The Design of Model for Tibetan Language Search System

Wang Zhong
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
wangzhong@lzu.edu.cn

Abstract-In this paper, the prototype of the Tibetan language search system is built and the solutions to key issues for this model are proposed.The characteristics of Tibetan language of web pages are analyzed and extracted. The web page encoding are converted to standard Unicode, which permits for better recognition of Tibetan words for web page and the efficiency for searching informations in Tibetan will be significantly improved. The emergence probability, as well as semantic features, are considered for the Tibetan words classification system, The capability of eliminating unknown words and ambiguity problem are enhanced. This design will increase the search efficiency and help users get better searching results.

Keywords:Tibetan; word segmentation; indexing; URL database; encoding conversion

I INTRODUCTION

Tibetan is is used as the basic language in Tibetan areas, mainly including Tibet, Qinghai, Sichuan, Gansu, Yunnan and some other countries such as India, Nepal, Bhutan, Sikkim. Tibetan is a kind of alphabetic writing including consonants, vowels and punctuations. Fig.1 shows the grammatical structure of Tibetan and Fig.2 takes one Tibetan letter 'LIAO' as an example.

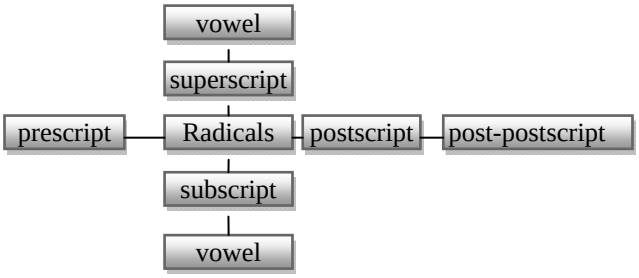

Figure1. grammatical structure of Tibetan

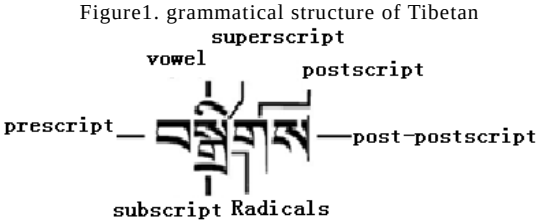

Figure2. letter LIAO's structure

With the rapid development of information and network technology, the dissemination of Tibetan information becomes more and more widespread. There are many research groups which focus on expanding research on the Tibetan information processing. Yu etc. proposed segmentation method based on particle and its grid connection features [1]. Lu etc. put forward the intelligent keyboard layout for Tibetan with "a key to multi-character" and "a key to place"[2]. Yong etc. proposed the conversion method of Tibetan from non-standard to the national standard (GB18030) [3]. Rui etc. used OpenType fonts to solve the overlay characters in Tibetan [4].

TABLE I SUPPORT FOR TIBETAN

Search System	Support Tibetan	Tibetan API	word segmentation	Sorting
Google	YES	YES	NO	NO
BaiDu	NO	YES	NO	NO
Bing	NO	YES	NO	NO
Yahoo	YES	YES	NO	NO
HotBot	YES	NO	NO	NO
Lycos	YES	NO	NO	NO
Excite	YES	NO	NO	NO

