Industry Experience: Chinese Names Duplicate Records Detection

Thong Tong Khin and Badrul Affandy Ahmad
Software Development Lab,
Mimos Bhd,
Kuala Lumpur, Malaysia.
e-mail:{thong.tkin, badrul.affandy}@mimos.my

Wang Xiaomei
Xiamen University Malaysia,
Sepang, Malaysia.
e-mail: xmwang@xmu.edu.my

Kandiah Arichandran
Kuala Lumpur, Malaysia.
e-mail: arichandran@gmail.com

Abstract— **The Soundex method is the preferred method for duplicate detection process on Malaysian Chinese names. The names are written in English text, but are phonetically translated from various Chinese dialects. When using the Russell Soundex method, it is found that the number of duplicates is high and the number of false positives is also high. The adaptive nature of Soundex method provides an avenue to optimize it for foreign language names, such as Chinese names. Through a series of tests, this study has optimized the Soundex codes for general Malaysian Chinese names. The test results have shown that a few short Chinese surnames contribute to false positives.**

Keywords: duplicate detection; Chinese names; Soundex; false positive.

I. INTRODUCTION

Normally, during data cleansing projects, duplicate detection of foreign language names is done by using the Soundex method [1]. When using the original Soundex method for Malaysian Chinese names, it is found that it produces a large number of duplicates and false positives. Be it in the government or private office use, Chinese names are written in English text, but they actually come from various phonetics of respective dialects. It is possible that the same Chinese character names are being translated into different spelling names (Romanized) due to different phonetics systems of Chinese dialects. Chinese names are normally written either as three words or two words. Sometimes, English names are also prefixed to the names. The different name formats would affect the matching results with Soundex codes. Emma Woo [4] described ten most popular Chinese format names in America. Some of the Chinese American formats also apply in Malaysia. This study excludes other ethnicities including Malay, Indian, and Kadazan, because their names are suitable for string matching algorithm application.

There is a large number of duplicate detections approaches such as Levenshtein [13], edit distance [3] and Soundex. In the Levenshtein method, edit distance and name comparison methods match results based on substitution, deletion, insertion or transposition of characters. In this study, we selected the Soundex method because it can process phonetic data effectively with its matching Soundex codes. The Russell Soundex method was invented by Robert Russell and O'Dell in 1918 [2]. A large number of studies have been conducted to optimize Soundex rules [5] [6] [7] [8]. A cross-language algorithm was developed to measure the similarity of Asian words phonetically [9]. The algorithm showed positive results for Asian names but it had limited success when using Chinese names. Soundex-Pinyin is a spelling correction system to detect Chinese strings, but this system requires Pinyin input [10]. In this study, a newly improved algorithm based on the existing Soundex algorithm is proposed to optimize the detection results of names. An investigation is also done to improve the algorithm on Chinese names detection while it is developed as a plugin component applicable to any data cleansing or ETL (Extract Transform Loading) process workbench. The duplicate detection process reads Malaysian Chinese names as input and then applies the Soundex method. The matching records are written to a duplicate list, while the non-matching records are written to a non-matching list. The matching list is further analyzed by a subject matter expert to determine Chinese word relevance. If it is relevant and matched, the

 ISBN: 978-1-61208-552-4

outcome is true positive, otherwise it is considered false positive, for example Chan and Chin surnames.

The results of this study are supported through the use of a confusion matrix, which provides false positives and accuracy readings [12]. False positive readings provide information on the ability of the Soundex method to detect the correct duplicates. The accuracy reading provides to the user the confidence level of the Soundex method in general.

The rest of the paper is structured as follows. In Section II, the problem statement is formulated. In Section III, the method to customize Soundex rules is described. In Section IV, the results using the new Soundex are presented and discussed. Finally, in Section V, this paper concludes with a summary and ideas for future enhancement.

II. PROBLEM STATEMENT

A. English text based Soundex method

This study uses the Soundex method for duplicate name detection; thus, the matching is done on the Soundex codes. The Russell Soundex method was developed with the following rules, to produce a four-digit alphanumeric code:

- Step 1: The first letter of the name is selected as the first digit of code, but all occurrences of characters A,E,H,I,O,U,W,Y are omitted.
- Step 2. Assign the codes to the letters as in Table I.
- Step 3. If two or more letters with the code were adjacent in the original name, omit all but the first.
- Step 4. Convert to the four-digit format by adding training zeros if there are less than three digits, or by dropping the right most digits if there are more than three.

