Research Report on Tibetan Collation (TU)

Planned Completion Date: 07/15/08

Title: Research and implementation of Tibetan Collation based on Unicode and GB Tibetan

Researchers: Ngodrup NyimaTrashi DruJie GeSangDuoJi QunNuo RenqinNuobu BianBa WangTui Yong Tso

Presentation: Algorithm, Software, Documentations and papers

Abstraction:

Sorting issue in Tibetan language is one of the key components of Tibetan Information Technology, and is also one of the most important indicators to represent the degree of the development of Tibetan information technology. It not only reflects the path of the development of Tibetan information technology, but even more importantly provides unprecedented technical supporting for the works like file searching, information searching, text sorting and etc.

Taking a view of characteristics of Tibetan language, the paper designs a Tibetan sorting algorithm by analyzing Tibetan grammar rules and basic sorting rules of major Tibetan dictionaries. The algorithm resolves the issues of Tibetan sorting through four key modules, respectively the Root-letter Recognition Algorithm, Priority Algorithm, Sorting-related Digital Code String Access Algorithm and QuickSort Algorithm. In the process of designing the priority algorithm, the paper separates the priority algorithm into three different modules as structure priority, component priority and character priority by considering the complexity of Tibetan language and needs of Tibetan sorting. Since consonant characters of Tibetan language present the orderly nature, the paper proposes creatively the Root-letter Recognition Algorithm and three Priority Algorithms according to the basic rules of Tibetan sorting.

The Root-letter Recognition Algorithm is firstly to accurately extract the root letter from each syllable in Tibetan and assign the syllable into the corresponding sorting group by recognizing the root letter. Then the structure priority deals with the sorting issues for those words with different structure and same root letter in each syllable, the component priority deals with the sorting issues for those words with same structure and different components in each syllable, and the character priority deals with the sorting issues for those words with same structure and components, but different component elements in each syllable. Thus, the algorithm not only resolves fundamental issues of Tibetan sorting, but decreases the complexity of time and space, which presents strong vitality.

In Tibetan, different phrases have different numbers of syllables, and each syllable might include seven component elements. The median of digital code strings that generate these component elements reaches to over 28 bit, and the number increases dramatically while the numbers of syllables increase. Considering the facts listed above and complexity of Tibetan language, the maximum length of the syllable is set in the process of designing algorithm. While the numbers of syllables increase, new issues are emerged in the storage of digital code strings. 32 bit PC is unable to hand directly the numerical sequence for it exceeds the limit of digital capacity. In order to

overcome the problem, the paper turns the original digital-formatted code strings into text format, which successfully solves the issue.

This algorithm has been accomplished in Windows 2000/ NT/XP Operating System.

Keywords: Tibetan Sorting, Root-letter, Structure Priority, Component Priority, Character Priority, Algorithm

The Contents for project:
(1) to study the Tibetan language sorting rules;
(2) A Study on Root-letter Recognition Algorithm based on Tibetan syllables;
(3) A Study on structure priority and Character Priority algorithm ;
(4) A Study on one syllable and multi-syllable sorting algorithm in Tibetan;

Innovations:

Although Microsoft Corporation has designed Tibetan collation algorithm based on the ISO/IEC10646 Tibetan code character set, and has developed Tibetan collation software, but its collation principle and the result simply do not conform to Tibetan sorting principle, thus has caused its algorithm and the software cannot put in the practical application. At the same time, although the country also has the expert to conduct the research, but was still at the research stage, had not discovered that the collation software realizes. Therefore, realizes regarding Tibetan writing collation algorithm and the software, internationally was still at the blank condition. Therefore, this project based on the ISO/IEC10646 Tibetan code character set, through the deep research Tibetan collation rules, profits from the experts in Tibetan collation aspect part research results, the original creation collation algorithm, fundamentally has solved the problem which Tibetan collation.

