Advanced Science and Technology Letters
Vol.63 (NGCIT 2014), pp.139-142
http://dx.doi.org/10.14257/astl.2014.63.29

Study of the private information similarity and the distinguishing module development Improvement of Korean name string similarity algorithm

Byung-Hui Jeong[1], Jin-Tak Choi[2,1], Reeyan KyooHo Lee[1], Dong Ha Park[1], Chang Il Kim[1],

[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, ChangIlKim }@ge.com

Abstract. For patient information identification, demographic fields (name and address), and medical information (medical history) are compared in order to measure the similarities. The improvement of algorithm related to Koreans' Hangul name is necessary in the individual distinguishing part in order to introduce MPI system matching Korea. Since Hangul is a type of combination, we cannot match the names more exactly by the one-dimensional distance comparative function which was used in existing English. This paper suggests the algorithm for the Korean name comparison.

Keywords: Korean, name, similarity,

1 Introduction

For patient information identification, demographic fields (name and address), and medical information (medical history) are compared in order to measure the similarities.

The improvement of algorithm related to Koreans' Hangul name is necessary in the individual distinguishing part in order to introduce MPI system matching Korea. Since Hangul is a type of combination, we cannot match the names more exactly by the one-dimensional distance comparative function which was used in existing English.

This paper suggests the algorithm for the Korean name comparison.

[1] *Corresponding author.*

ISSN: 2287-1233 ASTL

2 Past Work

There is the editi distance algorithm of English alphabet as the most basic algorithm. It is the algorithm that judges the similarity of two words depending on how many times the insertion, deletion, and substitution happen which are necessary for 2 words to become same each other.

GrpSim[2] algorithm is a method to distinguish the part in more detail which always has the same distance in case of the substitution of phoneme at KorED. It generated the groups following the features of Hangul such as bilabial, alveolar sound, palatal consonant, velar, and glottal sound, and let them have a higher similarity value when two phonemes were in the same group when comparing them and improved the similarity judgment between words. It was the similarity method not existing edit distance method and used the method that as the similarity was high, the probability of being same words was high. OneDSim, OneDSim2[3] were a method that improved the problem that the similarity dropped when the consonant and vowel are separated like '가마ㄴ히' in GrpSim above. It suggested OneDSim that used the method of putting the distinguishing sign between syllables and OneDSim2 that used the method of giving the weighted value to the first phoneme. It accomplished the improvement in distinguishing the vulgarism through the suggested algorithm..

3 Our Approach

This paper suggests 3 Hangul name similarity algorithms of Koreans.

First, the family name used in Korea is judged using the Koreans' family name information registered in the Statistics Korea as the Table 1, and the penalty value is given to unregistered family names to raise the discrimination capacity.

Second, similar phoneme pair was made based on the location of keyboard, Hangul's feature such as initial law, and the feature of name, and if two phonemes were in the similar phoneme pair although they did not match up, the higher score was given.

Table 2. Medial (vowel) similar phoneme pair table.

phoneme Pair	weighted value	phoneme Pair	weighted value	phoneme Pair	weighted value
(ㅁ,ㅂ)	Nb1	(ㄱ,ㅅ)	Nb3	(ㄴ,ㅁ)	Nb3
(ㅁ,ㅍ)	Nb1	(ㄱ,ㄷ)	Nb3	(ㄷ,ㅇ)	Nb3
(ㅂ,ㅍ)	Nb2	(ㄱ,ㄹ)	Nb3	(ㄹ,ㅎ)	Nb3
(ㄷ,ㅌ)	Nb1	(ㄱ,ㅇ)	Nb2	(ㅅ,ㅎ)	Nb3
(ㅅ,ㅈ)	Nb2	(ㄴ,ㅇ)	Nb1	(ㅈ,ㄷ)	Nb3
(ㅈ,ㅊ)	Nb1	(ㄹ,ㅇ)	Nb1	(ㅊ,ㅌ)	Nb3
(ㄴ,ㄹ)	Nb1	(ㅇ,ㅎ)	Nb1		

Table 3. Initial sound (consonant) similar phoneme pair table.

phoneme Pair	weighted value	phoneme Pair	weighted value	phoneme Pair	weighted value
(ㅐ,ㅖ)	Nb1	(ㅜ,ㅠ)	Nb1	(ㅕ,ㅡ)	Nb2
(ㅐ,ㅒ)	Nb2	(ㅡ,ㅣ)	Nb2	(ㅜ,ㅡ)	Nb2
(ㅐ,ㅚ)	Nb1	(ㅚ,ㅣ)	Nb1	(ㅗ,ㅡ)	Nb2
(ㅖ,ㅒ)	Nb2	(ㅚ,ㅡ)	Nb1	(ㅜ,ㅡ)	Nb2
(ㅖ,ㅚ)	Nb3	(ㅜㅣ,ㅣ)	Nb1	(ㅏ,ㅑ)	Nb1
(ㅒ,ㅚ)	Nb3	(ㅜㅣ,ㅜ)	Nb3	(ㅗ,ㅛ)	Nb1
(ㅜㅣ,ㅚ)	Nb2	(ㅝ,ㅕ)	Nb1	(ㅘ,ㅏ)	Nb1
(ㅕ,ㅒ)	Nb1	(ㅝ,ㅜ)	Nb3	(ㅘ,ㅗ)	Nb3

Third, the low frequency alternative phoneme Table 4 and Table 5 was provided that could replace the family name, name, and hardly used phonemes that were created based on Hangeul name data of 8122 Seoul public servants, and when the low frequency phoneme was compared and the opposite phoneme was included in the alternative phoneme, the higher similarity than when they did not match up was given.