TABLE I. RUSSELL SOUNDEX

Letters	Code
B, F, P, V	1
C,G,J,K,Q,S,X,Z	2
D,T	3
L	4
M,N	5
R	6
A,E,H,I,O,U,W,Y	Omitted

The Russell Soundex is accurate to detect character similarities between names after vowels and coded characters are dropped. It is suitable to detect duplicate names due to its nature to detect names based on closest phonetic sounds. The Soundex was originally developed to uncover specific relative's names in American history journals, but in recent time, it is used in record linkage analytics [11]. The major problem was that the English based Soundex did not produce good results when it was applied for Chinese names. The original Soundex's rules were not suitable for detecting Chinese names. In the first duplicate detection test with Russell Soundex code, the result was not encouraging because there was a high a number of duplicates and false positives.

B. Multiple formats of Chinese names

The collection of 300 and 527 names from our experiment data also provides us with some information on how the actual scenario might be found in government or private agencies when processing Chinese names. There are a few name formats to consider in the Soundex process that might affect the accuracy. In particular, when working with 4 digit codes, the prefix with English name, for example Franky Cheah, and double names with alias symbol for example Chin@Ah Kow, have very limited success in producing true positive duplicates. The name formats encountered in our sample data for testing are described below.

1. Chinese name with surname first: Lou Sheng, Lin Hui Ling.
2. Chinese name composed of three or more separate words: Kwai Yung Chui Ja
3. Hyphenated Chinese disyllabic name, with an uppercased second syllable: Han-Sheng Lin
4. Combination of an American given name and a Chinese middle name: Jean Yun-Hua King
5. Family name in the middle of the name: Abraham Ng Kamsat
6. Chinese name with alias of second name: Ngoh Swee Lan @ Ng Swee Lan

III. METHODS

A. Customize Soundex rules

In this study, we present three rules, namely Soundex 13, Soundex 20 and Soundex 21, that show significant results. By using data integration tool, such as Pentaho or Talend, the Soundex script is applied to an ETL flow. Matching the duplicates was done by comparing the four-digit alphanumeric codes against duplicate items. The duplicate matching result is further examined manually in order to determine the number of duplicates and false positives. We tested two sets of names (300 and 527) and the results of the Russell Soundex test is given in Table IV. Apparently, there was a high number of false positives in the duplicates results. The duplicate items, in general, were closely spelled names.

In Table II, the Soundex 21's rules are described as follows:

1) Vowels are considered for Chinese names due to the fact that vowels are the core in Chinese syllabic structure. Vowels cannot be omitted in any Chinese syllables and the simplest Chinese syllabic structure is composed of a vowel and a tone, for instance, "I" in International Phonetic Alphabet (IPA) or *yi* in Hanyu (Pinyin). There are two values for vowels, 6 for "I" and 7 for "U". The two characters "Y" and "W" (value 8) are also included in the matrix as these two glides are regarded as allophones of the high vowels /i/ and /u/ under certain conditions in Chinese.

 ISBN: 978-1-61208-552-4

TABLE II. CUSTOM SOUNDEX RULES

Code	Soundex 13	Soundex 20	Soundex 21
1	P,F,B,V,M	P,F,B,V,M	B,F,P,V
2	D,T,L	D,T,L	D,T,L
3	J,Q,X,K,G	J,Q,X	J, Q,X
4	Z,C,S,R	Z,C,S,R	Z,C,S,R
5		K,G	K,G
6	I	I	I
7	U	U	U
8	W,Y	W,Y	W,Y
Omitted	A,E,H,O,N	A,E,H,O,N	A,E,H,O,N, M

2) For consonants, we grouped them according to the place of articulation. For instance, value 1 is for labial consonants " B, P, F, V"; value 2 is for denti-alveolar consonants "D, T, L"; value 3 is for alveolo-palatal consonants " J, Q, X"; value 4 is for denti-alveolar consonants " Z, C, S" and fricative "R"; value 5 is for velar consonants "G, K". Among these consonants, "V" is a special one as it is used by Malaysian Chinese names due to the fact that many names are spelled in dialects. However, the pronunciation of "V" is not used in Mandarin.

