Research on Vertical Search Engine Based on Tibetan News Sites

Zhiqiang Han [a], Guixian Xu [b,] * and Wei Sun [c]

School of Information Engineering, Minzu University of China, Beijing 100081, China

[a]hanzq@sohu.com, [b]xuguixian2000@sohu.com, [c]645281904@qq.com

Corresponding author: Guixian Xu

Keywords: Tibetan, news sites, vertical search engine.

Abstract. In the paper, we detailed introduce the features of Tibetan language and related technologies that we use to deal with Tibetan language with computers. A specified architecture thinking method is proposed by analyzing the development of the Tibetan news sites and taking the development of vertical search engine into consideration as well.

Introduction

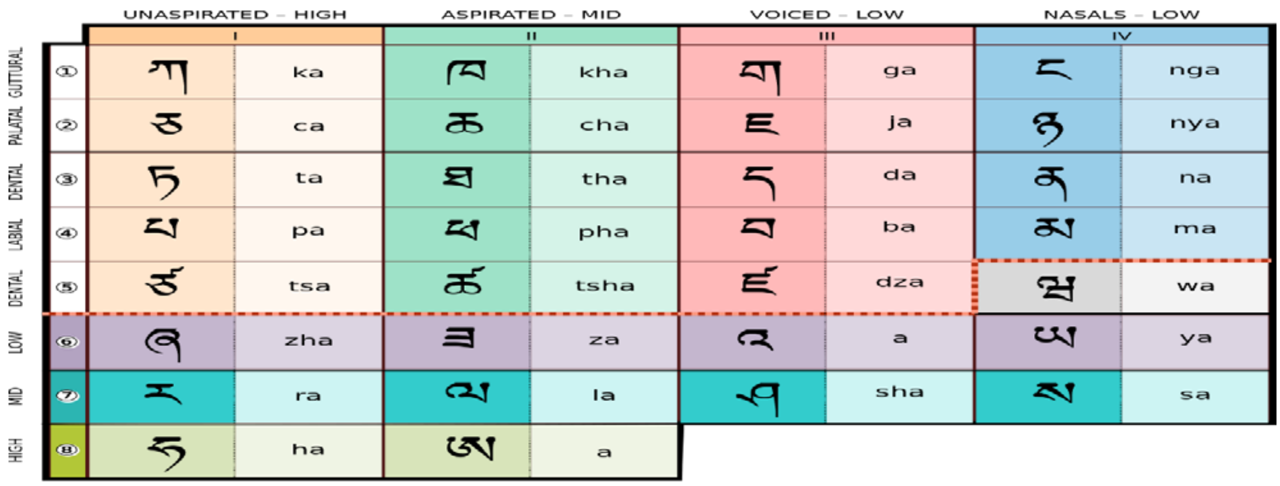
As the Internet technology matures, the growth of the network shows nonlinear growth trend and its speed varies which tremendously influence our social life and work. In this scenario, it's extremely important for us to find the real effective information quickly and accurately. But the traditional search engine, such as Google and Baidu which belong to full text search engine, cannot present the contents that we need accurately, reasonably and comprehensively within a specified professional field. For the past few years, vertical search engine emerges as a new type search engine, and it provides the possibilities to search information in a designed and specialized field for us. As a result, people pay more and more attention to it. Our country is a nation that has multiple nationalities, multiple cultures and a time honored history, among which Tibetan areas is vast and the population is large. Tibetan sites have been emerging since the foundation of the country. There are still many problems to deal with in the contents search work of the Tibetan sites, in which the vertical search engine field has been in a blank state ever since. It is extremely unfavorable for either the unity of 56 nationalities or the spread of Tibetan culture.

Based on the intrinsic characteristics of the Tibetan language and the whole process vertical search engine obtains data and extract useful information, this paper emphatically analyses the possibility to build a vertical search engine architecture that has high reliability and high accuracy based on Tibetan news sites.

Related work

Structure characteristics of Tibetan letters. Tibetan current situation. Tibetan language is the general and native language of the people in the Tibetan areas of our nation. The number of Tibetan people will be above 6 million currently and their residential area locates mainly in Tibet Autonomous Region, Qinghai Province, Gansu Province, Sichuan Province and Yunnan Province etc. Tibetan language belong to Tibetan branch of Sino-Tibetan of Tibeto-Burman languages, and at present the Tibetan language in our nation is mainly divided into three literal dialects: Tibetan, Kang, Amdo. Tibetan language derives from the ancient Sanskrit and the western text 13 hundred years ago. Tibetan characters can be transliterated into Devanagari word by word. Also, Tibetan is similar with Chinese pinyin and they are all alphabetic writings. Tibetan consists of consonants, vowels and punctuation marks. The number of consonants is 30, and the signs and pronunciations are listed as follows in Fig. 1 [1].

The signs and pronunciations of vowels are listed as follows in Table 1 [2].



