The Methods Comparison of Collecting Mongolian Websites

[1,2]Zhijuan Wang, [1]Yinghui Feng

[1]*The College of Information Engineering, Minzu University of China, Beijing, China*
[2]*Minority Languages Branch, National Language Resource Monitoring & Research Center, Beijing, China*

Corresponding author is Zhijuan Wang

Abstract
Not more than 10 lines. Mongolian websites are very important for studying the Mongolian language and Mongolian culture. The domain names and character encodings of Mongolian websites are complex. The common methods couldn't collect all Mongolian websites in all character encodings. This paper introduces the features of Mongolian websites firstly. Then five methods are introduced: getting websites from the management institutions of domain name, obtaining websites information from navigation websites, the method based on meta-search engine, the method based on hyperlinks and the hybrid method. At last, several methods are compared in terms of the recall rates, the precision and F-measure. The experiment results show that the hybrid method has high recall rate (79%) and precision (97%).
Keywords: MONGOLIAN WEBSITES, HYPERLINKS, MONGOLIAN CHARACTER ENCODINGS, HIGH-FREQUENCY WORDS

1. Introduction

Along with the development of Internet technologies and information construction of Mongolian, the total number of Mongolian websites is becoming larger. These network resources are important for studying Mongolian language and Mongolian culture. Moreover, they are useful for realizing full text search engine in Mongolian. The challenge comes from the fact that the domain names and character encodings of Mongolian websites are complex. It is difficult to collect all Mongolian websites in all character encodings. Therefore, it is necessary to study the methods of collecting Mongolian websites according to the features of Mongolian websites (Dai Yu-gang, 2007).

The rest of the paper is organized as follows. Section 2 presents the features and definition of Mongolian websites. Then two common methods of collecting websites are discussed in Section 3. Section 4 introduces three methods of collecting Mongolian websites in detail. In Section 5, several methods are compared in website numbers, recall rates, precision rate and F-measure. Finally, the conclusions are given in Section 6.

2. The Features and Definition of Mongolian Websites

The first Mongolian website was set up on December 18, 2006 by Inner Mongolia Menksoft Software Company. The website is the first portal of Mongol-Chinese bilingual comprehensiveness in China. It is the first time that the website has realized the vertical typesetting from left to right.

2.1. The Features of Mongolian Websites

Through previous researches, the features of Mongolian websites are summarized as follows:

A. The number of Mongolian websites is small

The number of Mongolian websites is smaller compared with Chinese and English websites. The

number may be less than 1000. Therefore, it is necessary to collect Mongolian websites as many as possible.

B. The domain names are complex

A website is composed of a web server and a domain name. Generally, a website has a separate second-level domain, such as "sohu.com". The third-level domain – "news.sohu.com" – is a sub-website of "souhu.com" rather than a separate website. However, for Mongolian websites, the domain names are very complex. Three conditions are listed below:

- The URL of Mongolian website is a second-level domain (2LD).

For example, "http://www.hlberb.com/" is a Mongolian website and its URL is 2LD.

- The URL of Mongolian website is a third-level domain (3LD).

For example, "http://www.ttcy.com" website has three language-versions: Chinese, Mongolian and Slavic. The URL of Chinese version is 2LD ("http://www.ttcy.com/"), but the URL of Mongolian version is 3LD ("http://mo.ttcy.com/").

- The URL of Mongolian website is a subdirectory of 2LD.

For example, the domain name of Naimanqi people's government website in Mongolian is "http://www.nmqnw.cn/mgl". It is a subdirectory of 2LD.

Therefore, the domain names of Mongolian websites can be 2LD, 3LD or the subdirectory of 2LD.

C. The Mongolian character encodings are complex

From early 1980s, scholars from China, Mongolia (such as National University of Mongolia), Germany (such as Berlin's Free University), and Japan (such as Tokyo University of Foreign Studies) have developed different Mongolian character encodings. There are more than ten kinds of Mongolian character encodings and six kinds of them are used commonly in mainland of China. The encoding types and range are shown in Table 1. (Jin Wei, 2009)

In order to see the using status of character encodings in Mongolian websites, 50 Mongolian websites are selected randomly and their encodings are counted. Figure 1 displays the percentages of Mongolian websites in different character encodings.

