Research on Automatic identification of Chinese minority language websites

Haifeng Liu [1, a], Yuanyuan Yang [2, b], Jing Li[3, c] and Zhiqiang Han[1, d]

[1]School of information and engineering, Minzu University of China, Beijing 100081, China

[2]School of minority language and literature, Minzu University of China, Beijing 100081, China

[3]School of Kazak language and literature, Minzu University of China, Beijing 100081, China

[a]liuhaifeng_2011@126.com,[b]wateryoo919@aliyun.com, [c]61714561@qq.com, [d]hanzq@sohu.com

Keywords: minority language websites, identification, feature.

Abstract. This paper presents features of Chinese minority language text collection on websites, analyses the problems of webpage identification of Chinese minority language text, and proposes three automatic identification methods. Based on these methods, designs and realizes software to identify Chinese minority language text such as: Mongolian, Tibetan, Uyghur, Kazak, Kirgiz, Yi script Tai Lue script, Korean, Russian, Zhuang script and so on.

Introduction and Related works

It is generally thought that Mustonen (1965) [1] proposed to identify different texts according to the characters of various languages is the beginning of text identification. The early research mainly relied on language rules, with the development of computer, methods of natural language identification changed from analyzing language rules to statistical analysis. Cavnar (1994) [2] presented N-gram text automatic identification, which is classical method based on statistical analysis. Cavnar used N-gram to test 3478 texts in 8 languages, the rate of correct identification reached 99.8%. The same year, Dunning reached 99.9% combining markov model and N-gram. After that, some scholar applied statistics algorithm such as relative entropy [3] and SVM [4] to text identification, and used skills like smoothing technique to make the identification rate reached 99.998% [5]. The increasing of webpages of various language texts appealed scholars to research multi-text identification skills between different language families or the same language family[6][7][8][9], the correct identification rate is on the rise.

In China, we use the method combing rules and statistics to identify Tibetan [10] [11], Mongolian [12] and Uyghur texts [13] [14], correct identification rate can reach 100%, 80% and 97%. But research on other minority language texts' identification is less.

Features of minority websites identification

Present problems. Compared with Chinese and English characters, minority language characters have obvious features. Some minority languages have various characters of the same language by the influence of history. The computer skills of dealing with minority language character are fall behind. Half of the minority websites are folk websites, whose source code is not standard, some minority language characters' encoding is not unified, which makes the encoding of the same language not compatible. Present problems of automatic identification mainly come from features of minority language itself and immaturity of supporting technology.

Features of minority websites. (1) The same language has different characters. Such as Mongolian (Traditional Mongolian, Tod Mongolian, New Mongolian), Uyghur (Arabic character, Latin character), Kazak (Arabic character, Latin character, Kirill character), Tai Lue (New Tai Lue, Old Tai Lue).

(2)The same character has different encoding. Tibetan and traditional Mongolian have the most encoding forms. Tibetan has Unicode, Founder, Ascll (11 forms), HuaGuang, Tibetan University,

Tonguer, Pandita and so on. Traditional Mongolian has Unicode, Menk, Hussein, Fonder, Minggatu, Oyuta, Burigude and so on.

(3)The same character with different encoding has cross and overlap region. In Tibetan, part of GB2312 encoding has cross-field; In Mongolian, part of Unicode encoding has cross-field.

Table1. Part of GB2312-based Tibetan encoding

encoding	First byte scope	Trail byte	Syllable point encoding
Founder DOS	0xC0-0xEE	0x21-0x7E	0xC032
Founder Windows	0xAA0xAC,0xB0-0xDE	0xA0-0xFE	0xAAAC
HuaGuang DOS	0xB0-0xFB	0x21-0x7E	0xE162
HuaGuang Windows	0xB0-0xFB	0xA1-0xFE	0xE1E2
Tonguer encoding	0x81-0xEE,0xF5	0x210x7E,0x40-0xFE	0xA6E6
Tibetan University encoding	0xAA-0xAF,0xF8-0xFB	0xA1-0xFE	0xFABB