		UNASPIRATED – HIGH		ASPIRATED – MID		VOICED – LOW		NASALS – LOW	
		I		II		III		IV	
GUTTURAL	①	ཀ	ka	ཁ	kha	ག	ga	ང	nga
PALATAL	②	ཅ	ca	ཆ	cha	ཇ	ja	ཉ	nya
DENTAL	③	ཏ	ta	ཐ	tha	ད	da	ན	na
LABIAL	④	པ	pa	ཕ	pha	བ	ba	མ	ma
DENTAL	⑤	ཙ	tsa	ཚ	tsha	ཛ	dza	ཝ	wa
LOW	⑥	ཞ	zha	ཟ	za	འ	a	ཡ	ya
MID	⑦	ར	ra	ལ	la	ཤ	sha	ས	sa
HIGH	⑧	ཧ	ha	ཨ	a				

Fig. 1

Table 1 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)
◌ེ	འགྲེང་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)

The punctuation marks of Tibetan mean section sign and dividing line and syllable-dividing marks of syllable point, phrases, the single hanging operator at the end of the sentences and the double hanging operators at the end of the chapters are all included in it. With the deepening of the reform and opening-up policy, the punctuation marks in Chinese can be found in Tibetan as well.

Structure of Tibetan words [3].Tibetan is different from both Chinese and English, so the method to deal with Tibetan sites should be considered separately. Each syllable in Tibetan has a base word, by which the center consonants of the syllable is decided. Vowel affix can be added above or under the base word, which means different vowel. Added-words can be found just around the base word that plays different roles. The written order of Tibetan is in accordance with the modern mainstream words: laterally written from left to right.

Words and expressions in Tibetan consist of syllable and can be classified as single-syllable word and polysyllabic word. Punctuation marks separate each syllable, while there are not obvious boundary separators between words. Tibetan words can be classified as 13 categories, i.e. verbs, nouns, adjectives, pronouns, numerals, quantifiers, adverbs, conjunctions, prepositions, auxiliary, modal, interjection and onomatopoeia. The ways of word formation in Tibetan are to compound and to derive words, in which the derived ways account for most. In the derived words, there are mainly suffix, less prefix and infix. 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.

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[4].

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 [5]. Later on a new approach is added into the segmentation methods in Tibetan, i.e. methods based on combination of rules and statistics [6]. 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.

Distribution and features of Tibetan sites. Current situation of Tibetan sites [7] 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 [8]. *"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 just as Fig. 2 shows.

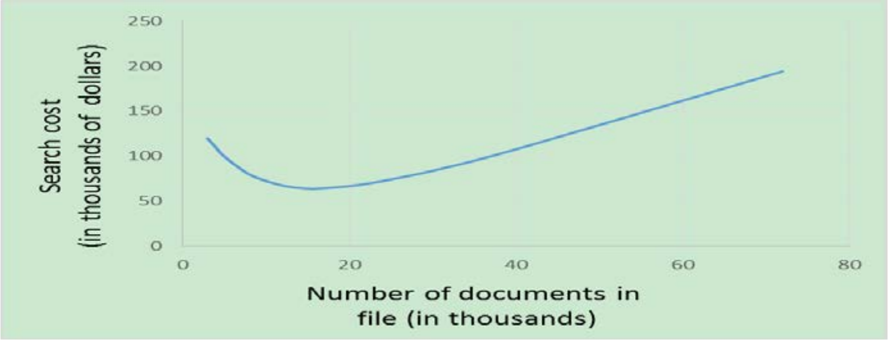


Fig. 2

Classifications of Tibetan sites [9].The report also gives us the current Tibetan sites' distribution as Fig. 3 shows:

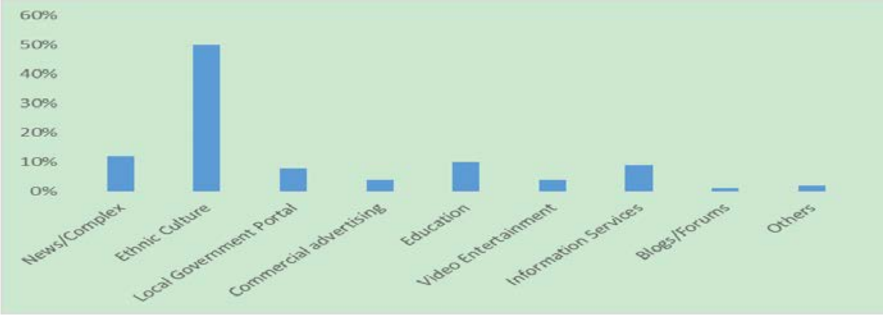


Fig. 3

Languages cases of Tibetan sites. The languages used in Tibetan websites has been in the statistical analyze in the report as Fig. 4 shows.

