Report on Misspelled Location Names for Approximate String Search Analysis

Chao Wei

SID:748116

Abstract

Commonly used local name in tweets may be different form the official name or misspelled. In this report I explore approximate string matching techniques to find location name in tweets based on gazetteer. Global Edit Distance and N-gram Distance are important matching strategy to solve the approximate string search problem, which are used in my system.

1 Introduction

Social networks such as Twitter and Facebook received hundreds of millions of messages sent by users per day. A fundamental amount of matching work has focused on exact match. However, if users misspell the name of location they are looking for, their match will not successful.

N-gram is a statistical analysis of text of speech content to find a number of some sort of item in the text. The Global Edit Distance transform the string of interest into each dictionary entry, using the operations Insert, Delete, Replace, and Match. These two methods contrast words without regarding the language used. Another kind of strings search strategy is very language sensitive: The method Soundex compares words with regard to their phonetic similarity.

In this report, I focus on the effectiveness of these methods and also identify their advantages and disadvantages of the different methods.

2 Global Edit Distance

Global Edit Distance is a fundamental method for measuring similarity between string, which calculate the minimum number of character insertions, deletions, and replace needed to transform one string to the other. (Damerau 1964)

Damerau-Levenstin-metric for calculating the minimum number or error for the two words s and t.

$f(0,0) = 0$

$f(i,j) = \min \{ \quad f(i-1,j)+1,$

$f(i,j-1)+1,$

$f(i-1,j-1) \; + \; d(s_i,t_j),$

$f(i-2,j-2) \; + d(s_{i-1},t_j) + d(s_i,t_{j-1}) \; + 1 \}$

The function d is used to measure distance of letters. There is a simple measure used in following.

$$d(s_i,t_j) = \begin{cases} 0, \; if \; s_i = \; t_j \\ 1, \; if \; s_i \neq \; t_j \end{cases}$$

The function $f(i,j)$ is used to calculate the minimum number of difference that differentiate the first i character of the first word from the first j character of the second word.

By application of dynamic programming techniques, the time complexity of my java system has been reduced to O(n^2).

2.1 Experiment Results and Discussion

For each location name based on gazetteer I searched in the tweets query, the results of Global Edit Distance measuring which are greater than 2 have great difference between original name. There are some examples presented in following.

Query String :Acorn Hill Childrens Center
Tweet String :UC DAVIS Student protesters gathering at Mondavi Performing Arts Center for speech on fee hikes tonight Dutton Hall no longer occupied.
Distance: 9

Query String :Chardon Fire Department
Tweet String: Nutritionist Washington DC Nutritionist Req Number Department WIC Clinic SchhttpwwwshurinfohcnutritionistWashingtonDC
Distance: 9

Query String :Allenmore Hospital
Tweet String :Thanks all for the kind thoughts passed on to Lucy whos doing hospital duty Weekend lab sched means a likely stay of several days.
Distance: 6

Query String :Averys Hill
Tweet String : Very tired Watching the Hills before bed
Distance: 3

Additionally, considering of the context of words, the shortest Global Edit Distance will not always perform best match. For instance, the distance between “example” and “the example of very popular string operations” is 34; the distance between “example” and “expected ampersand lesson” is 18. Obviously, the first match is more accurate.

Moreover, the performance of Global Edit distance performance is good when dealing with nearly similar words, for example casual misspellings, but if the Global Edit Distance is longer than the word itself, it has no worthy meaning as similarity value.

In addition, there are no misspelled location names finding in tweets, only the exact matches. In this case, the precision might be 0.

3 N-grams

An *n*-gram is a adjoining sequence of *n* items from a given sequence of text. The items can be phonemes, syllables, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus. When the items are words, *n*-grams may also be called shingles.

A “unigram” is an n-gram of size 1, a “bigram” is an n-gram of size 2, size 3 is “trigram”. Trigrams and bigrams perform the best results in matching similar words to a given word described by Salton (1988) and Zamoa (1981) .My system is based on bi-gram algorithm.

The following is bi-gram language model:

A string of English words can be represented by W = $w_1 w_2 w_3 w_4 \dots w_n$

$P(w_1 w_2 w_3 w_4 \dots w_n) \cong P(w_1) P(w_2 \mid w_1)\, P(w_3 \mid w_2)\, P(w_4 \mid w_3) \dots P(w_n \mid w_{n-1})$

3.1 Experiment Results and Discussion

For each location name based on gazetteer I searched in the tweets query, the results of bigram measuring which are greater than 2 have great difference between original name. There are some examples presented in following.

Query String :Angola Wesleyan Church
Tweet String :Im tellin yall the singin neva stops I woke up this mornin to sing at school an now I find myself singing at a church gathering.
Distance: 10

Query String :Athanasiou Valley Airport
Tweet String :Im at the airport Im all jitters so excited.
Distance: 11

Query String :Berua Church
Tweet String :Im tellin yall the singin neva stops I woke up this mornin to sing at school an now I find myself singing at a church gathering.
Distance :5

Query String :Best Western Shaheen Motel
Tweet String :When BHO accepts the Peace prize will Kayne West jump up to the microphone and tell us who really should have won nobel tcot
Distance :13

To sum up, there are some exact matching of location names, but zero misspelled names finding in tweets. In this case the precision of n-gram system might be 0.

In addition, n-gam model is much simple than Global Edit Distance model, and easily to perform. The n-gram method works well on English, but quite useless for small alphabets languages. The reason is that there are 26 alphabets in English, so that there are 26^2 = 676 possible bigrams for that alphabet. For other language with small alphabets this number would be much smaller. It is hard to distinguish the difference from these limited combinations.

4 Conclusion

In this report two different strategies for calculating the similarity of strings have been introduced and compared. Moreover, I have described that the advantages and disadvantages of two methods.

Future work will have to deal with the combination of some strings matching techniques such as edit distance, soundex, vector space model, cosine similarity.

5 references

Damerau, Fred J. *(March 1964), "A technique for computer detection and correction of spelling errors", Communications of the ACM, ACM, 7 (3): 171–176.*

Salton, G.; McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Zamora, E.; Pollock, J.; Zamora, A. (1981). The Use of Trigram Analysis for Spelling Error Detection. *Information Processing and Management 17(6)*, pages 305–316.