Status of Activity: finished

Software： GangJian (Snow eyes) Tibetan Collation software is made in TU and used in MS-Office for Tibetan Collation

Papers：

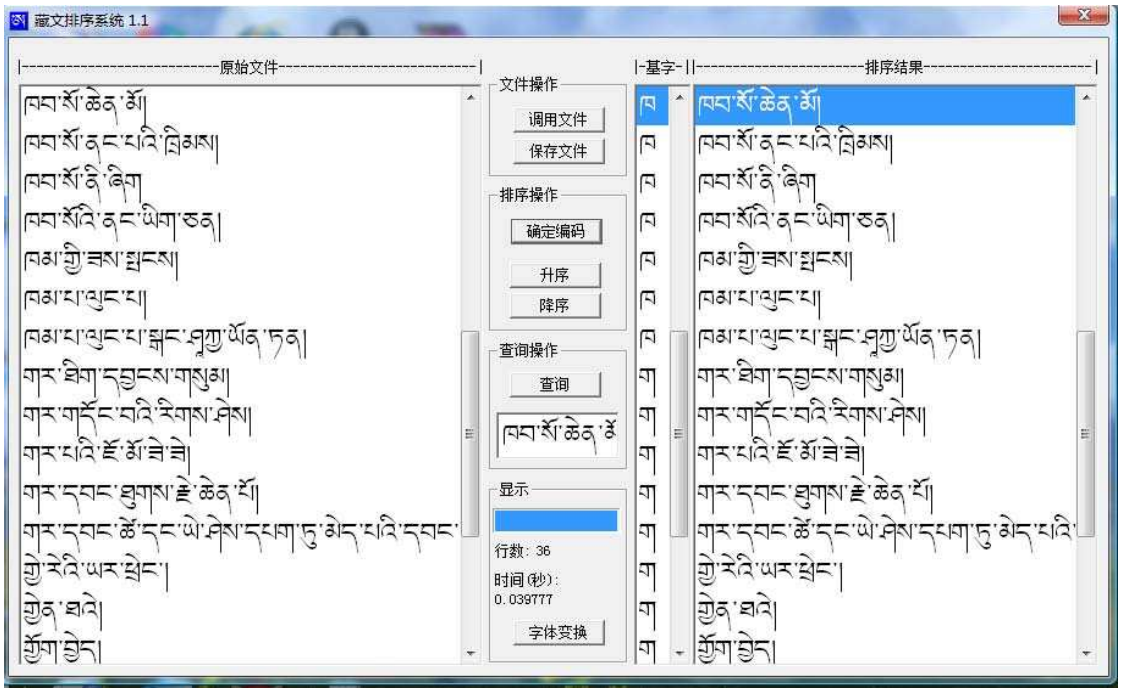
(1) BianBa WangTui, A Study on Root-letter Recognition Algorithm Based on ISO/IEC10646 International Standards for Tibetan Coded Character Set（基于国际编码标准小字符集藏文排序算法研究之识别基字算法）, Journal of Tibet University. 2007.3

(2) Druggye Ngogdrup, A Method for Ordering Tibetan Text Based on Tibetan Coded GB（基于藏文编码 GB 的藏文排序方法研究）, Journal of Tibet University，2008.1

