Research on Tibetan News Sites' Web Crawler and Search Engine

Han Zhiqiang
School of Information Engineering
Minzu University of China,
Beijing 100081, China
hanzq@sohu.com

Sun Wei
School of Information Engineering
Minzu University of China,
Beijing 100081, China
645281904@qq.com

Xu Guixian*
School of Information Engineering
Minzu University of China,
Beijing 100081, China
xuguixian2000@sohu.com
* Corresponding Author

Abstract—*In this paper, researchers detailedly introduce the features of Tibetan language and related technologies that researchers use to deal with Tibetan news web pages with computers. To get the content of the Tibetan news, researchers used web crawler to download Tibetan news pages which were the bases of this project. Researchers used an open source web crawler named scrapy and rewrote the crawl part to make the crawler work more accurately and efficiently. To search the Tibetan content in a way, researchers defined and counted every statistical data that was useful and helpful to enhance the performance of the search engine. Researchers used solr, another open source software, as the user interface of this system. The crawler and search engine were combined by the web pages to provide the data retrieval service. Comparing with other works, researchers' work adopted a safe and stable enough framework to enhance the user experience in using Tibetan search engine. Researchers' work played a positive role in the spread of Tibetan culture and promoted the development of the Tibetan language news in the field of search engines.*

Keywords-Tibetan; news sites; web crawler; solr; search engine.

I. INTRODUCTION

With the development of the internet, Tibetan web page data is getting larger and larger. At the same time, the amount of Tibetan news data no matter domestic and livelihood, sporting events or business news is increasing rapidly[1].

As we had the need to classify and organize the Tibetan news, web crawler and search engine came into being. So far both Chinese and English search engines have growing mature. While the development of Tibetan language search engine is limited due to some reasons[2].

After the Tibetan informatization construction, Tibetan web pages are getting much more than before. The approaches to obtain Tibetan information are getting larger and larger, meantime it's harder for us to search for the exact information we need. Tibetan search engine, a way to retrieve Tibetan information, is still unavailable till now. Like the Chinese language, the web crawler based on Tibetan language is the foundation of the Tibetan information processing technology. It starts working from a series of urls to crawl and collect web pages. Finally the Tibetan corpus is built after denoising. The corpus is the original data of our research. Although Chinese Academy of Sciences and Northwest University for Nationalities have already done some research, they haven't achieve a breakthrough. The literature describing Tibetan crawler is not enough and a little bit brief. Tibetan web pages automatic discovery and collection system is not announced yet. Since the Tibetan character set of international standards release too late and to achieve Tibetan display based on the small Tibetan coded character set international standards is relatively complex, this situation lasts for years. As a result, almost all the previous Tibetan software use a set of Tibetan encoding they defined alone[3]. No transfer rules is defined between these software. This is one of the difficulties in Tibetan applications. Simultaneously, as for search engine, to collect the newest and changed Tibetan web pages is its main work as well. Thus the crawler system has the ability to crawl refreshed pages and achieve incremental collection.

We need to judge the encoding of a web page before we want to get its content. First we need to know if the web page is in Tibetan. If the answer is yes, then we need to remove the unrelated contents and then save them in the international unified standards encoding.

II. STRUCTURE CHARACTERISTICS OF TIBETAN LETTERS

A. *Tibetan current situation*[4]

There are so many Tibetan people living in China and most of them resident in the Tibetan areas, such as Qinghai Province, Tibet Autonomous Region, Sichuan Province, Yunnan Province and Gansu Province etc. They speak Tibetan all their lives, which makes Tibetan language native and general in the Tibetan parts of our country. Since the new country was founded in 1949, there have been more and more Tibetan people who could enjoy the new happy life in China. Tibetan language evolved from the Tibetan branch of Sino-Tibetan of Tibeto-Burman languages. Now Amdo, Tibetan, and Kang are three literal dialects that are frequently used in the mainland. Every Tibetan character has the same meaning with a word in Devanagari. Meanwhile you can find the fact that the pinyin we use when we are young is quite similar to Tibetan. Alphabetic writings are their common features. Punctuation marks, vowels and consonants constitute the Tibetan language. The number of consonants is 30, and the signs and pronunciations are listed as follows in Fig. 1 [5].