3) The three vowels A, E, O and two other letters M,N are omitted due to their high frequency in word histogram (Table V).

In Soundex 13, the main difference from Soundex 21 is the labial consonants "B,P,F,V,M" that include the M for Chinese phonology nasal sound. Value 5 is left blank for consistency. The omitted letters are "A,E,H,O,N". In Soundex 20, the main difference from Soundex 21 is the labial consonants "B,P,F,V,M" in which there is the additional M for Chinese phonology nasal sound. The omitted letters are "A,E,H,O,N". After many test iterations, it was found that Soundex 21 produces better results than Soundex 13 and Soundex 20.

IV. RESULTS

Given the experimental data from two data sets, namely 300 name lists and 527 name lists, four tests were conducted using Russell Soundex and the customized Soundex versions 13, 20 and 21.

The result data is divided into actual and predicted output in order to calculate true positives (TP), true negatives (TN) and false positives (FP) readings. The accuracy and false positives formulas are given in Table IV. With the help of a Chinese language expert, we identified the duplicates by taking into consideration at least two words and also close spelling which should represent the same Chinese character in reality. Actual duplicates are recorded in a data matrix. After the Soundex tests, prediction observation is recorded such as TP, TN, FP. The prediction data is either yes or no. Thus, each set of data then has the confusion data matrix ready for examination.

A. Duplicate detections with Russell Soundex

The results in Table IV show duplicates, accuracy, and false positives. The non-duplicate list shows false negatives, whereby supposedly matched names are not detected.

The number of duplicates found for 300 and 527 names are 45 and 85, respectively. The false positives percentage was 46.3% for the first set of 300 names. The family name Chong and Cheong were close, but they did not represent the same Chinese character, similarly Lai and Lee, also Chin and Chan. The family names heavy with vowels and short in length would likely cause false positives if their first names were the same or similar in spelling.

Referring to Table III, the names Lim Jing, Tan Sin Yee, Wong Meow Fah, were given the respective codes L525, T525 and W525. However, the algorithm also gave the same codes for Lim He Jian, Tan Jenny and Wong Nyet Yin which were not related in reality. The names were detected as closely matched because Russell Soundex omitted many characters that Chinese names usually have.

TABLE III. EXAMPLE FALSE POSITIVES WITH RUSSELL SOUNDEX

Set	Code	Name	Chinese Characters
1	L525	Lim Jing	林 静
	L525	Lim He Jian	林 何 健
2	T525	Tan Sin Yee	陈 欣 宜
	T525	Tan Jenny	陈 珍 妮
3	W525	Wong Meow Fah	黄 妙 花
	W525	Wong Nyet Yin	黄 月 英

B. Duplicate detections with customized Soundex rules

For the first stage of the project, we had our Soundex method testing on the names using a revised phonetic algorithm. For the 300 names list, Russell Soundex detected a higher number of duplicates and accuracy readings as compared to Soundex 13, Soundex 20 and Soundex 21. It is also noticed that, in the set of 300 names, the false positives of Soundex 21 were fewer than Soundex 20. For the 527 names list, the Soundex 13 had the most number of duplicates. For the 527 names list, the Soundex 13 had 38.3% of false positives and that was higher than Soundex 20 and 21. Both Soundex 20 and 21 had almost the same number false positives. In general, the customized Soundex results had lower false positives and number of duplicates than the results from Russell Soundex. The accuracy of custom rules is also generally higher than using Russell Soundex. The number of false positives for Soundex 20 and Soundex 21 was almost the same. The effect of letter M was not significant. The Soundex 21 was far better than Soundex 20 in accuracy readings of the 300 names list, but both had the same accuracy for the 527 names list.