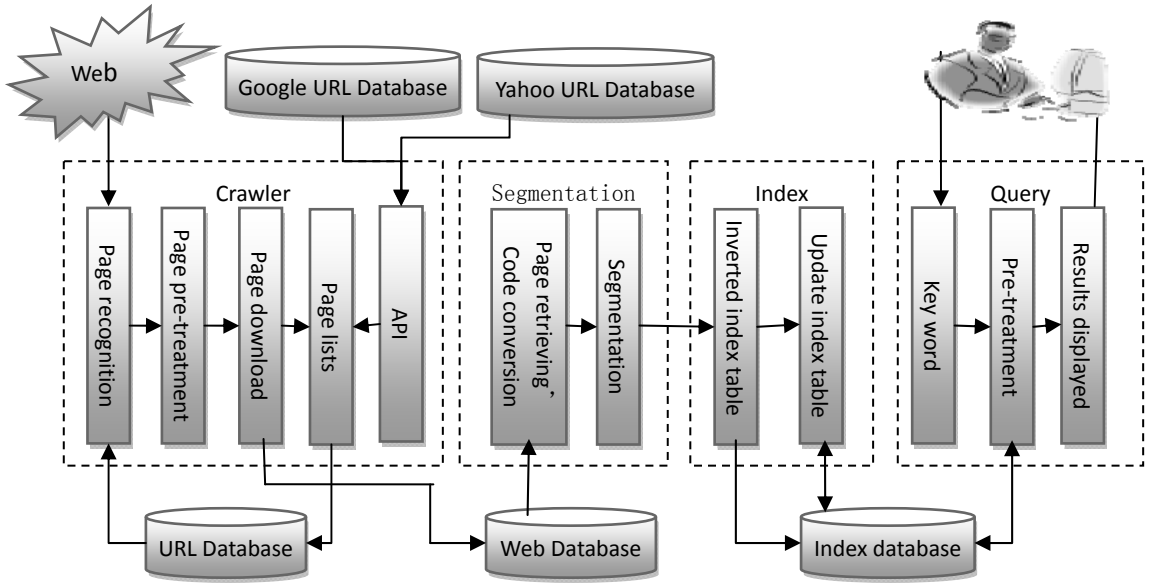

Figure3. System Block Diagram of TibetanSearch

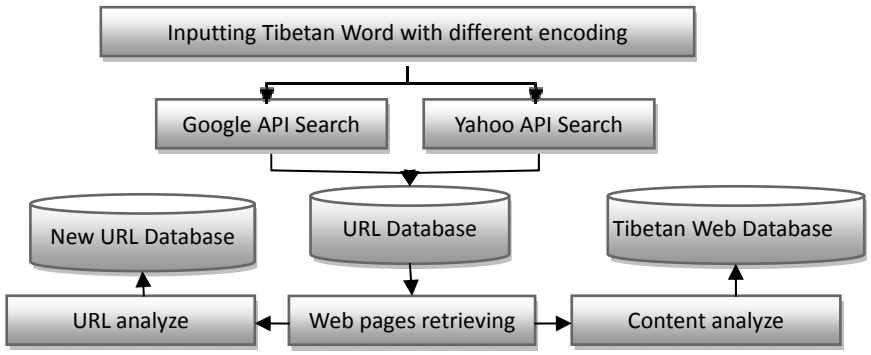

Figure4. Module for collecting the Tibetan Pages

The Search engine system based on a certain strategy, use specific computer programs to gather information from the Internet, and then the information is organized and processed to provide users the search results. Most of search engines still can't support Tibetan. Some systems such as Google, Yahoo are able to perform simple queries for Tibetan, but there is no word segmentation and sorting according to Tibetan keyword. Table I summarizes the support of popular search systems for Tibetan.

This paper designs a model for Tibetan Search. The crawling program is used to collect Tibetan information, complete accurate word segmentation of Tibetan, and establish index structure. This specific design will help people get results quickly and provide sorting search results according to a certain sorting algorithm.

II TIBETAN SEARCH MODEL

Tibetan search system 'TibetanSearch' is composed of crawlers system which identifies and crawls web pages of Tibetan, segmentation system, indexing system and query system which provides users the search results. The system's main structure is given in Fig. 3. We can use the API of Google or other search system to gather the