Table 4. Low frequency alternative phoneme table for family name.

Vowel (6)		**Consonant (4)**		**Final consonants (22)**							
ㅋ	ㄱㅌㅁ	ㅒ	ㅐㅖㅔㅣ	ㄲ	ㄱ	ㄻ	ㄹㅁ	ㅂ	ㅁ	ㅋ	ㄱㅁ
ㄲ	ㄱㄷㅅㄹ	ㅙ	ㅚㅗㅐ	ㄳ	ㄱ	ㄾ	ㄹ	ㅄ	ㅁ	ㅌ	ㄴㅇ
ㄸ	ㄷㅈㄱㅇ	ㅞ	ㅚㅜㅔ	ㄵ	ㄴ	ㄺ	ㄹ	ㅅ	ㄱㄹ	ㅍ	ㄹㅁ
ㅃ	ㅂㅁ	ㅟ	ㅡㅣ	ㄶ	ㄴ	ㄿ	ㄹ	ㅆ	ㄱㄹ	ㅎ	ㄹㅇ
ㅆ	ㅅㄱㅎ			ㄷ	ㄱㅇ	ㅀ	ㄹ	ㅈ	ㄴㅁ		
ㅉ	ㅈㅂㄷㄴ			ㄹㄱ	ㄱㄹ	ㄾ	ㄹ	ㅊ	ㄹㅇ		

Table 5. Low frequency alternative phoneme table for name.

Vowel(6)		**Consonant (3)**		**Final consonants (19)**							
ㅋ	ㄱㅌㅁ	ㅒ	ㅐㅖㅔㅣ	ㄲ	ㄱㅅ	ㄻ	ㄹㅁ	ㅄ	ㅂㅅ	ㅎ	ㄹㅇ
ㄲ	ㄱㄷㅅㄹ	ㅙ	ㅚㅗㅐ	ㄳ	ㄱㅅ	ㄾ	ㄹㅂ	ㅆ	ㅅㄱ		
ㄸ	ㄷㅈㄱㅇ	ㅞ	ㅚㅜㅔ	ㄵ	ㄴ	ㄺ	ㄹㅅ	ㅈ	ㄴㅂ		
ㅃ	ㅂㅁ			ㄶ	ㄴ	ㄿ	ㄹ	ㅋ	ㄱㅁ		
ㅆ	ㅅㄱㅎ			ㄷ	ㄱㅇ	ㅀ	ㄹ	ㅌ	ㄴㅇ		
ㅉ	ㅈㅂㄷㄴ			ㄹㄱ	ㄱㄹ	ㄾ	ㄹ	ㅍ	ㄹㅁ		

This algorithm calculates similarity by recurrence formula in [3].

4 Experiment

When judging the similar name by fixing 1,000 similar name data as 96%, the degree improved in the existing algorithm was judged through the error rate that judged 300,000 non-similar data as being similar. As a result of experiment, the error rate judging the non-similar data as similar data was 5% in the comparative algorithm, 2.7% in the suggested algorithm compared to the existing algorithm.

Table 7. Experiment result.

	GrpSim	OneDSim2	BRCSim
Similar name Judging Rate	96%		
Error rate	6.2%	5.3%	2.7%

5 Results

In this paper, the name similarity algorithm was suggested by adding the specialized algorithm on Hangul name of Korean based on OneDSim2 and GrpSim that were the algorithm for distinguishing the loanwords and vulgarism. It was confirmed through the experiment that the algorithm suggested in this paper showed a better performance in distinguishing the name than the existing algorithm.

This algorithm is going to be applied to MPI system in Korea and will be used to judge Koreans name similarity, and through this, the Hangul name similarity-measuring algorithms of various fields will be improved continuously.

References

1. Kang Ho Roh.: Edit Distance and Similarity Algorithms for the Korean Alphabet. (2010)
2. Kangho Roh, Kunsoo Park, Hwan-Gue Cho, Sowon Chang, Edit Distance Problem for the Korean Alphabet with the Phoneme Classification System, Journal of KIISE: Computer Systems and Theory, Vol.37, No.6, pp.323-329, (2010)
3. Kang Ho Roh, Kun Soo Park, Hwan Gue Cho, So Won Chang.: Similarity and Edit Distance Algorithms for the Korean Alphabet using One-Dimensional Array of Phonemes. Journal of KIISE : Computing Pracices and Letters, Vol 17, No 10, pp.519-526 (2011)

