Collection of Tibetan Network
Chang-zhi WANG, Guixian XU* and Hui WANG
Information Engineering College, Minzu University of China, Beijing, 100081, China
*Corresponding author

Keywords: Web crawler, Pretreatment, Tibetan corpus.

Abstract. With the development of Tibetan information technology, technologies about Tibetan web crawlers was extremely important. We elaborate different pages pretreatment rules according to the different sites and make the collected Tibetan Web text dump for Tibetan documents, by constructing a Web crawler to crawl different Tibetan websites, Experiments show that it can quickly and effectively to build large-scale Tibetan corpus, build the foundations for Tibetan information processing technology by self-made software and the module of pretreatment.

Introduction

With the development of the Internet, the search engine helps people retrieve information, becomes the main entrance to the Internet among the people. As the core function of the search engine, web crawler is responsible for downloading the information from the Internet to the local. As the development of information, the number of web pages is increasing, the requirement of web page information search is getting higher, so the functional requirements of web crawler must be improved. This paper has important practical significance to explore the related technology of network collection.

At present, both Chinese and English web crawler [1] researches techniques are quite mature, but there are few related crawler researches based on Chinese minority language. In recent years, Tibetan and other minority language pages are more and more. The obtained Tibetan information resources on the Internet automatically using web crawler, provides corpus foundation for Tibetan text classification, machine translation, data mining and so on.

Background

The method of traditional Tibetan collection, can be traced back to 1960s. Tibetan literature acquisition program called PL-480 undertaken by the Library of Congress, in the first artificial mode for acquisition [2].After the seventies, many universities in the United States and Centre for the study of religion did collected the Tibetan documents, and stored it on microfilm, also open these data for the whole world. It is obviously not feasible that the traditional way to collect information in today's information era.

Web crawler began to appear in 1990s, the current technology has matured, is an important part of the search engine. Because of the Tibetan search engine commercial value is not high, this research is relatively less, only a handful of universities and scientific research institutions do this research. The related work of various studies on Tibetan web page collection, mainly related to the identification and conversion of Tibetan web page processing code [3], Tibetan segmentation scheme [4] and other technology. Currently the Tibetan text collection on Internet usually use custom web crawler. Crawler using a given initial URL began to download page, and extract the effective links from the pages, through the links between web pages, using the appropriate access strategy, can access the Internet all the Tibetan information page in theory.

Web pages are very blended, extracting effective information from the web page, the need to deal with the noise of the page. We adopt JSOUP parser and Dom4j parser to preprocessing the obtained Tibetan web page.

This paper will introduce the Crawler4j which is an open source crawler to collect texts from
Tibetan related sites and to preprocess obtained pages for structured documents, save for the Tibetan raw corpus, as the corpus of Tibetan information processing technology research.

The Proposed Method

Principles of Crawler

Web can be seen as a directed graph of net structure [5], each page as a node, the URL of the page as the unique identifier in a whole net, and other URLs in the page can be as the directed edge, point to the other web pages node. Web crawler complete the access to each of the web node through these edges with direction. On the crawling strategy of web crawler, it is obvious that the traversal strategy of the directed graph, includes depth first and breadth first algorithm. Depth first algorithm is a recursive process, the crawler will consume too much system resources in the actual process, may lead to the trapped problem of crawler. Here crawler adopts the Breadth-First strategy, and the idea is that the initial URL is highly relevant with the theme of web page in a certain range.

Download Tibetan Page

About Tibetan web page download, we set the links of popular Tibetan portal sites as initial URL, the crawler according to the robots protocol to determine which page is accessible, which is prohibit access. The web page can be downloaded to the local through the download module of web page .URL extractor pick up the urls that not visited by system, and push these urls into the URL queue that ready to be accessed, while the system also need to maintain a visited URL queue to avoid repeat visit of the Tibetan page. As more and more crawling web pages, the maintained URL queue has been to become larger, the system will become more and more slow. In order to solve this problem, the hash value of URL can be calculated by using the hash function [6] such as MD5, which can be used as an identifier of the visited URL, so that it can reduce the storage space. It can improve system performance that we can use the bloom filter [7] algorithm to solve the problem of whether the URL have accessed. In this system, the open source crawler uses Berkeley DB embedded database to store the URL, so as the memory overflow problem not appear , and the program can take the last unfinished tasks continue when again to work , solve the problem of restart work from the initial URL every time.