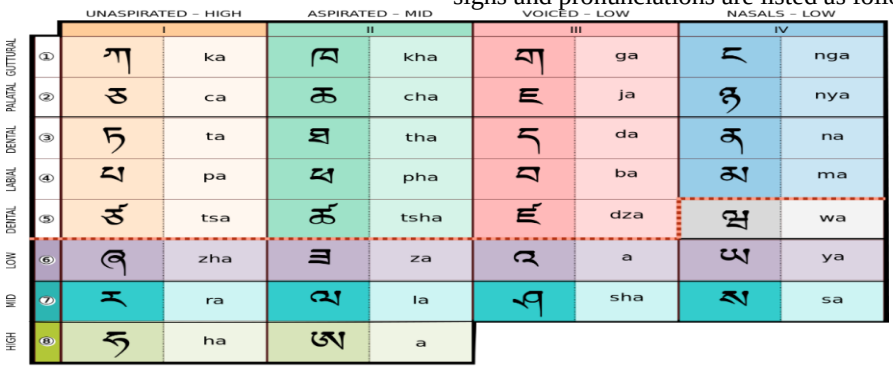


Figure 1.

The signs and pronunciations of vowels are listed as follows in Table I [6].

TABLE I. Vowel symbols

Dependent vowel signs	Symbol Name	Pronunciation	Examples
◌	-	[a] ([ɛ] -d, n, l, s)	ཀ (ka),ཏ (ta), ཨ (a)
◌ི	གི་གུ (gi gu)	[i]	ཀི (ki), ཏི (ti), ཨི (i)
◌ུ	ཞབས་ཀྱུ (zhabs kyu)	[u] ([y] -d, n, l, s)	ཀུ (ku), ཏུ (tu), ཨུ (u)
◌ེ	འཁྲེང་པོ ('kreng po)	[e]	ཀེ (ke), ཏེ (te), ཨེ (e)
◌ོ	ན་རོ (na ro)	[o] ([ø] -d, n, l, s)	ཀོ (ko), ཏོ (to), ཨོ (o)

Complicated as it is, the punctuation marks of Tibetan consist of many elements, such as syllable-dividing marks of syllable point, dividing line, ection sign, phrases, the double hanging operators at the end of the chapters and the single hanging operator at the end of the sentences. These years because of the good and encouraging policy, such as the reform and opening-up policy, Tibetan, Chinese and even English language are becoming similar to each other.

B. *Structure of Tibetan words*[7]

We cannot consider Tibetan the same way as English or Chinese. As a result, we should treat Tibetan as a special case. Base word can be found in every syllable in Tibetan, and it determines the center consonants of the syllable. There is a vowel affix above or beneath the base word, and every style means a new vowel. In Tibetan, it's very common to see the compound words everywhere which play import part in expressing meanings in sentences. The written order is laterally written from left to right, which is consistent with the modern mainstream words.

Polysyllabic word and single-syllable word are two branches of Tibetan words and expressions. Words often appear together in Tibetan and you can hardly tell them from each other, but punctuation marks can be found between syllables. Like Chinese, Tibetan words have many classifications, such as adjectives, verbs, pronouns, nouns, quantifiers, adverbs, numerals, auxiliary, modal, conjunctions, interjection, prepositions and onomatopoeia. Tibetan words can be compounded and derived, thus lots of new words are made. Suffix, less prefix and infix are three important parts of derived words. In the traditional Tibetan grammar, there are 9 typical suffixes and few prefix, infix and postfix. There are six kinds of compounded words, i.e. "nouns and noun", "a noun and a verb", "verbs and verb", "adjective and the adjective", "nouns and adjectives" and "adjective and verb". Grid is used to connect sentences in Tibetan, and there are eight kinds of grids, i.e. nominative, industry grid, as the grid, for the grid, from the grid, genitive, in Georgia and vocative.

C. *Tibetan word segmentation*

There are several scholars who have been doing the related researches in the Tibetan Word Segmentation work, but owing to the fact that Tibetan grammar is different from the widely spread languages, e.g. Chines and English, the problem of Tibetan Word Segmentation remains to be done.

Both Tibetan and Chinese are similar with each other, i.e. there are not typical separate marks that can be easily handled by computers just like spaces in the English between words. Tibetan has a strong grid grammar theory, as a result we cannot directly use the segmentation theories and technologies of Chinese or English for reference. Early word segmentation methods can be roughly divided into two broad categories, i.e. statistical approaches and rule-based methods. Later on a new approach is added into the segmentation methods in Tibetan, i.e. methods based on combination of rules and statistics. Statistical approach needs to build a model first to count for the automatic segmentation system, then we can get the parameters of the model. After that we will chose the very lexical bundle from all the possible lexical bundles based on the experiments' statistics as the final output result, and the result must appears with highest probability. While the rule-based method, using an algorithm, take advantage of rules and word list to match the candidate words in the text with the words in the word list. The candidate words will be segmented to the output as the final result if they successfully match and conform to the requirements of the rules. The third resolution combine the advantages of the previous two resolutions, and avoid the defects of them, as a result it can perform quite well both in property and accuracy rate. The new method will achieve the correct segmentation after the previous work, and identify the people names, place names and organization names by adopt different rules. If the method works well, then the speed of the segmentation will be greatly improved.