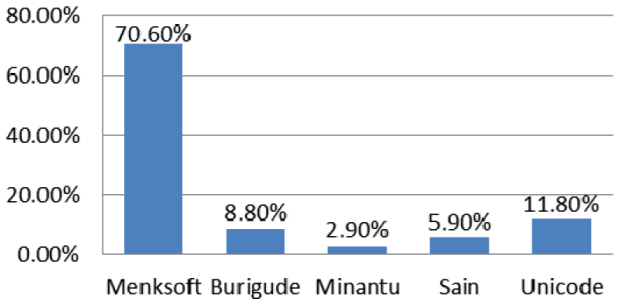

Figure 1. The Percentage of Mongolian Encodings Used in Websites

As can be seen in Figure.1 above, five Mongolian encodings are used in Mongolian websites and the percentage of Mongolian websites in Menksoft is the biggest (70.6%). The percentage of Mongolian websites in Unicode is 11.8%. And the percentage of other encodings is less than 10%. It is clear that at least five encodings are used in Mongolian websites.

2.2. The Definition of Mongolian Websites

From above all, we can see that Mongolian websites are not similar to the common websites, such as Chinese or English websites. In order to collect Mongolian websites accurately, the definition of Mongolian websites should be given according to the features of Mongolian websites.

The definition of Mongolian websites is: the domain name of a Mongolian website can be a second-level domain, a third-level domain or subdirectories of second-level domain. a Mongolian website has two-stage or more than two-stage directory and the percentage of Mongolian characters of every webpage must be above 90%.

3. The Common Methods of Collecting Websites

There are two common methods to collect websites: getting websites information from the management institutions of domain name and getting websites information from navigation websites.

Table 1. Mongolian Character Encodings Used in Mainland of China

Name of encoding	Encoding type	Range of character encoding
Menksoft	Hybrid encoding	0xE264-0xE34F
Unicode	Phonetic encoding	0x1800-0x18AF
Fangzheng	Hybrid encoding	0x766D-0x781B
Oyuta	Hybrid encoding	0xE250-0xE377
Sain	Hybrid encoding	0xE246-0xE355
Mingantu	Hybrid encoding	0x254-0xE33E
Burigude	Shape encoding	0xE246-0xE29F

3.1. Getting Websites Information From The Management Institutions of Domain Name

Every domain name should be registered from management institutions of domain name. In mainland of China, China Internet Network Information Centre (CNNIC) is responsible for Chinese domain name registry. (Wang Zhi-juan, 2013)

In previous researches, we have noticed that the domains of Mongolian websites are complicated. The domains of some Mongolian websites are second-level, and these websites can be gotten from CNNIC directly. However, there are many Mongolian websites that their domains are third-level or subdirectories of second-level. Therefore, this method can't be used to collect all Mongolian websites.

3.2. Obtaining Websites Information From Navigation Websites

Obtaining websites information from navigation websites is a useful method to collect websites in a certain language or field. In this paper, the biggest Mongolian navigation website –"http://www.mongol.cn/lan/mn/index.html" – is taken for example, to show the percentage of Mongolian websites in all websites. The navigation website provides 100 hyperlinks of Mongolian website. However, Figure 2 shows that 46% can't be opened, and 14% websites are not Mongolian websites, and only 40% of hyperlinks are Mongolian website. So this approach isn't promising for collecting all Mongolian websites.

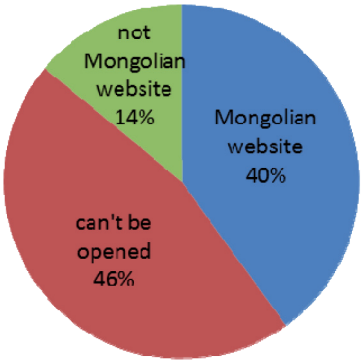

Figure 2. The Percentage of Mongolian Websites in Mongolian Navigation Website

4. Methods of Collecting Mongolian Websites

4.1. Collecting Mongolian Websites Based on Meta-Search Engine

All Mongolian web pages contain Mongolian high-frequency words. So Mongolian high-frequency words (such as "᠋᠋ ᠋᠋ ᠋᠋") can be used to collect Mongolian websites (Ma Hong-yuan *et al*, 2012. Feng Yan-hui *et al*, 2011. Li Hong-mei *et al*, 2008).

Because of multiple character encodings of Mongolian, it is necessary to collect web pages containing high-frequency words in different character encodings in meta-search engine. Although the method can collect some Mongolian websites, it is not effective for Mongolian websites in all kinds of character encodings.

Taking a high-frequency word "᠋᠋" as an example, which means "of", the searching results are shown in Figure 3. It is clear that, for the same high-frequency word, Google could not collect Mongolian websites in all character encodings.