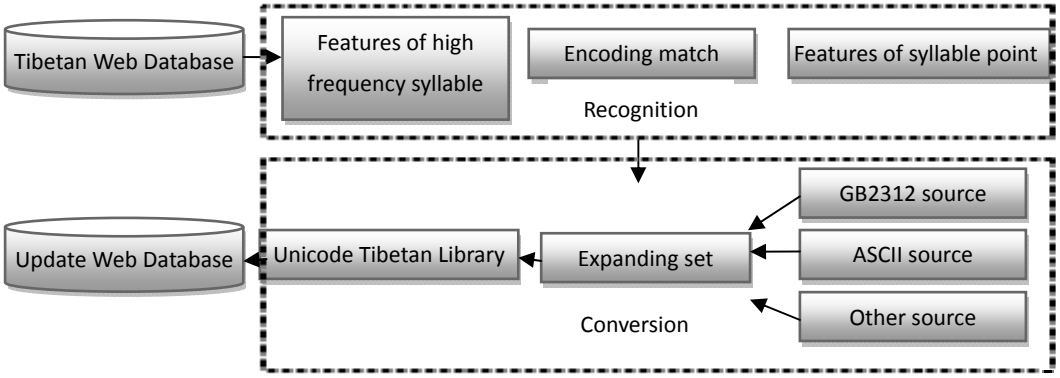

Figure5. recognition and conversion

Tibetan pages; meanwhile, we collect Tibetan information from Internet with Crawlers [5]. If there is no Tibetan information, a threshold for search depth should be set and this search should be discarded when the search depth exceeds the threshold. If the target page contains information in Tibetan, word segmentation and index need to be done for Tibetan in the next step. TibetanSearch make word segmentation for users' input keywords, then get relevant web pages by indexed table, sort those pages, and send the results to the users.

III MODULE DESIGN OF SYSTEM

The Design of Tibetan-search system which can recognize and retrieve the Tibetan pages from Internet, unify the encoding, build a Tibetan word segmentation system and index system based on inverted indexed tables in order to meet users requirements for Tibetan search.

A Module for collecting the Tibetan Pages

Those Tibetan pages will be collected by combining Google, Yahoo API and crawler system. URL Database is built by collecting different encoding of Tibetan word (Duplicate pages in the database will be discarded by re-hash merge algorithm), and the corresponding pages are downloaded for further word segmentation processing (Fig.4). At the same time, the crawler starts to scan from the URL of entrance page[6], and constantly sends the pages to Tibetan Web database using depth-first or breadth-first algorithm until it satisfies stop condition.

B Module of encoding recognition and conversion

In this module, the crawler system needs to identify whether the collected pages includes Tibetan information. Both the recognition process of Tibetan web page and the determination process of Tibetan page encoding need the Tibetan features to make a judgment since it's not easy to get the correct encoding information for Tibetan from the HTML files in attribution set encoding and charset [4, 7, 8].

The final purpose of Tibetan encoding conversion is to unify all non-standard encoding to Unicode standard using various Tibetan encoded control tables. The flowchart is shown in Fig.5.

C Module of word Segmentation

The target is divided to sentences using punctuation, the Tibetan sentence is divided to block by case-auxiliary. In the block, referring to the dictionary database [8,9], word segmentation are completed using forward and reverse maximum matching method. If the phrase has a high probability, add it to the dictionary database. After completing probability analysis, if there is still ambiguity problems, a semantic analysis with the database of semantic rules should be conducted. Finally the output will be used for the indexing module; all block diagrams are shown in fig. 6.