III. DISTRIBUTION AND FEATURES OF TIBETAN SITES

A. *Current situation of Tibetan sites*

With the development of the information globalization, more and more Tibetan sites appear, and the contents they present are getting extremely large with various social network relationships buried in them. Relative to Chinese or English sites, the scale of the Tibetan network resources are still small. The number of Tibetan sites is about 180 in which there are still some websites to which we cannot easily have access. "Research on Internet Development in China Minorities" from the National Social Science Fund Project shows us that with the increase of the number of the files in the websites, the cost to search on the web will first drop a little bit, then it will increase rapidly.

B. *Classifications of Tibetan sites*

The report also gives us the current Tibetan sites' distribution[8] as Fig. 2 shows:

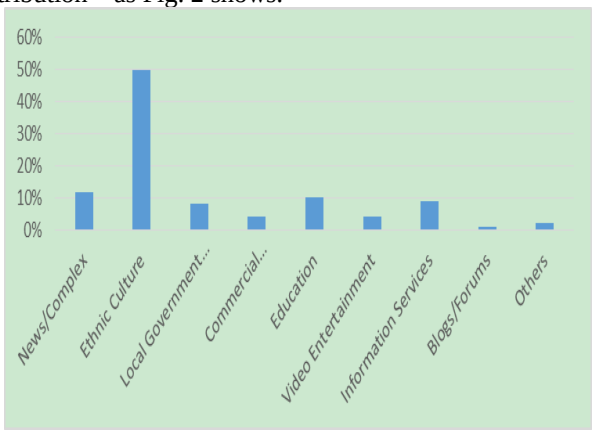


Figure 2.

C. *Languages cases of Tibetan sites*

The languages used in Tibetan websites has been in the statistical analyze in the report as Fig. 3 shows.

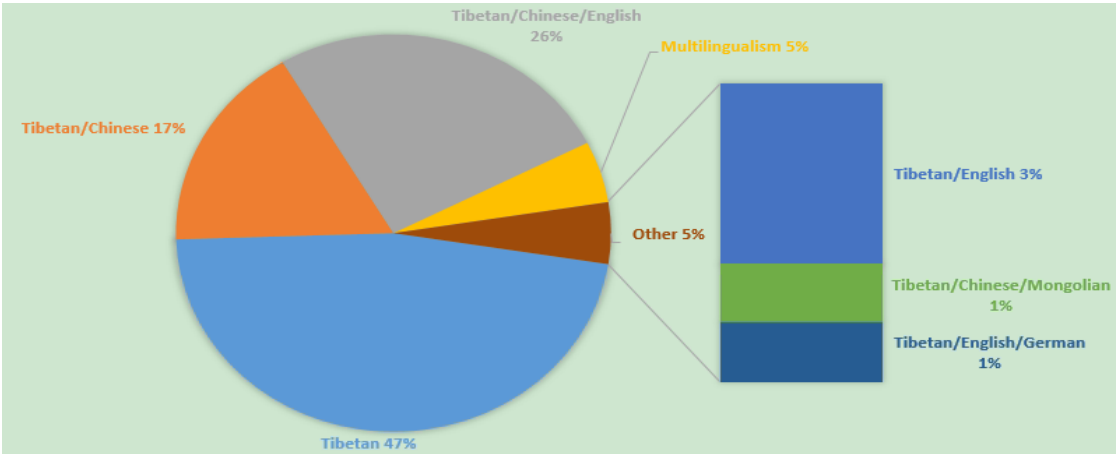


Figure 3.