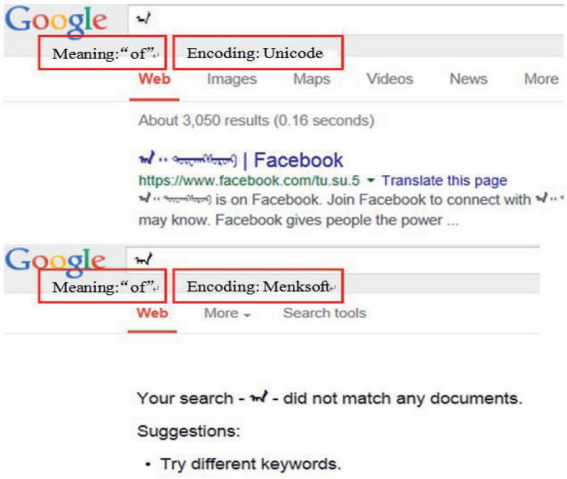

Figure 3. The different searching results of "᠋᠋" in unicode and menksoft

Different meta-search engines were tested for collecting Mongolian websites in all kinds of character encodings. The results are given in Table 2. Some meta-search engines can only collect Mongolian websites in Unicode. Some meta-search engines can

Table 2. Searching Results of the Method Based on Meta-search Engine

Meta-search engine	The number of websites	The number of Mongolian websites		Type of character encoding
Google	32	30	64	Unicode
360	48	41		Burigude/Menksoft/Sain/Mingantu/Unicode
Bing	42	40		Burigude/Menksoft/Sain/Mingantu/Unicode
Baidu	29	25		Unicode
Yahoo	25	22		Unicode

collect Mongolian websites in five character encodings. But its number is less than the total number of Mongolian websites. Therefore, the method based on meta-search engine isn't effective in collecting Mongolian websites.

4.2. Collecting Mongolian Websites Based on Hyperlinks

As we have introduced in Section 3, the common methods are less effective in searching Mongolian websites. Though web crawler can be used to collect Mongolian websites (Wang Qi *et al*, 2005. Chen Zhumin *et al*, 2009), the number of hyperlinks fetched by web crawler is growing exponentially. Some studies show that the proportion of internal links in a web page reaches 60%-80%, and some are even higher (Jin Wei, 2009. Wang Xiao-yu *et al*, 2003. Belhumeur, P. N *et al*, 1997. Qiu Jun-ping *et al*, 2005). About 90% external links are friendly links which are useful, meaning that the probability of Mongolian websites hyperlinks is very high. Based on this, the method based on hyperlinks is proposed.

The algorithm based on hyperlinks designs a special filter to extract external links. Firstly, a URL is input and depth is set in a web crawler. Secondly, <a> and <frame> tags are gotten using "OrFilter" (Dice Holdings, Inc. Htmlparser, 2006). Thirdly, URLs corresponding to the tags are got. Finally, repetitive URLs are filtered.

In this method, recognizing the language of web pages is conducted manually. Five character encodings are included in 103 websites. As hyperlinks are irrelevant to the character encodings, Mongolian websites in all kinds of character encodings can be collected with the method based on hyperlinks.

4.3. Hybrid Method of Collecting Mongolian Websites

The method based on hyperlinks can collect Mongolian websites in all kinds of character encodings. However, the process of recognizing the language of web pages costs lots of time. To minimize manual intervention, statistical method is used to recognize language.

Before recognizing language, Mongolian high-frequency words must be gotten. Some scholars from Inner Mongolia University researched the characteristic symbols of Mongolian, founding that some auxiliary verbs (e.g. "᠂ ᠃ ᠄") can be used to recognize Mongolian web pages (Wang Rui, 2008. Bai Shuangcheng *et al*, 2013. Wang Ling *et al*, 2013. Nasun urt, 2011). About 1000 texts are used to do word frequency statistics. There are about 400000 words and 20000 kinds of words. The word weights are got according to TF*DF and the top 20 high-frequency words are selected.

The goal is to judge the language of candidate web pages according to Equation 1, and web pages are selected if their scores above a threshold. Finally, use the URL of Mongolian web page to find the URL of Mongolian website.

$$\text{score}(\text{web page}) = \sum_{i=1}^{n} tf(i) \quad (1)$$

In Equation 1, tf(i) is the term frequency of high-frequency words. This hybrid method combines the method based on hyperlinks with statistical method. Mongolian websites in all kinds of character encodings can be collected and the total efficiency of collecting Mongolian websites is improved greatly.

