Advanced Science and Technology Letters
Vol.63 (NGCIT 2014), pp.134-138
http://dx.doi.org/10.14257/astl.2014.63.28

How to calculate the Korean name candidate using the phonetic similar code

Byung-Hui Jeong[1], Jin-Tak Choi[2,1], Reeyan Kyoo-Ho Lee [3], Dong-Ha Park [3],
Young-Wook Yoon[3]

[1]BRC Co., Ltd. 203-3 Songdo-dong, Yeonssu-gu, Incheon, Republic of Korea,
buxbany@gmail.com
[2] Incheon University, 119 Academy-ro, Yeonsu-gu, Incheon, Republic of Korea
choi@incheon.ac.kr
[3] GE Healthcare IT Korea Technology Center. Meet you all Tower 4th floor, 7-50 Songdo-dong, Yeonssu-gu, Incheon, Republic of Korea
{ Reeyan.Lee, DongHa.Park, YOUNG_WOOK.YOON }@ge.com

Abstract. MPI system being used diversely abroad is also using the method that distinguishes individuals in combination with the private information. The system to judge the combination of private information quickly and exactly is necessary in this system. In this paper, Suggest the phonetic similar code algorithm of Korean name data for the combination of private information and the algorithms that can extract correct the candidate group by using Korean name features.

Keywords: Korean, name, phonetic,

1 Introduction

The spill of private information is emerging as a social issue all around the world. It is one solution not to save the sensitive information as possible to solve this problem. Korean government initiated a law of prohibiting the resident registration number storage and is going to enact soon to solve this problem. In accordance with this flow, the hospitals are also introducing the system by which we can distinguish patients' information without saving the sensitive private information in the hospital information system.

MPI system being used diversely abroad is also using the method that distinguishes individuals in combination with the private information. The system to judge the combination of private information quickly and exactly is necessary in this system.

In this paper, Suggest the phonetic similar code algorithm of Korean name data for the combination of private information and the algorithms that can extract correct the candidate group.

[1] Corresponding author.

ISSN: 2287-1233 ASTL

2 Past Work

As the method to search all names when searching the large-capacity names is inefficient, the method to extract the candidate group is needed.

The phonetic algorithm method exists in this way and the representative method is Soundex algorithm. The remaining consonants except for the first consonant after leaving only consonants after removing vowels in English words are changed into the code value using the similar consonant group suggested by the algorithm.

You compare the codes that came out in this way and judge if they are same each other, they are similar and if they are different each other, they are not similar. There is a method that used Soundex method to Korean[1].

It is the group that was suggested after leaving the first consonant after dividing syllables into consonant/vowel phonemes and removing the vowels. It is the method to change into the code value. However, there is a method that sets up the condition in each phoneme and creates the detailed codes rather than simply using the code of group because the accuracy of this method becomes the problem[2][3][4].

However, since this study is conducted focusing on Korean search of loanwords and the name characteristic is not reflected, it is impossible to extract the exact candidate group in the name candidate extraction.

3 Our Approach

This paper reorganized the groups usually used in Korea name using the Koreans' name nature in order to solve the problems that we cannot extract the exact candidate group when extracting Koreans' Hangul name candidates mentioned before, and in case of usually used syllables, exceptional syllables were made to have them included in several groups.

3.1 Algorithm

If we give the similar code only by Korean consonants, the discrimination capacity in Korean names that have relatively short length and less phonetic confusion so the similar vowel information was grouped to conduct the encoding.

The syllables were divided into phonemes and all of the final consonants were removed and then the code was created only using the initial sound and final consonants.

3.2 Consonant/vowel group

As Table 1, the part usually used in Korean rules and Korean names were grouped to create the classification code. The vowel was composed for the bit calculation.

Table 1. Consonant(initial sound) / vowel classifying code.

