibetan syllable	The frequency of Tibetan words	The coverage of web pages
1	ཡིས།	6.4%	97%
2	དང།	4.1%	80%
3	ར།	2.2%	77%
4	ནས།	1.7%	81%
5	ས།	1.3%	72%
6	ཞེད།	1.2%	53%
7	ཚོད།	1.1%	58%
8	ཀྱི།	1.1%	52%
9	ལ།	1.1%	54%
10	རེད།	1.0%	62%

It is important to judge the language of web pages. Method 5.1-5.3 can not judge the web page is in Tibetan or not. The last two methods can work better.

VI. KEY TECHNOLOGY 3: FIND THE URL OF TIBETAN WEBSITES

Figure 6 the flow chart of finding the URL of Tibetan website by the URL of Tibetan web page

Since a lot of Tibetan web pages with different URL are available. It is possible to find the URL of Tibetan websites using the URL of Tibetan web page. The relationships between the URL of Tibetan web page and the URL of Tibetan website are described below.

- The URL of Tibetan web page is 2LD
- The URL of Tibetan web page is the lower URL of Tibetan website
- The URL of Tibetan web page is the top-level of Tibetan website

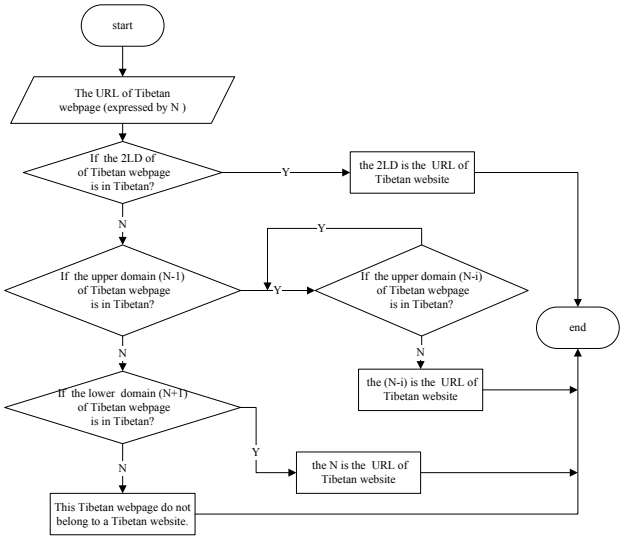

Figure 6. The flow chart of finding the URL of Tibetan website by the URL of Tibetan web page

Figure 6 is flow chart of finding the URL of Tibetan website by the URL of Tibetan web page.

Firstly, get a URL of Tibetan web page (expressed in N). And judge the 2LD of Tibetan web page is in Tibetan or not. If yes, this 2LD is the URL of Tibetan website.

If the 2LD of Tibetan web page is not in Tibetan, judge the upper domain (expressed in N-1) of Tibetan web page is in Tibetan or not. If yes, judge the upper domain (expressed in N-i) is in Tibetan or not until N-i-1 is not in Tibetan. So, the N-i is the URL of Tibetan website.

If the upper domain (expressed in N-1) of Tibetan web pages is not in Tibetan, judge the lower domain (expressed in N+1) of Tibetan web page is in Tibetan or not. If yes, the N is the URL of Tibetan website. If not, this Tibetan webpage do not belong to any Tibetan website.

VII. CONCLUSION

Tibetan websites is very important for sharing Tibetan information and spreading the Tibetan culture. Collecting Tibetan websites as more as possible is helpful to study Tibetan and Tibetan culture. Three key technologies of collecting Tibetan websites are discussed in this paper. Firstly, using high frequency syllables in different Tibetan character encoding to searcher the web pages may be in Tibetan as many as possible. Then, using Tibetan syllable dot and the properties of HTML to judge the language of web pages is in Tibetan or not. Thirdly, according the relationships between the URL of Tibetan web page and the URL of Tibetan website, the URLs of Tibetan websites are gotten. About 100 Tibetan websites in three Tibetan character encodings (Unicode, BZD and Tongyuan) are collected by these methods. These Tibetan websites have been included in 2011 Language Situation in China (publish by The Commercial Press). These key technologies of collecting Tibetan websites are also useful for collecting websites in other minority language.

REFERENCES

[1] Chen yuzhong, Yu shiwen, “Research and Prospect of Tibetan information processing technology”, China Tibetology, vol.4, pp.97-107, 2003.