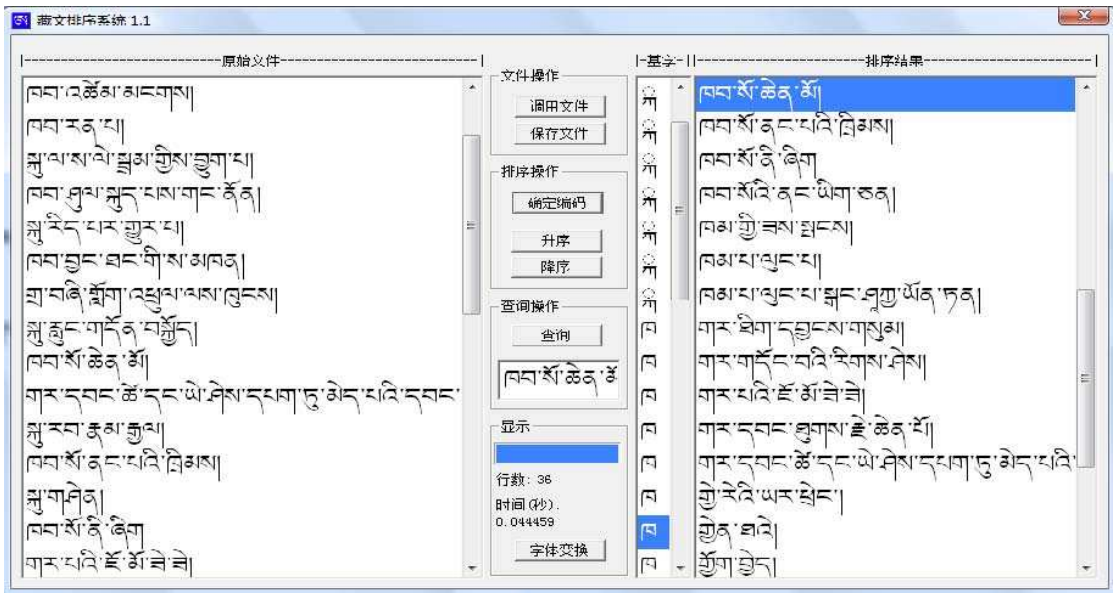
Result:

This software testing method is through the different test object, looked that test result and dictionary sorting to be whether same, if is the same, then tests successfully, otherwise examines or calculates the error rate.

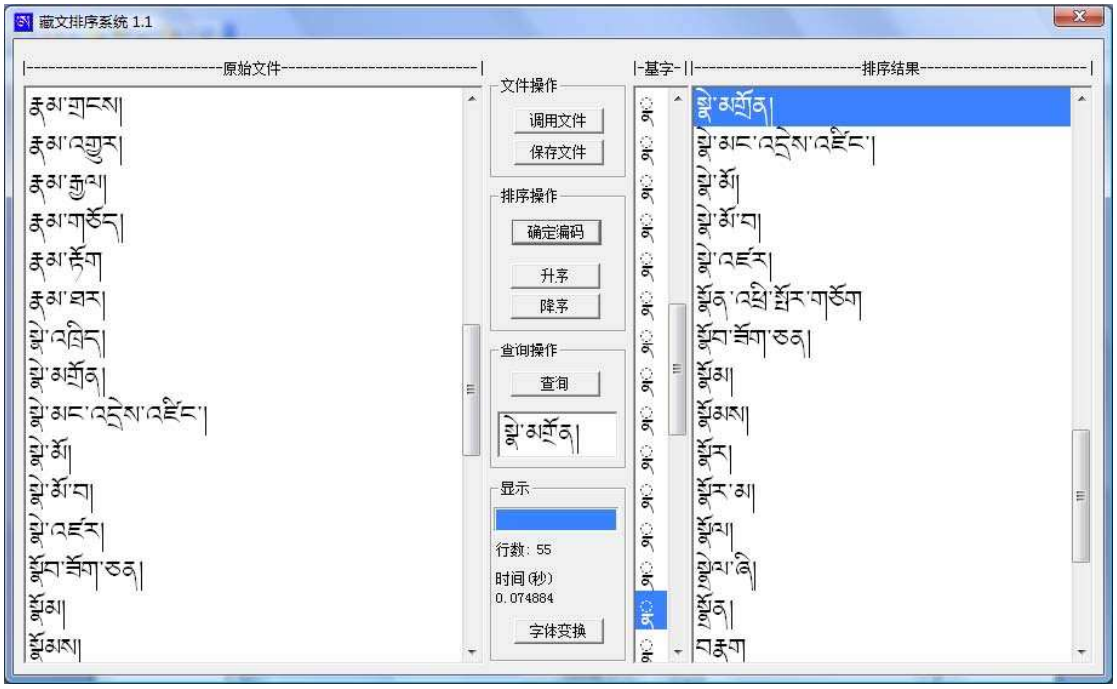
1. Donggha Tibetan Studies dictionary test



2. Tibetan Studies dictionary test



3. Beijing Nationality Publishing house tests (北京民族出版社测试)



The test number of lines and uses the time(测试行数及所用时间)