Consonant group	Code(Hex)	Vowel group	Code(Hex)
ㅁ,ㅂ,ㅍ,ㅃ	0x9	ㅐ,ㅒ,ㅔ,ㅖ,ㅚ	0x01
ㄷ,ㄸ,ㅌ	0xA	ㅟ,ㅙ,ㅞ	0x02
ㅅ,ㅆ	0xB	ㅓ,ㅕ,ㅝ	0x04
ㅈ,ㅉ,ㅊ	0xC	ㅏ,ㅑ,ㅘ	0x08
ㄱ,ㄲ,ㅋ	0xD	ㅗ,ㅛ,ㅜ,ㅠ,ㅡ,ㅢ,ㅣ	0x10
ㄴ,ㄹ,ㅇ,ㅎ	0xE		

3.2 Exceptional syllable group

Usually used syllables like Table 2 can raise the accuracy when being included in several groups rather than in only one group so the vowel code was set up as FF to be compared with candidates of all vowel groups through the bit calculation.

Table 2. Exceptional group classification.

Consonant group	Exceptional group
0x9	'봉','붕','병','문','만','매','미','목','묵','범','봄','복','맹'
0xA	'덕', '득'
0xB	'성', '송', '승', '선', '숭', '순', '상', 손', '숙', '석','수','서','소'
0xC	'정', '종', '중', '장', '전', '점', '철', '칠'
0xD	'경', '걸','길','굴','갈', '귀', '구','기','그','건'
0xE	'용','영','형','녕','령','원','율','률','열','렬','윤','연'

4 Experiments

It was judged how much the accuracy was improved based on the number of similarity judgment of similar names and the similar judgment error rate of non-similar names through the code based on 1010 pairs of similar names and 300 thousand pairs of non-similar names the user generated manually, and it was judged how much the candidate group was evenly distributed through the encoding of 1.5 million cases of names.

KODEX algorithm is used as the comparative algorithm which is not specialized in regard to the loanword.

The following Table 3 shows the similarity judgment result on the similar names and non-similar names using KODEX.

Table 3. Experiment results.

KODEX			Our Approach		
similarity judgment	similar names (1,010 Pair)	non-similar names (3million pair)	similarity judgment	similar names (1,010 Pair)	non-similar names (3million pair)
Simmilar	396	507	Simmilar	986	5,520
Non-Simmilar	614	299,493	Non-Simmilar	24	294,480
accuracy		99.6%	accuracy		98.2%
responsiveness		39.2%	responsiveness		97.6%
Specificity		99.8%	Specificity		98.2%
15 million name convert result			**15 million name convert result**		
Total code count		16,881	Total code count		8,132
Max Same Count		19,944	Max Same Count		28,698
Less than 2 Same Count		6,015	Less than 2 Same Count		2,555
Average of Count per Code		88.61	Average of Count per Code		183.96

5 Results

As a result of experiment, while the rate of extracting similar name pairs as the candidate was about 39.2% in case of KODEX, in the suggested method, the similar name pair was extracted as 97.6% candidate.

On the other hand, while the accuracy dropped from 99.6% to 98.2% in case of non-similar name, in the candidate extraction for the name, it was the more important element that how many similar candidate groups were extracted than the error value so the exact candidate extraction was improved.

Considering that the average number of suggested algorithm per code was 183.96 compared to KODEX and the number of codes under 2 reduced to 2555, we can say the candidates were evenly extracted.

Through it, the candidate calculating method of Korean name that used the phonetic code was improved in the similar code extraction compared to existing phonetic method. Currently, the algorithm suggested in this paper will be loaded in the actual MPI system to be improved more based on the experience.

References

1. Kang Byung-Ju., Lee Jaeseong., Choi Key-Sun.: Phonetic Similarity Meausre for the Korean Transliterations of Foreign Words. Journal of KISS (b):software and applications. B / v.26 no.10, 1999, pp.1237-1246 (1999)
2. Park Jong Hyeok.: An Enhanced Algorithm for Equivalent Foreign Word Transliteration Detection, (2004)
3. Hyo Kyong Kim.: A Phonetic Matching Algorithm For Korean Spell Checker. (2006)

4. Sook Hyeon Ko, Jae Sung Lee.: An Enhanced Context Sensitive Algorithm for Equivalent Foreign Word Transliteration Detection. Proceedings of the 19th Conference of Hangul and Korean Information Processing, pp.114-121 (2007)