A Chinese String matching Algorithm Based on Edit Distance for Electric Power Customer Service

Xusheng Liu, Songhe Mu *, Zhiming Li, Zixing Yang, Wei Han and Zhiwei Wang

Information operation and maintenance centre of customer service centre in State Grid Co., Ltd, Tianjin, China

*Corresponding author e-mail: gwkfzx_xxywzx@163.com

Abstract. With the rapid development of information technology, electric power customer service needs to extract accurate information from the data centre to provide customers urgently. String matching is the most straightforward solution. There are some problems in the editing distance definition of traditional string matching algorithm when dealing with Chinese text. In this paper, the pronunciation characteristic of Chinese characters is integrated in the definition of edit distance and a string matching algorithm based on edit distance for electric power customer service data centre is proposed. Finally, through simulation analysis, we demonstrate that this algorithm has good reliability when dealing with Chinese string matching problem. Thus, it is suitable for electric power customer service to obtain the target string.

Keywords. electric power customer service, string matching, fuzzy matching, edit distance.

1. Introduction

With the rapid development of information technology, people have a higher demand for information acquisition, especially the electric power industry users. With the explosive growth of power data, how to extract the information which customer needs from mass data has become a big problem for power customer service. The main functions of electric power customer service include business acceptance, information inquiry, electric power failure sheet and complaint sheet processing, etc. All these functions require accurate positioning of the customer's information into the data sheet [1]. As the most direct information matching method, string matching technology has become a research hotspot in the information age. Electric power customer service urgently needs a string matching technology suitable for this system to solve the above problems.

There are two types of string matching: perfect match and fuzzy match. Perfect match means that the result of searching is the same as the string to be searched. Fuzzy matching refers to the calculation
of the matching degree between the searched string and the string to be searched. In reference [2], a threshold-based fuzzy matching method was adopted to define the similarity comparison function and similarity threshold to search for the target character. Reference [3] adopts the fuzzy matching method based on top-k, defines the candidate string set and applies the range method to match k most similar strings from the set.

The key to fuzzy matching is to calculate the matching degree between the specified string and the target string. According to the existing research, the string matching degree is mainly calculated based on edit distance, and there are many ways to express edit distance. In reference [4], the hamming distance is used as the editing distance. This method requires two strings to have the same length, so as to compare the number of different characters in the same position. In reference [5], the similarity degree of the q-gram set was calculated to measure the string difference. In this method, the word segmentation of the two strings was firstly processed, and the number of the same substrings of the two word segmentation sets was calculated by sliding window. Reference [6] regards the string to be compared as a vector and calculates the cosine value of the Angle between the two vectors. In other words, the cosine similarity of the two strings represents the edit distance. In reference [7], the edit distance is calculated by using the idea of dynamic programming. This method establishes an edit distance matrix and calculates the elements in the matrix in the order from top left to bottom right through the recursive formula of edit distance. Finally, the elements in the bottom right corner are the edit distance of two strings.

At present, the method of dynamic programming is mainly used for the calculation of edit distance, but the traditional recursive formula of edit distance is not suitable for the matching of Chinese string, so it can’t be used for the string matching of electric power customer service data centre. In this paper, the recursive formula of edit distance is improved by fusing pronunciation characteristic of Chinese characters, and a string matching algorithm of electric power customer service data centre based on edit distance is designed. This algorithm greatly improves the accuracy of Chinese string matching and can solve the problem of target string search for electric power customer service.

2. Calculation of matching degree

2.1. Edit Distance Recursion Formula

Edit distance is the minimum cost required to convert two strings, including adding, deleting, and replacing a character [8]. The larger the edit distance is, the greater the difference between the two strings is.

