International Journal of Software Engineering and Its Applications
Vol. 9, No. 1 (2015), pp. 87-94
http://dx.doi.org/10.14257/ijseia.2015.9.1.07

Efficient Calculation of Korean name Candidate Using the Phonetic Similar Code

Geun-Young Park[1], Byung-Hui Jeong[2], Jin-Tak Cho[3] and Eun-Young Jung[4]

[1]*Incheon University, 119 Academy-ro, Yeonsu-gu, Incheon, Republic of Korea*
[2]*BRC Co., Ltd. 203-3 Songdo-dong, Yeonssu-gu, Incheon, Republic of Korea,*
[3]*Incheon University, 119 Academy-ro, Yeonsu-gu, Incheon, Republic of Korea*
[4]*Gachon University Gil Medical Center, 1198 Guwol Dong, Namdong-gu, Incheon, Republic of Korea*
[1]*hains8285@hanmail.net,* [2]*buxbany@gmail.com,* [3]*choi@incheon.ac.kr,* [4]*eyjung@gilhospital.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 [1] 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.

2. Related 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.

Our approach is based on phonetic matching algorithms such as blow we describe these algorithms in detail.

The phonetic algorithm method exists in this way and the representative method is Soundex algorithm [2]. 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.

The Metaphone [3] algorithm based on the Soundex algorithm is a one-byte character system designed to compensate for the shortcomings of Soundex, which provides a high

ISSN: 1738-9984 IJSEIA

recall factor but has low accuracy. It is identical to Soundex in its mechanism of deleting vowels (A, E, I, O, and U) from each word, except for the initial vowel, but it does not delete them if a vowel is repeated, with the exception of a posteriorly appearing vowel; and with this, the substitute condition for the initial character is maintained and processed.

The Kodex [4] algorithm assigns an identical code number to similar words by categorizing them according to their Korean consonant, on top of the basic Soundex algorithm, in an attempt to resolve the confusion in the designation of loan words. The Kodex algorithm shows excellent performance in its search speed and recall factor because it was developed on the principle of simplicity. It also has some shortcomings, though: its low accuracy and its substitution of consonants with no supplementary information on vowels with no exception, thereby generating many words with identical code values. This algorithm is aimed at performing Korean designation of loan words.

The Ekodex [5] algorithm succeeded in improving the accuracy of the Kodex algorithm by making words regular according to the usage of each Korean word and language habit, and by using the vowel information on the initial and final consonant, but it limits the type of searchable loan words within the distance of two bytes between the query word and the search word. It has other shortcomings: indifferent substitution with representative vowels, and lack of detailed sub-rules on the phonetic value of the "ㅇ" vowel in the middle of a word.

The Ckodex [6] algorithm improved the accuracy of the Ekodex algorithm by applying detailed sub-rules depending on the context based on the Ekodex algorithm. However, it does not apply any rule on the initial law or exceptional rule, because it was developed to extract similar words from the general words. As the accuracy of the Ckodex algorithm can be improved by processing detailed sub-rules per initial consonant, middle vowel, and final consonant of Korean text based on nine basic rules, general performance improvements in processing similar words may be expected, but this complex algorithm requires much time to process such words because it is aimed at more quickly extracting candidates for the name similarity comparison.

Most of phonetic matching algorithms for Hangul focus on match Korean transliterations of foreign words [7-9].

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.

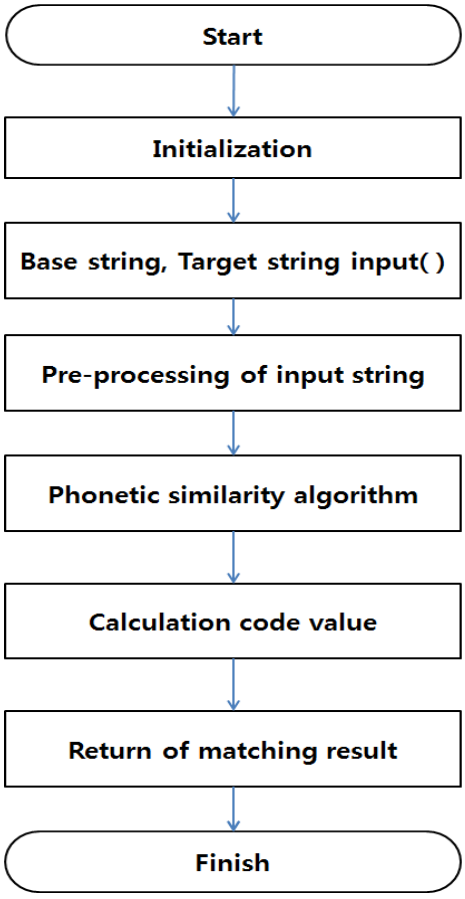


Figure 1. Whole Process of Proposed Method

The overall process stream minimizes some potential problems with the inputted data set through preliminary processing of the input stream. The code values are calculated using the phonetic similarity conversion algorithm, which generates codes based on the consonant/vowel classification codes and the exceptional group classification. Then the candidate words are extracted through a matching process using the calculated code values.