IV. WEB CRAWLER IN TIBETAN LANGUAGE

A. *Principles of website crawler in Tibetan language*

Web crawler has many names, such as Crawler, bots, robots and wanderer. We may define a web crawler from two angles, broad and narrow[9]. From the narrow perspective, it takes standard http protocol to traverse the Web information space and rely on hyperlinks and Web document retrieval approaches to complete the work. From the broad perspective, it is a web document retrieval program relying on http protocol. Web crawler shows highly ability in web pages extraction, especially its ability to download web pages which is indispensable in search

engine. The realization to visit a site relies mainly on making requests to HTML documents. It traverses a web site and moves from one to another, completes the automatic indexing and then save the index to databases. It works by analyzing the HTML language tags' structures without user intervention.

B. Algorithms to crawl pages in Tibetan language

1. Breadth-first search algorithm[10]

Breadth-first search algorithm, also called BFS, is the simplest one in the graph search algorithms, such as Dijktra algorithm and Prim algorithm. The algorithm complete the graph traversal by crawling along the width of the tree. Once it finds the target, it stops immediately. The use of this algorithm is to enlarge the coverage of the web pages to the most. The main shortage is that it maybe download many unrelated web pages, which reduce the efficiency of the algorithm.

2. Depth-first search[11]

Depth-first search algorithm, also called DFS, keeps digging till the remotest node. For the newest node, if it starts with the source code and the edge between it and the very previous node is new as well, further exploration may be completed along this edge. If there's still node that is not discovered, then the process will be repeated until all the other nodes are found. At the most of the time, trapped problem may occur during the process. As a result, it does not have the completeness and optimization.

3. My strategy

What we use is scrapy that is written in python and is open source as well. We rewrite the program to adapt to crawl the Tibetan web pages. First we start scrapy by adding some start urls, then it will find all the news pages in this list page and next list page. Second it will find all the news pages in next pages and the next page of next page. Meanwhile it will start crawling pages that it found just now, because it works asynchronously. To get all the critical information, it finds the target by using xpath tool. Sometimes we need to construct an xml http request to get the content we need such as "more details". Finally we store all the items into database for further use.

C. Unified encoding of Tibetan sites[12]

Because of the fact that Tibetan has so many encodings, the data process work cannot be done without processing unified news encoding. To make the search part go smoothly, we need to transfer all the news into unified encoding, such as "Unicode". We adopt international standards for Tibetan character encoding.

V. SEARCH ENGINE TECHNOLOGY

A. History and current situation of search engine[13]

World-Wide-Web did not show up until 1990s, once search tools like Archie and Gopher are used to query the files distributed in the hosts. With the development of the Internet technologies the first generation of search engine was born, and Yahoo is the most typical one. After we step into 21st century, to search on the Internet before we go out or do something is becoming a special life style. At present Google and Baidu are two primary search engines we use in daily life.

B. Species and differences between search engineering

According to the way they work, search engines are divided into three main kinds, i.e. full text search engine, vertical search engine and meta search engine.

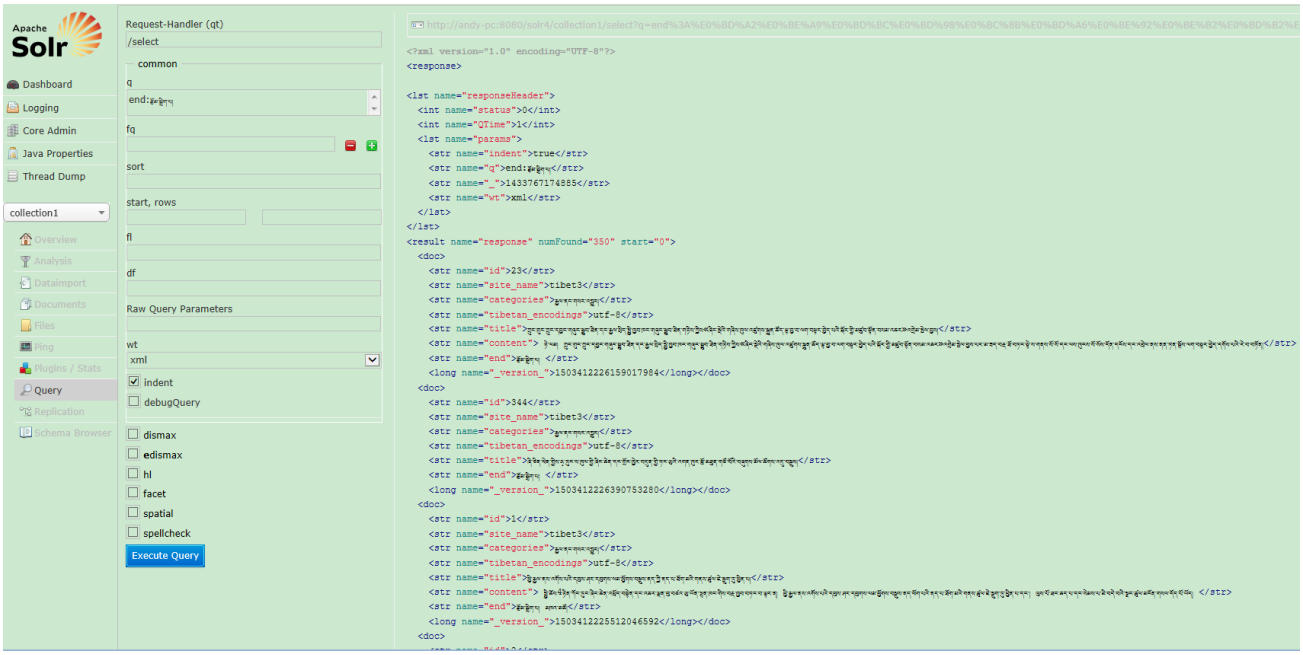
C. Advantages of vertically search engine