 ISBN: 978-1-61208-552-4

TABLE IV. SOUNDEX WITH 4 DIGIT RESULTS

Soundex Type	Number of Names	Number of Duplicates	Accuracy % ((TP+TN)/ Total)	False Positive % (FP/(FP+ TN))
Russell	300	45	57	46.3
	527	85	46	64.8
13	300	24	79	18
	527	86	65	38.3
20	300	32	78	23.8
	527	75	69	31.7
21	300	32	83	19.4
	527	75	69	31.8

TABLE V. WORD HISTOGRAM OF 5 CHARACTERS IN THE 527 NAME LIST

Characters	6 digits	7 digits	8 digits
A	276	330	357
E	255	320	360
H	251	285	311
O	196	216	230
N	242	317	396

There were false positive results due to the presence of English names prefix, the closely matched spelling of different family names, such as between Chia and Chai, and the dissimilarities between first names (second and third) of the same family name (first).

When applied to Chinese names, the Soundex result had a few false positives because of misspelled names and English name prefixes and closely matched name consonants. The closely matched name occurred when all vowels and other omitted characters were removed, such as, in Soundex 13, Tee Yan Qi and Tan Hwa Jie revealed a similar code, namely T735.

C. Conclusions

Soundex 21 was generally far better than Russell Soundex in producing fewer duplicates and false positives. The false positives of Soundex 21 were slightly fewer than Soundex 13, but almost the same as Soundex 20. The Soundex 21 had higher accuracy reading than Soundex 20. This result gave confidence that the Soundex 21 was able to produce a good duplicate detection result.

V. CONCLUSION

When using Russell Soundex in the duplicate detection, the result produced a high number of false positives. The number of false positives is reduced while the percentage of accuracy is increased when the code rules are customized and improved in Soundex 13, Soundex 20 and Soundex 21. As a result, the duplicate detection, especially with Soundex 21, produced an acceptable result. Future research is needed on automating the intelligence to detect and verify Chinese names as part of a duplicate error correction system.

ACKNOWLEDGMENT

The authors of this paper would like to thank MIMOS Bhd for their involvement in ETL project at a commercial bank.

REFERENCES

[1] J. Soo, O. Frieder, “On foreign name search” European Conference on Information Retrieval, pp. 483-494, Springer Berlin Heidelberg. 2010.

[2] R. Russell and M. Odell. "Soundex." US Patent 1, 1918.

[3] K. Rieck and C. Wressnegger, “Harry: A Tool for Measuring String Similarity” Journal of Machine Learning Research 17, pp. 1-5, 2016.

[4] E. W. Louie, “Chinese American Names: Tradition and Transition” McFarland, 1998.

[5] R. Shah, "Improvement of soundex algorithm for indian language based on phonetic matching." International Journal of Computer Science, Engineering and Applications 4, No. 3, pp. 31-39, 2014.

[6] D. Holmes and M. C. McCabe, "Improving precision and recall for soundex retrieval." International Conference on Information Technology: Coding and Computing 2002, pp. 22-26, IEEE Press, 2002.

[7] A. H. Yousef, "Cross-language personal name mapping." International Journal of Computational linguitics Research, vol 4, issue 4 , 2013.

[8] D. Pinto, D. Vilarino, Y. Aleman, H. Gomez, N. Loya and H. Jimenez-Salazar, "The Soundex phonetic algorithm revisited for SMS text representation." International Conference on Text, Speech and Dialogue, Springer Berlin Heidelberg, 2012.

[9] O. Htun, S. Kodama, and Y. Mikami, "Cross-language phonetic similarity measure on terms appeared in asian languages." International Journal of Intelligent Information Processing 2.2, pp. 9-21, 2011.

[10] D. H. Li and D. W. Peng, "Spelling Correction for Chinese Language Based on Pinyin-Soundex Algorithm," International Conference on Internet Technology and Applications, Wuhan, pp. 1-3, IEEE Press, 2011.

[11] A. Karakasidis and V. S. Verykios, "Privacy Preserving Record Linkage Using Phonetic Codes." Fourth Balkan Conference in Informatics 2009, pp. 101-106, IEEE Press, 2009.

[12] V.A. Narayana, P. Premchand, A. Govardhan, “Performance and Comparative Analysis of the Two Contrary Approaches for Detecting Near Duplicate Web Documents in Web Crawling” International Journal of Computer Applications (0975-8887), Vol 59, No. 3, 2012.

[13] I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals” Doklady Akademii Nauk SSSR, 163(4), pp. 845–848, 1966.

 ISBN: 978-1-61208-552-4