Since assigning similar codes based purely on Korean consonants may undermine the capacity to discern relatively shorter Korean names and may pose no significant phonetic confusion, they should be coded by grouping together the vowel groups with similar information. Furthermore, code values must be assigned by combining consonants (initial consonants) and vowels using the probabilistic traits of typos and their phonetic characteristics, such as the initial law in the names. A more detailed explanation of the aforementioned grouping method follows.

To accelerate the code generation calculation speed, the algorithm allowed the generation of those codes capable of performing bit operations on the vowel codes. Whether it is TRUE or NOT is determined by performing the AND operation between the code values of the inputted characters and those of the base characters.

The exceptional group is formed by grouping together exceptional cases that cannot be classified into any group. Therefore, an FF classification code is assigned to them so that they would be included in all vowels in the calculation.

The algorithm prevented the potential occurrence of a “negative false” by applying another method, in which two codes are generated when the initial consonant and the final consonant change places. The mechanism behind this shall be further explained in the next paragraph.

Any overlap between the two codes generated as such is determined through a bit operation on them. If they are determined to be overlapping codes, they are extracted as candidates.

In the following paragraph, the method of creating groups and the details of the algorithm required to perform the aforementioned algorithms are explained.

3.1. Normalization of the Input Data

The process of normalizing the input data into the predefined complete Korean text format is as follows. The inputted Korean text strings are normalized according to the following rules to turn them into the complete Korean text format.

Field	Criteria	Rules
Name	Blank space	Delete the blank space within strings
	Exceptional characters	Delete the code values of other characters (special characters, Roman characters, and numbers), excluding Korean characters
	Omission of consonants/vowels	Cluster those connected non-complete phonemes and recompose them in the order of C-V-C
		In case a consonant is dropped and there is no final consonant in the preceding character, attach it as a final consonant to the preceding character.

3.2 Separation of Input Strings

To apply the rules suggested for the inputted strings, a process of dividing the strings into surnames and first names, and another process of dividing the syllables into phonemes for the separated character strings, are required. As the syllables in Korean text are constructed formulaically, they can be divided into phonemes using appropriate formulas.

3.2.1. How to Separate Consonants from Vowels in Korean Text

As Korean letters are 2-Byte combination characters, a total of 11,172 individual characters can be composed with Korean letters, which lessens the overall capability to discern among them. To resolve this dilemma, basic consonants must be separated from vowels in each Korean character and matched so that 1-byte characters may be used liberally in the similarity algorithm. The matching process is as follows.

The Korean characters are composed of initial consonants, middle vowels, and final consonants, and the unicodes used to express 11,172 Korean characters fall in the range of 0xAC00~0xD7A. The initial consonants, middle vowels, and final consonants can be separated or combined in accordance with certain formulas that were formed based on the unicodes with their methods, as follows.

- Initial consonants: (Unicode of the inputted Korean character – 0xAC00) / (28 x 21)
- Medium vowels: [(Korean unicode – 0xAC00) / (28 x 21)] / 28
- Final consonants: (Korean unicode – 0xAC00) / 28

Indices of initial consonants, medium vowels, and final consonants can be created using the aforementioned calculation methods, which makes it possible to separate them all.

3.3 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. This group was made based on the location of keyboard [10], Hangul's feature such as initial law [11], and the feature of name.

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		

3.4 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. We extract this group by experimental approach by using our name dataset which is Seoul government employee name list.

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'

3.5 Exchange of Initial Consonant and Final Consonant

Potential occurrences of "negative false" are sometimes observed when the aforementioned method is used. For instance, 홍길동(Hong Gil Dong) and 홍길몽(Hong Gil Mong) are very similar, but they cannot be included in the same group through comparison of their phonetically similar codes. If the difference in a phoneme but not in a syllable makes it impossible to extract such potential candidates, it would be difficult to expect some desirable outcomes. To tackle such dilemma, another method is proposed in which two codes are generated for each name.