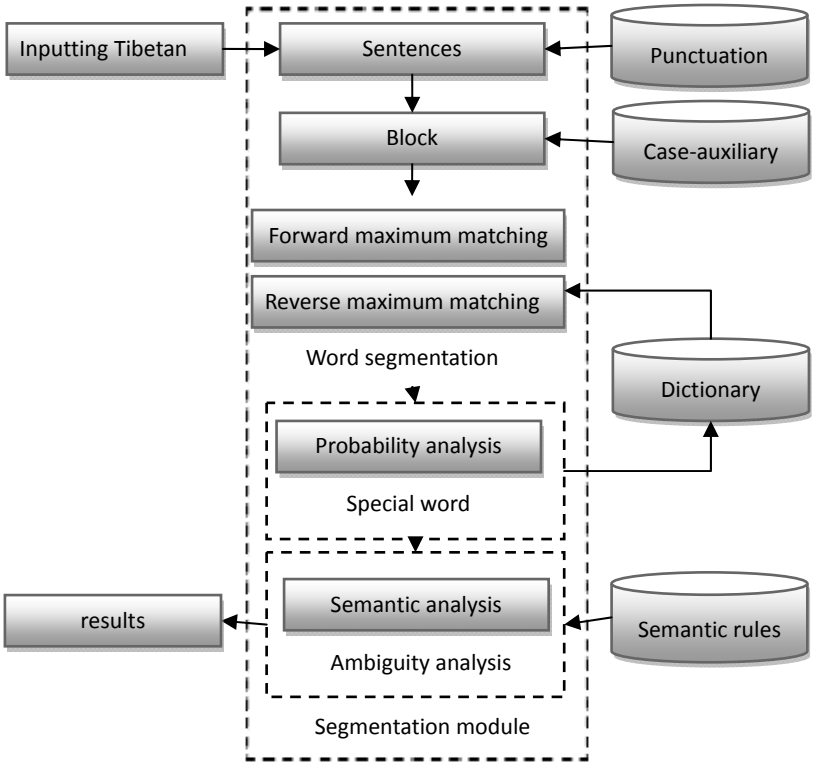

Figure6. Module of word segmentation

D Module of indexing

Because of the coexistence of Tibetan, Chinese and English webpages, there are direct links between these three kinds of pages, so it's better to consider these three types of pages to build the indexing. Blocks are indexed in the order of Tibetan, Chinese, and English. Fig.7 shows the organizational structure of this module.

E Module of output results

The weight of each web page is composed of the weight of key words and the weight of sentences. According to the weight obtained from target pages and the specific requirement of users，the selected information will be pre-treated and displayed(Fig.8).

Based on the evaluation system designed by Cleverdon [10], the special database of Tibetan Web pages will be built from all web pages crawled by Google's API. Proper query words need to be proposed by Tibetan experts. The fair results will be marked.

IV CONCLUSIONS

The model of search system of Tibetan language whic h unifies the recognition of webpage and Tibetan webpag e transcoding is proposed and designed. The Tibetan page s are processed by combining maximum matching, the pro babilistic method and the semantic segmentation. The acc uracy of segmentation will be improved significantly by u sing the probabilistic method for the unknown words and t he semantic analysis for dealing with the ambiguity probl em[11,12]. The sorting will be scientific and effective by using the unique index structure specially designed for Ti betan words. This will be more concise, convenience and friendly for users.

ACKNOWLEDGEMENTS

Acknowledgment This work is partly supported by National Natural Science Foundation of China under Grant No.61405083, the Fundamental Research Funds for the Central Universit ies, Grant No.lzujbky-2013-187, lzujbky-2013-42.