Table2. Part of Unicode-based Mongolian encoding

encoding	Encoding scope
Oyuta	0xE250-0xE377
Burigude	0xE246-0xE29F
Menk	0xE264-0xE34F
Hussein	0xE246-0xE355
Minggatu	0xE254-0xE33E

(4)Writing of source code in webpage is not standard. Some of the identification of character set of meta in minority webpage is optional, like" charset" and "encoding", which makes great inconvenience to correct decoding, picture 1 is an example of messy code page after decoding of http://www.nmqnw.cn/mgl/

Fig.1 Messy code webpage example

Identification

There are 10 on-line minority language text websites : Mongolian, Tibetan, Uyghur, Kazak, Kirgiz, Yi script ,Tai Lue script, Korean, Russian, Zhuang script. This paper chooses the most widely used character of each language as research object, that is traditional Mongolian, Tibetan, Arabic character of Uyghur, Kazak, Kirgiz, Yi script, New Tai Lue script, Korean, Russian, and Zhuang scrip.

Research shows there are three approaches with high accuracy of identification of minority language text after decoding the source code of webpage correctly.

(1) feature character –based approach

Feature character is a character set which will not appear in other language text or can distinguish a certain kind of text. For example, the syllable point and droop symbol in Tibetan will not appear in other languages generally. Meanwhile, these two feature character take a high percentage in Tibetan. Thus we can identify Tibetan by feature character-based approach.

Certain text character set which lies in certain encoding section with monopoly Unicode encoding scope can be also handled feature character, for example: Tibetan with Unicode encoding, traditional Mongolian, Korean, Yi script, Tai Lue script and so on can be identified by judging the Unicode encoding scope of character set.

(2) label attribute of webpage-based approach

Generally speaking, the attribute which marked “encoding”, ”charest”, ”font-family” in source code of HTML webpage will decide the encoding type of a certain webpage. For minority language texts which are not in the scope of Unicode encoding and have different encodings, especially Mongolian and Tibetan, the attribute of “font-family” in source code of webpage can generally indicate the texts in the webpage. In China, the common font-type of “font-family” in Tibetan webpage are: BZDBT, BZDMT, BZDHT, TIBETBT, TIBETFG, TIBETCT, TIBETZT, TIBETHT and so on. In Mongolian webpage, the common font-type of “font-family” are: SYMN2008, Sy2008, symn2008f, HuderUI-Saiyin, Saiyinwebcagantig, Menksofet2012, Menksoft2007, MenksoftQagan, Menksoft2013regular, MENKSOF0, MenksoftQagan_mirror, Huritai, MGT-MHWT-OT and so on. Some Mongolian webpages that contains < meta name="generator" content="MenkCms Portal-http://www.menksoft.com"> can also indicate it is Mongolian webpage. Making full use of the information that indicates text identity in the webpage can rapidly assist to identify the language and text in the webpage.

(3)N-gram-based approach

In computational linguistics and probability theory, N-gram is continuous array that contains N minimum partition unit in given test or speech array. Minimum partition unit can be phoneme, syllable, letter, character and so on. N-gram is classical in automatic identification of textual text; it has good effect on those texts can be identified neither by feature character-based approach nor label attribute-based approach. For example, the characters in Uyghur, Kazak, Kirgiz texts are mostly similar, feature character cannot identify them, but N-gram can realize it.

Compositing the advantages and disadvantages of above approaches, this paper descries the steps of identifying minority webpage text:

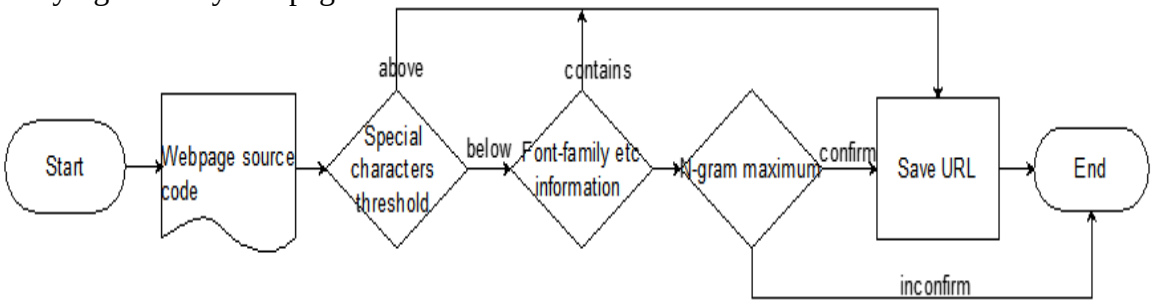

Fig. 2. Flow chart

Step 1. Count the frequency of feature character in the webpage, if the frequency reach threshold vale of a certain language word, this webpage can be identified as this language and store this URL, if not, go into step 2.

Step 2. Test information like META of source code in webpage, if find information like “font-family” that can indicate a certain of language word, this webpage can be identified as this language and store this URL, if not, go into step3.

Step 3. Identify by Bigram approach, if the maximum result of identification is this language text, this webpage identification can be recognized as successful and store this URL, and the process of identification finish.

Conclusion

Using the feature character-based approach, webpage label attribute and N-gram based approach to identify traditional mongolian, Tibetan Uyghur, Kazak, Kirgiz, Yi script Tai Lue script, Korean, Russian, Zhuang script 10 minority language textual word, can realize comparatively ideal result without manual intervention. So far, the average correct recognition rate can reach above 95%, before using in-service still needs to promote correct recognition rate. The following work: firstly, adding more identification connector of language based on promoting correct recognition rate, getting ready for other minority language which have not been appeared in internet; secondly, promoting the recall ratio in collection of minority language websites.

Acknowledgement

This paper is supported by the Innovative Research Project of MinZu University of China. The project title is "Extraction and classification of Kazak configuration suffixes based on grammar rules (Project No.K2014012) ".

References

[1]Mustonen S. Multiple discriminant analysis in linguistic problems [J]. Statistical Methods in Linguistics. 1965, 4: 37-44.

[2]Cavnar W B, Trenkle J M. N-gram-based text categorization [J]. Ann Arbor MI. 1994, 48113(2): 161-175.

[3]Sibun P, Reynar J C. Language identification: Examining the issues [J]. 1996.

[4]Kruengkrai C, Srichaivattana P, Sornlertlamvanich V, et al. Language identification based on string kernels[C]. IEEE, 2005.

[5]Brown R D. Finding and identifying text in 900+ languages [J]. Digital Investigation. 2012, 9: S34-S43.

[6]Yamaguchi H, Tanaka-Ishii K. Text segmentation by language using minimum description length[C]. Association for Computational Linguistics, 2012.

[7]Chew Y C, Mikami Y, Nagano R L. Language Identification of Web Pages Based on Improved N-gram Algorithm [J]. International Journal of Computer Science Issues (IJCSI). 2011, 8(3): 47-58.

[8]King B, Abney S. Labeling the languages of words in mixed-language documents using weakly supervised methods[C]. 2013.

[9]Lui M, Lau J H, Baldwin T. Automatic Detection and Language Identification of Multilingual Documents [J]. 2014.

[10]Wang Sili. Automatic finding and collection of Tibetan websites [D].MinZu University of Northwest, 2010.

[11]Identification of Tibetan website and encoding [Z].Google Patents, 2007.

[12]Research on Mongolian website crawling and encoding identification transfer [D]. Mongolian University, 2008

[13]Method and system of automatic identification of Uyghur websites [Z].Google Patents, 2010

[14]Mairidan· Wushouer, Weinila· Mushajiang. Research on automatic recognition of Uyghur, Kazak, Kirgiz in e-dictionary software [J]. XinJiang University Journal (science).2011(01):88-92

[15]Jinliang, Sandanma, Yuying. Encoding of traditional Mongolian and analysis of application [J]. Chinese Journal. 2012(07), 16-17.