The system also uses multi thread technology, so that the crawler has a more powerful ability to grab, set the time of connecting the web and reading to avoid unlimited waiting.

The entire crawler frame is shown in figure 1:

![Figure 1. Frame of Tibetan crawler.](image-url)
Preprocessing of Tibetan Page

The set of acquired original web pages may contain a large amount of information that is not related to the content of the text, such as the HTML markup language of web pages. Removing web noise is very important for the work of the system.

In the technology of web page denoising, the text content in the web page is usually divided by the HTML tag, which makes the system cannot directly deal with the text directly. There are a lot of advertising, images, audios and other useless information in the page, these interference information is called noise. In order to remove them, there are mainly based on web structure [8] method, template based [9] method, based on visual information [10] method. At present, the better effect of WEB subject extraction method is based on web page structured information, according to the types build DOM tree, through the DOM tree to extract the main part of the web page.

We using JSOUP tool parsed Tibetan web page into a DOM tree, and according to the different characteristics of Tibetan website, making different pretreatment rules, through a selector, finally collected from the HTML document to the title, author, time and other the corresponding text related information.

Tibetan Text Persistence

The related Tibetan pages will be downloaded to the local through the web crawler, then after denoising converted into structured XML documents. We use an open source XML parser package called Dom4j, which is used to get related information from after the documents preprocessed, saved directly to the local disk.

According to the different sites it collect Tibetan information and classification preservation, lay the foundation for the Tibetan text classification study.

Experiment

Crawling on several popular Tibetan websites. For example: China Tibet Net (http://www.tibet.cn/), Tibet Xinhua channel (http://tibet.news.cn/), people's net Tibet channel (http://tibet.people.com.cn/) and others. Remove duplicate and does not conform to the standard web pages, we included a total of the 82763 pages, some pages in the preprocessing stage will filter out due to part of the web pages are picture, audio or other, so the page were collected 65465 Tibetan documents through the preprocessing. A web page directly downloaded to the local, as shown in figure 2:

![Figure 2. Original web page style.](image)

It can be known from the figure 2, the original is a lot of useless information in the web page. The text stored as XML documents through the web pretreatment technology. Document save
format as shown in figure 3:

```html
<xml version='1.0' encoding='UTF-8' ?>
  <file>
    <domainname>http://xizang.news.cn</domainname>
    <chineseName>Tibet Xinhuas channel</chineseName>
    <pinyin>xinzhoupd</pinyin>
    <language>Tibetan</language>
    <encodingtype>UTF-8</encodingtype>
    <provider>http://xizang.news.cn</provider>
    <corpusType>Network</corpusType>
    <filename>xinwenzhongxin.2014-09-05.c_1336225552.html</filename>
    <title>བྲུག་སྲི་བྱུང་བོའི་བོད་ིབ་ཁམས་ངོ་བོ</title>
    <subtitle/></subtitle>
    <author/>....</author>
    <time/publish>2014-09-05</time/publish>
  </file>
</xml>
```

Figure 3. Format of corpus text saving after preprocess.

Experiments show that it adopts the multi thread technology can quickly and efficiently obtain all page from Tibetan website through the web crawler. Through the preprocessing technology, we can collect the text related information to the structured document.

**Conclusion**

In this paper, we collected Tibetan information base on the name, source, date of release, content and other labels, stored structured document through the self-made crawler software and pretreatment system. This research is the foundation of Tibetan information processing, provides sufficient data for the study of extraction hot words of Tibetan and Tibetan text classification. The system crawling the Tibetan website mainly uses the Unicode encoding, not relating to the relevant page code detection and code conversion. Subsequent research will improve the coding problem.

**Acknowledgement**

This research was financially supported by "Beijing Social Science Fund" (NO.14WYB040), “the National Key Technology Research and Development Program of the Ministry of Science and Technology of China” (No. 2014BAK10B03), “the National Natural Science Foundation of China” (No. 61309012, No. 61331013).

**References**