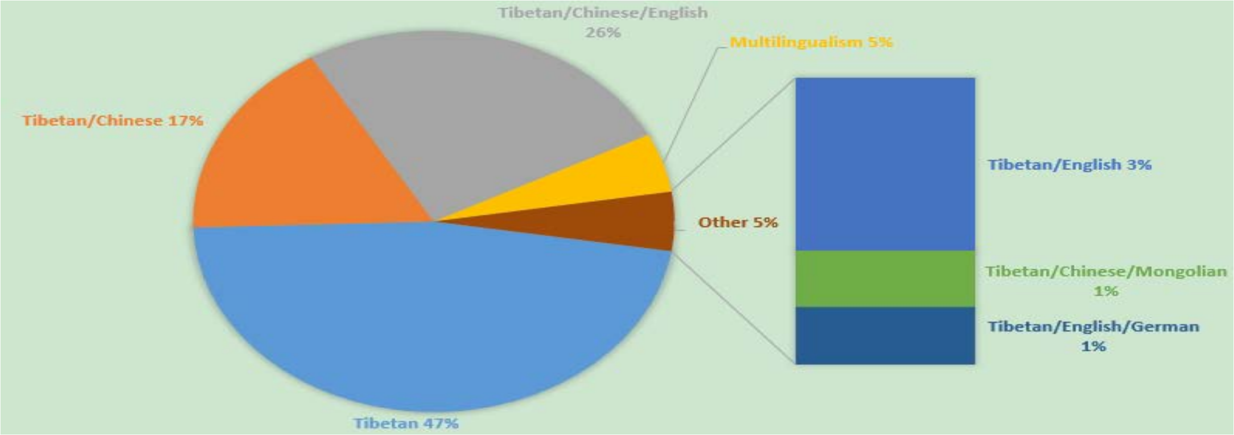


Fig. 4

Vertical search engine technology. History and current situation of search engine [10] 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.

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 [11].

Advantages of vertical search engine [12]. 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.

Application in Tibetan website of vertical search engine. Possibilities to build a vertical search engine in Tibetan sites**.** Vertical search engine has been playing its unique role no matter in Chinese or in English so far [13]. It develops in almost all fields, i.e. electricity commerce, medical treatment, finance, real estate and tourism, and has a strong momentum. After my research, I am aware of the fact that vertical search engine has not stepped into the search field of Tibetan sites. However the people who speak Tibetan has been increasing by years and the development potential is unpredictable. It may contribute to the development of vertical search sites in Tibetan areas.

Necessities to build a vertical search engine in Tibetan sites. By the above statistics on Tibetan websites we find that the Tibetan websites' number is increasing these years. Yet it's still hard for us to find the real useful things in a short time by searching in so much information. In this case, vertical search engine points out a bright way for us [14]. By secondary processing information, we can ensure the accuracy of the information and it greatly reduces the time it takes to search information on the Internet with the work efficiency improved at the same time.

Summary

The number of people who use Tibetan every day is increasing and the word processing work gradually develops. There are several algorithms that are used to process Tibetan websites and the difficulties we meet when using computers to handle the web pages are getting less. The development of vertical search engine offers us the possibility of technical support and implementation to apply the vertical search engine in the development of Tibetan news sites.

Acknowledgement

This paper is supported by the Fundamental Research Funds for the Central Universities. The project title is "The Research on Hot Topic Discovery and Tracking Technology of Tibetan Network". This work is also supported by "Beijing Social Science Foundation (No.14WYB040)".

References

[1, 2] Information on http://en.wikipedia.org/wiki/Tibetan_alphabet

[3] Lirong Qiu: Intelligent Networks and Intelligent Systems (ICINIS), 2013 6th International Conference on (Shenyang, 1-3 Nov. 2013), p.256-259.

[4] Guixian Xu: Intelligent System Design and Engineering Application (ISDEA), 2012 Second International Conference on (Sanya, Hainan, 6-7 Jan. 2012), p.610-612.

[5] Chen Yuzhong, Li Baoli, Yu Shiwen and Lan Cuoji: The First Session of the Seminar Students Computational Linguistics [C] (China, August, 2002).No 1.

[6] Meng Xianghe: Key Technology Research on Tibetan Websites Topic Detection and Tracking (MS., Northwest University for Nationalities, China 2013).

[7] Zhang Huaqiu, Yu Hongzhi, Chen Xinyi, Chen Xugang: Computer Science and Network Technology (ICCSNT), 2011 International Conference on (Harbin, 24-26 Dec. 2011), p.1411-1414.

[8] Maosong Sun: First National Seminar on Tibetan information processing (Qinghai Normal University, August 16th, 2014).

[9] Xiaodong Yan, Yuan Sun, Xiaobing Zhao, Guosheng Yang: Information Science and Engineering (ICISE), 2010 2nd International Conference on (Hangzhou, China, 4-6 Dec. 2010), p.3388-3391.

[10] Shih-Fu Chang, Chen, W., Meng, H.J., Sundaram, H., Di Zhong: Circuits and Systems for Video Technology, IEEE Transactions on (Sep 1998). vol.8, no.5, p.602-615.

[11] Information on http://en.wikipedia.org/wiki/Web_search_engine

[12] Xiao Dongmei: Researches In Library Science, (2003) No. 2, p. 87-89.

[13] Granados, N.F., Kauffman, R.J., King, B.: Hawaii International Conference on System Sciences, Proceedings of the 41st Annual (Waikoloa, HI, 7-10 Jan. 2008). p. 389-389.

[14] Xiao Dongmei: Researches In Library Science, (2003) No. 2, p. 87-89.