When the traditional method is used to calculate the edit distance, usually every operation on a character is regarded as an atomic operation and the cost of each atomic operation is one. The way to considerate the information is relatively simple. This method is suitable for English string matching, but the weight of each atomic operation needs to be considered when dealing with Chinese string matching. First, since a replacement operation can be regarded as a deletion and an increase, the cost of deletion and increase operation should be lower than that of a general replacement operation. For example, the conversion cost of "Beijing city" and "Beijing" should be less than that of "Nanjing" and "Beijing". In addition, the traditional substitution cost of one also ignores some features of the character itself. Here the phonetic characteristics of Chinese characters are integrated to improve the weight of substitution cost. For example, the first two characters of "Beijing customer" are different
from the first two characters of "platinum customer". However, due to the similar pronunciation, it is likely to represent the same string. Therefore, the cost of manipulating atoms is improved compared with the recursive formula of traditional editing distance.

According to the cost of operating atoms, the recursive formula of editing distance designed in this paper is shown in equation (1). The first two lines of this formula indicate that the edit distance increases by 0.5 when increasing and deleting operations are performed. That is, the atomic cost of deleting and increasing operations is set to 0.5. Where, \(S_1\) and \(S_2\) are two strings to be matched, \(S_1[i]\) and \(S_2[j]\) respectively represent the \(i\)th character of string 1 and the \(j\)th character of string 2, and \(d[i,j]\) represents the edit distance between the first \(i\) character of string 1 and the first \(j\) character of string 2. When \(i\) and \(j\) are the lengths of the two strings, \(d[i,j]\) is the edit distance of the two strings. The last two lines represent the change in edit distance during the substitution operation. The cost of the substitution operation is defined here as the ratio of the traditional edit distance \(d_p\) to the spelling encoding length of the two characters. Spelling code is an English string, so its edit distance can be calculated by the traditional edit distance formula.

\[
\begin{align*}
\text{if } S_1[i] = S_2[j] & : d[i,j] = \min\{d[i-1,j] + 0.5, d[i,j-1] + 0.5, d[i-1,j-1]\}, \\
\text{if } S_1[i] \neq S_2[j] & : d[i,j] = d[i-1,j-1] + d_p / \max(|S_1|, |S_2|),
\end{align*}
\]

(1)

2.2. String matching Degree

After the recursive formula of edit distance is defined, the edit distance of two strings \(d[|S_1|,|S_2|]\) can be obtained by dynamic programming. Then, the matching degree of the two strings is calculated based on the edit distance. The formula for calculating the matching degree \(r^s\) of strings 1 and 2 is shown in equation (2). Where, \(\max(|S_1|,|S_2|)\) represents the larger value of the length of two strings.

\[
r^s = \frac{\max(|S_1|,|S_2|) - d[|S_1|,|S_2|]}{\max(|S_1|,|S_2|)}
\]

(2)

3. String matching algorithm for electric power customer service

In the previous chapter, we improved the calculation of the matching degree in the string matching problem. However, this method can't be directly applied to the information query system of electric power customer service. Electric power customer service needs to match the customer's inquiry string to the target string in the sheet of data centre. To solve this problem, a novel string matching algorithm for electric power customer service is designed here and the process is shown in figure 1.

First, the algorithm selects the keyword set based on the sheet properties of the data centre. Then, it groups the customer's consultation string based on the keywords. Each set of short strings corresponds to a keyword that is exactly matched to query sheet properties in the sheet. After querying the corresponding sheet attribute, the matching degree calculation method in the previous chapter is applied to calculate the matching degree between the set of short strings and the strings under the sheet attribute. When each group of strings has been matched, the customer consultation strings of the groups are unified into a whole. According to equation (3), the matching degree between the total
string of customer consultation content and the total string constituted by the query path of the candidate sheet is calculated. Finally, the query path string with the maximum matching degree is selected as the target string and the corresponding data information is fed back to the customer.

\[ D = \sum_{i=1}^{I} \alpha_i \tau^s(i) / I \]  

In this formula, \( D \) is the total string matching degree, and \( I \) is the number of keywords, \( \alpha_i \) indicating whether the keywords of the string in group \( i \) are exactly matched to the corresponding sheet attribute. The value of \( \alpha_i \) is one if exactly matching, while the value is zero if not. \( \tau^s(i) \) indicating the matching degree between the customer consultation content string in group \( i \) and the candidate strings belong to the corresponding sheet attribute.