[2] YU Hong - zhi, LA Bin - jun, HE Xiang – zheng, “Tibetan Language Information Processing Techniquesunder Web Condition”, Journal of Northwest Minorities University for Nationalities(Natural Science), vol.26, No.1, pp.53-57, 2005.

[3] Wang Zhijuan,Qin Qiwen, “The Present Situation of Tibetan Websites”, the researches and advancements of information processing for Chines minorty languages and characters, Ethnic Publishing House, pp. 28~33,2012.

[4] An-jian-cai-rang, “Research of Tibetan Character Internal Codes Recognition Algorithm in the Multi-coded Environment”, Microprocessors, No.5, pp.69-71, 2009.

[5] LI Yong-hong,HE Xiang-zhen,AI Jin-yong,YU Hong-zhi, “Tibetan coding method and mutual conversion”, Journal of Computer Applications, vol.29, No.7, pp.2016-2018, 2009. http://dx.doi.org/10.3724/SP.J.1087.2009.02016

[6] LIU Hui—dan□RUI Jian—wu□WU Jian, “Encoding Detection and Conversion of Tibetan Web Pages”, the 25th Chinese Information Processing Society Conference Proceedings, pp.573-580,2006.

[7] Chen Qi, Li Yongzhong,”The study of Tibetan web crawling and unified coding conversion system”, Journal of Northwest Minorities University for Nationalities(Natural Science), vol.30, No.2, pp.22-26, 2009.

[8] ZHANG Liang,WANG Chun,“Design and implmentation of distributed web-crawling system”, Journal of Beijing Technology and BusinessUniversity(Natural Science Edition), vol.27, No.1, pp.37-41, 2009.

[9] LIU Yu-lian,ZHOU Chun-nan,ZHANG Qing,“Research and implementation of Web information retrieval system dynamic reconfiguration”,Information Technology, vol.28, No.7, pp73-75, 2004.

[10] Zhao Yin, “Design of the Mass WebPages collecting”,Northeastern University,2009.

[11] Liu Yulian, Zhou Chunnan, Zhang Qiang. “Research and implementation of Web information retrieval system's dynamic reconfiguration”, Information Technology, vol.7, No.7, pp.73-75, 2004.

[12] Pei Chunbao, Danzeng Basang, Ou Zhu, “Analysis and settlement that Tibetan handles problem in JAVA programming technology”, Tibetan Science & Technology, Vol.10, pp.68-70, 2008.

[13] Zhu Jie, Ou Zhu,Gesang Duoji. “Tibetan Web Information Extraction Based on DOM Pruning”, Computer Engineering, vol.32, No.24, pp58-60, 2008.

[14] Zha-xi-ci-ren, “Tibetan statistical analysis of Chinese Tripitaka·Danzhuer”, China Tibetology , No.2, pp. 122-133 ,1997

[15] WANG Weilan, “The Frequency-rank of Language Unit in Modern Tibetan”, Science Technology and Engineering, vol. 4 No. 5, pp.413-417, 2004.

[16] LU Ya-jun,MA Shao-ping,ZHANG Min,LUO Guang, “Researches of Calculations of Tibetan Characters, Pieces,Syllables,Vo cabulary and Universal Frequency and Its Applications”, Journal of Northwest Minorities University for Nationalities(Natural Science), vol.24 No.2, pp.32-42, 2003.

[17] Codes for the Representation of Names of Languages, Codes arranged alphabetically by alpha-3/ISO 639-2 Code

[18] http://www.labnol.org/internet/identify-language-of-text/8441

[19] Chao Hui, “The Tibetan high-frequency vocabulary of 2010 Tibetan newspapers and Tibetan Network (News)”, Language Situation in China: 2012 (CD), The Commercial Press, 2012.

AUTHORS

Wang Zhijuan is with the College of Information Engineering, Minzu University of China and Minority Languages Branch, National Language Resource Monitoring & Research Center, Beijing, China (e-mail: wangzj_muc@126.com).

This work was supported by National social science fund (No: 11CY016), National Science and Technology Supporting Item (No: 2009BAH41B04), the State Nationalities Affairs Commission fund (No: 10ZY07) and Open Project of National Language Resource Monitoring & Research Center (No: NMLR201104). Manuscript received 05 January 2013. Published as resubmitted by the author 02 June 2013.