In the proposed method, codes are generated according to the earlier-suggested algorithms, but the first code shall be processed according to the existing method, whereas the second code shall be generated using the code generation algorithm in which the initial and final consonants of each character consisting of an individual name are interchanged.

There are some additional rules on the codes generated after the exchange is completed: if there is no initial consonant, 0 is assigned as the consonant code to represent the lack of a final consonant.

It is possible to resolve this dilemma using this method, in which 홍길동(Hong Gil Dong) and 홍길몽(Hong Gil Mong) are not extracted as candidates due to their different codes.

4. Experiment

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.

Acknowledgements

This paper is a revised and expanded version of a paper entitled [How to calculate the Korean name candidate using the phonetic similar code] presented at [The 3rd International Conference on Next Generation Computer and Information Technology, October 24-26, 2014 at Liberty Central Saigon Hotel, Hochimin, Vietnam,]

This study was supported by the Senior-friendly Product R&D program funded by the Ministry of Health & Welfare through the Korea Health Industry Development Institute (KHIDI). (HI14C1435).

References

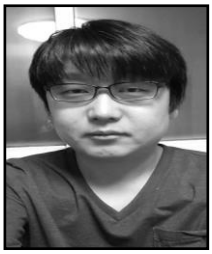
[1] R. Johnson, K. Schurenberg and R. Yeager, "System and method for implementing a global master patient index", **(2000)**.
[2] R. Russell and M. Odell, "Soundex", US Patent 1, **(1918)**.
[3] P. Lawrence, "Hanging on the metaphone", Computer Language, vol. 7, no. 12, **(1990)**, pp. 39-43.
[4] K. Byung-Ju, L. Jaeseong and C. Key-Sun, "Phonetic Similarity Meausre for the Korean Transliterations of Foreign Words", Journal of KISS (b):software and applications. B /, vol. 26, no. 10, **(1999)**, pp. 1237-1246.
[5] P. Jong Hyeok, "An Enhanced Algorithm for Equivalent Foreign Word Transliteration Detection", **(2004)**.
[6] S. Hyeon Ko and J. Sung Lee, "An Enhanced Context Sensitive Algorithm for Equivalent Foreign Word Transliteration Detection", Proceedings of the 19th Conference of Hangul and Korean Information Processing, **(2007)**, pp. 114-121.
[7] J.-H. Oh, S.-M. Bae and K.-S. Choi, "An Algorithm for Extracting English-Korean Transliteration pairs using Automatic E-K Transliteration", KIISE Spring Conference, **(2004)**.
[8] J.-S. Kim, K.-H. Kim and J.-H. Lee, "Retrieving English Words with a Spoken Word Transliteration", Journal of the Korean Society for Library and Information Science, vol. 39, no. 3, **(2005)**, pp. 93-103.
[9] S. Young Jung, S. Lim Hong and E. Paek, "An English to Korean transliteration model of extended Markov window", Proceedings of the 18th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, **(2000)**.
[10] S. Chang, S.-k. Kim and S.-C. Jung, "This slip of the tongue that slip of the pen: official documents", Ministry of Culture and Tourism, **(2000)**.
[11] I. Lee and S. Robert Ramsey, "The Korean Language", Suny Press, **(2000)**.

Authors

Geun-Young Park, received her M.S degree in Computer Science from Incheon National University, Korea.

Since 2013, She is currently working toward a Ph..D. in Computer Science and Engineering of Incheon National University.

Research interests: image processing, database



Byung-Hui Jeong

2010: M.S. degree in Information and Telecommunication Engineering, Incheon National University, Korea.

2014: Ph.D. degree of Computer Science and Engineering from Incheon National University, Korea.

2011~: manager of ICT research department, BRC Institute Korea.

Research Interests: Artificial Intelligence, Software Architecture, Database design, u-healthcare

Jin-Tak Choi
1991: Ph.D. degree in KyungHee University
1992: Exchanging Prof. of University of Pennsylvania
1987~: Professor of National University of Incheon
Research interests: Database, Computational statistics.

Eun-Young Jung, she received M.S. degree in Health Informatics, Gachon University, Korea, in 2001. She received Ph.D. degree of Medical Informatics from Ajou University, Korea, in 2012. She is currently manager of U-Healthcare Center, Gachon University Gil Medical Center, Korea. Her research interests include u-healthcare, Telecare and Health IT, Virtual Reality simulation.