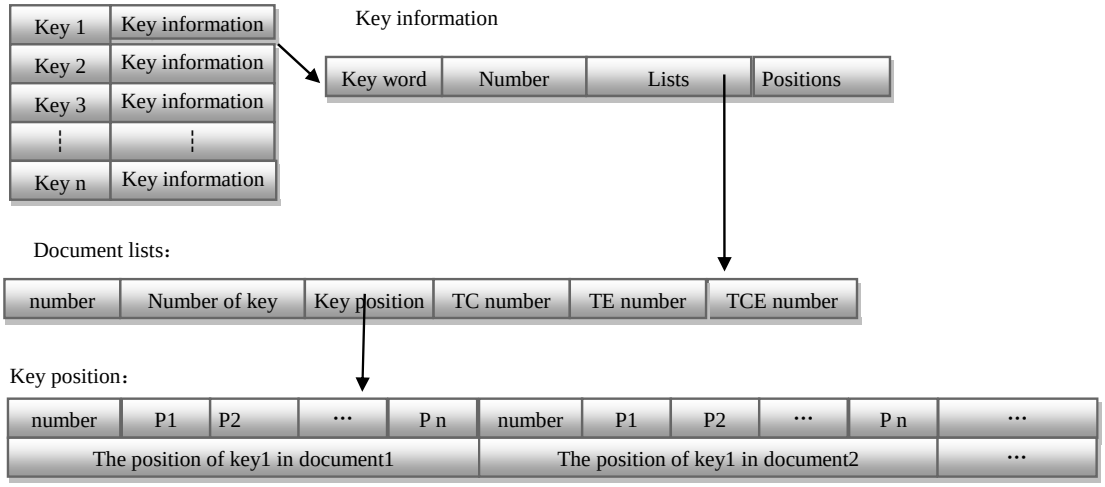

Figure7. Structure of indexing

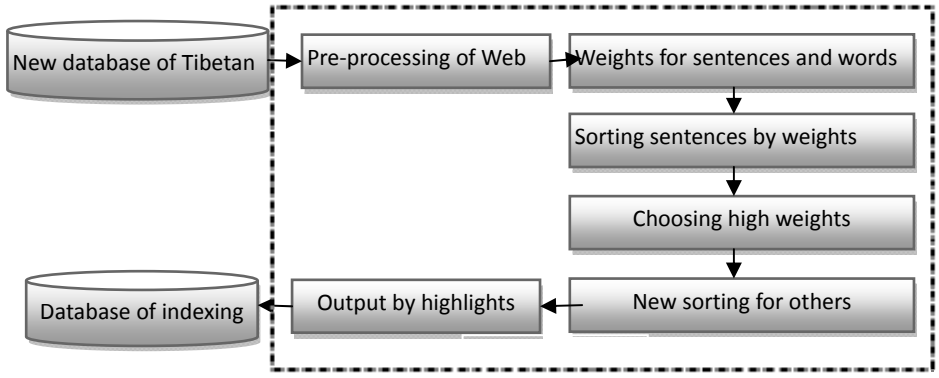

Figure8. Module of output results

REFERENCES

[1] Yu-zhong, C., L. Bao-li, et al. (2003). "The Design and Implementation of a Tibetan Word Segmentation System." Journal of Chinese Information Processing3: 15-20.

[2] Lu yajun, A study of Layout and Input Method of A General Tibetan Computer Keyboard .Journal of Chinese information Processing .2006,20(2):78-86

[3] Yong-hong, L., H. Xiang-zhen, et al.. "Tibetan coding method and mutual conversion." Journal of Computer Applications 2009, 29(7):2016-2018

[4] Rui jian-wu,Wu jian etc. .Study on Implementing Tibetan Operation System Based on ISO/IEC 10646, Journal of Chinese information processing.2005,19(5):59-66

[5] Akamine, S., Y. Kato, et al. "Development of a large-scale web crawler and search engine infrastructure. " 3rd International Universal Communication Symposium, IUCS 2009, December 2009, Tokyo, Japan, Association for Computing Machinery

[6] Sotiris Batsakis,Euripides G.M. Petrakis,Evangelos Milios et al.Improving the performance of focused web crawlers.Data & knowledge engineering,2009,68(10):926-945.

[7] Chun yan. Design and implementation of Tibetan encoding identification and conversion. Southwest Jiaotong University：Master thesis .2010

[8] Yu-zhong, C., L. Baoli, et al. (2003). "An Automatic Tibetan Segmentation Scheme Based on Case Auxiliary Words and Continuous Features." Applied Linguistics 1: 75-82

[9] Pu bu-tanzeng. Study of the Methods of Tibetan word segmentation Tibet University：Master thesis.2010.5

[10] Bar-Yossef, Z. and M. Gurevich (2008). "Random sampling from a search engine's index." Journal of the ACM (JACM) 55(5): 1-74.

[11] Li, G., J. Feng, et al. (2011). "Providing built-in keyword search capabilities in RDBMS." The VLDB Journal 20(1): 1-19

[12] Lin, R. R., Y. H. Chang, et al. (2011). "Improving the performance of identifying contributors for XML keyword search." ACM SIGMOD Record 40(1): 5-10.