5. Comparison

In this Section, four methods are compared in four aspects. Table 3 shows that the method based on hyperlinks and the hybrid method have high recall rates than other methods. In addition, the precision and F-Measure of hybrid method are the highest of the four methods. The total time is composed of computer and manual work. Considering the total time, hybrid method takes less time. Therefore, the hybrid method achieves a good performance.

6. Conclusion

Mongolian websites are very important for sharing Mongolian information and spreading Mongolian culture. Collecting Mongolian websites is useful to study Mongolian language and Mongolian culture. In this paper, several methods of collecting Mongolian websites are introduced and compared. The hybrid method is generic, and in such sense it can be applied to collect websites in other minority languages. Experiments have shown that the hybrid method has higher precision and recall rate. Moreover, this method takes less manual work. Future work will investigate more novel ways to recognize different minor-

Table 3. The Technologies of Collecting Mongolian Websites

Search method	The number of websites	Recall rate	Precision	F-Measure
navigation website	40	30.77%	40.00%	34.78%
based on meta-search engine	65	49.23%	85.71%	62.54%
based on hyperlinks	103	79.23%	74.10%	76.58%
hybrid method	103	79.23%	97.08%	87.25%

ity languages and classify minority texts. Ultimately, machine learning tools will be used to achieve more robust and intelligent methods.

Acknowledgements

This work is sponsored by Key Program of National Natural Science Foundation of China (No. 61331013), National Language Committee of China (No.WT125-46 and No. WT125-11), and also supported by Graduate Students Projects of Minority Languages Branch, National Language Resource Monitoring & Research Center (No.CML15A02).

References

1. Bai Shuang-cheng, Zhang Jin-song and Husile. A comparison study on word coding methods for Mongolialn IME. Journal of Chinese Information Processing 2013, pp: 169-174.
2. Belhumeur, P. N., Hespanha, J. P. & Kriegman, D, J. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, pp: 711-720.
3. Chen Zhu-min, Ma Jun, Han Xiao-hui and Lei Jing-sheng. Focused crawling oriented multigranular priority computation for URLs. Journal of Chinese Information Processing, 2009, pp: 31-38.
4. Dai Yu-gang, The study of Tibetan web page collecting technology. China Minorities Information Technology Institute--proceedings of the 11th national conference on China Minorities information technology, 2007.
5. Dice Holdings, Inc. Htmlparser. Available from:http://htmlparser.sourceforge.net/javadoc/org/htmlparser/filters/OrFilter.html[Accessed 2015], 2006.
6. Feng Yan-hui, Hong Yu, Yan Zhen-xiang, Yao Jian-min and Zhu Qiao-ming, A novel method for bilingual web page mining via search engines, Journal of Chinese Information Processing, 2011. pp:71-78.
7. Jin Wei. Research of Mongolian information retrieval model. Inner Mongolia: Inner Mongolia University, 2009.
8. Li Hong-mei, Ding Zhen-guo, Zhou Shui-sheng and Zhou Li-hua. Clustering method of web search results. Journal of Chinese Information Processing, 2008, pp: 56-63.
9. Ma Hong-yuan and Wang Bing. Query results caching and prefetching in web search engines based on user characteristics. Journal of Chinese Information Processing, 2012, pp: 19-26.
10. Nasun urt. Construction and application of mongolian languages knowledge bank. Journal of Chinese Information Processing, 2011, pp: 162-165.
11. Qiu Jun-ping and Duan Yu-feng. Study on Webometric (Part 3) – Study on the distribution of measurements on indexes of university websites hyperlinks. Journal of the China Society for Scientific and Technical Information, 2005, pp: 407-413.
12. Wang Ling, Dawa Yidemucao and Wu Shouer Silamu. An investigation research on the similarity of Uyghur Kazakh Kyrgyz and Mongolian languages. Journal of Chinese Information Processing, 2013, pp: 180-186.
13. Wang Qi, Song Guo-xin and Shao Zhi-qing. Link-based ranking algorithms in information retrieval. Journal of East China University of Science and Technology, 2005, pp: 455-458.
14. Wang Rui. 2008. Mongolian Web Spider, Text encoding recognition and conversion research. Inner Mongolia: Inner Mongolia University.
15. Wang Xiao-yu and Zhou Ao-ying. Linkage analysis for the world-wide Web and its application. Journal of Software, 2003, pp: 1768-1780.
16. Wang Zhi-juan. The key technologies of collecting Tibetan websites. International Journal of Emerging Technologies in Learning (iJET), 2013, pp: 30-34.