Owing to the unique way to collect information, vertical search engine effectively improve the quality of the source information. In addition, because of the fact that the subject areas it is designed for is relatively narrow, it's conducive to efficiently and accurately organize information at the earlier stage. As a result, the quality of the target information is improved as well.

Because of the highly target and specialization of the vertical search engine, the high pertinence and high reliability are ensured. As a result, the target information is located accurately and rapidly.

D. An example of the system

After months of working, we have already finished the crawl part and made it work together with the search engine smoothly and efficiently. The original interface is as follows, later on we will make a web interface and at that time we can visit the Tibetan news pages from browser.



VI. CONCLUSIONS

After the research on the Tibetan letters, distribution and features of Tibetan sites, web crawler in Tibetan language and search engine technology, we find that it's possible and quite meaningful to build a Tibetan language oriented crawler and search engine. With a high efficiency crawler and search engine system, Tibetan culture can be spread faster and wider.

ACKNOWLEDGMENT

This work is supported by "Beijing Social Science Foundation (No.14WYB040)".

REFERENCES

[1] Guixian Xu; Dunhao Zhong; Xu Gao; Yuan Lin; Xiaobing Zhao; Guosheng Yang, "Tibetan Web Information Collection System," Intelligent Networks and Intelligent Systems (ICINIS), 2012 Fifth International Conference on , vol., no., pp.236,238, 1-3 Nov. 2012

[2] Xiang Chuncheng; Weng Yu, "A Template-Based Tibetan Web Text Information Extraction Method," Intelligent Networks and Intelligent Systems (ICINIS), 2011 4th International Conference on , vol., no., pp.218,221, 1-3 Nov. 2011

[3] Gao Hongmei.The Design and Realization of Tibetan Web Pages' Crawler[J].China Computer&Communication,2012,09:36-37.

[4] Chen Yuzhong.The Situation of Tibetan Information Processing Technology[J]. China Tibetology,2003,04:97-107.

[5-6] Information on http://en.wikipedia.org/wiki/Tibetan_alphabet

[7] Liu Huidan.Encoding Deteetion and Conversion of Tibetan Web Pages [J]. Chinese Information Technology,2015,01:170-177.

[8] Xiaodong Yan; Yuan Sun; Xiaobing Zhao; Guosheng Yang, "A Tibetan web Text Clustering model," Information Science and Engineering (ICISE), 2010 2nd International Conference on , vol., no., pp.3388,3391, 4-6 Dec. 2010

[9] Duan Bingying.Study and Design of Web Crawler in Search Engine[D].Xi'an University of Electronic Science and Technology,2014.

[10] He Bai. Research and Implementation of Distributed and Multi-topic Web Crawler System. [J]. Computer Engineering,2009,19:13-16+19.

[11] Shu Meiliu.Search Strategy and Achieve of the Topic Search Engine Spider[J]. Computer Systems & Applications, 2010,03:49-52.

[12] Wu Qiang; Xiao Wei; Bian Ba Wangdui; Pu Dun, "Implementation of Tibetan Search Engine Based on XML Documents," Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on , vol.2, no., pp.17,19, 23-25 March 2012

[13] Shih-Fu Chang; Chen, W.; Meng, H.J.; Sundaram, H.; Di Zhong, "A fully automated content-based video search engine supporting spatiotemporal queries," Circuits and Systems for Video Technology, IEEE Transactions on , vol.8, no.5, pp.602,615, Sep 1998