![Diagram of string matching process for electric power customer service](image)

**Figure 1.** Figure of string matching process for electric power customer service.
4. Simulation and performance evaluation

4.1. Simulation parameter setting

In this paper, 100 customer voices are selected for recognition, and the corresponding 100 customer consultation strings are obtained. These strings are grouped according to the keyword set, which involve different quantity of keywords. The number of the keywords satisfies the uniform distribution of [10, 40].

To verify the advantages of the algorithm in this paper, we choose the string matching algorithm based on traditional editing distance in literature [7] as the comparison algorithm. In this paper, two indexes, matching difference degree and matching accuracy, are selected to evaluate the performance from point and surface perspectives. The matching difference represents the difference value between the matching degree of the target string and the interference string. The interference string here is the string that has the highest matching degree except the target string under the corresponding report property. Matching accuracy represents the percentage of the strings that have been successfully matched to the target string.

4.2. Simulation results and analysis

By matching the group strings with the traditional algorithm and the algorithm in this paper, we can get the matching difference of each group string under the two algorithms. Here are some examples of typical strings to be matched. Partial instances are shown in table 1.

| Keywords   | String to be matched | Difference degree of traditional algorithm | Difference degree of the algorithm in this paper |
|------------|----------------------|--------------------------------------------|-----------------------------------------------|
| City       | Tianjing             | 0                                          | 0.17                                          |
| Organization | Jianlanyuan         | -0.27                                      | 0.14                                          |
| Time       | April in 2019        | 0                                          | 0.14                                          |
| Index      | Power consumption    | -0.05                                      | 0.23                                          |

As can be seen from table 1, when the traditional editing distance calculation method is used for string matching, the matching difference degree may be zero or negative. This means the interference string is more similar with the string to be matched than the target string, resulting in string matching errors. However, the algorithm in this paper considers the character pronunciation when calculating the edit distance. Thus, in the case of several examples of traditional algorithm matching errors, the matching difference degree of the algorithm in this paper is all positive. This shows that the algorithm in this paper can be well applied to Chinese string matching, making up for the shortcomings of the traditional algorithm.

In order to describe the accuracy of target string matching by the algorithm in this paper more intuitively, the average matching accuracy under the number of keywords is selected as the measurement index in this paper. The simulation results are shown in figure 2. As we can see from this figure, as the number of keywords increases, the difficulty of string matching also increases. Thus, the matching accuracy presents a declining trend. However, the string matching algorithm in this paper has better performance of matching accuracy than the traditional algorithm under the same keyword set because of the optimization of edit distance recursion.
Figure 2. Figure of string matching accuracy.

To sum up, the string matching algorithm based on edit distance for electric power customer service proposed in this paper has higher reliability in dealing with Chinese string matching problem, and the data feedback to customers is more accurate.

5. Conclusion

This paper proposes a string matching algorithm for the information query of electric power customer service. This algorithm integrates the pronunciation characteristic of Chinese characters into the recursive formula of edit distance, which solves the problem that traditional string matching algorithm is not suitable for Chinese string matching. Through the simulation analysis, it is demonstrated that the algorithm can improve the reliability of matching target string to a great extent. When applied to the electric power customer service consulting system, the algorithm in this paper can feedback more accurate data to customers.

References

[1] Tian Y, On how to improve 95598 electric power customer service, China Market, 2013, pp. 36-37.
[2] Deng D, Li G and Wen H, An efficient matching-based method for error-tolerant autocompletion, Proceedings of the VLDB Endowment, 2016, pp. 828-839.
[3] Beebe N L and Clark J G, Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results, Digital Investigation, 2007, pp. 49-54.
[4] Russo L M, et al, Approximate string matching with compressed indexes Algorithmica, 2003, pp. 347-369.
[5] A. Behm, S. Ji, C. Li and J. Lu, Space-constrained gram-Based indexing for efficient approximate string search, 2009 IEEE 25th International Conference on Data Engineering, Shanghai, 2009, pp. 604-615.
[6] Liu H and Wang L, Method of recognizing Chinese characters by convolution network combined with cosine correlation, Computer Engineering and Application, 2019, pp. 1-7.
[7] Wei L and Wu X C, Application of dynamic programming string matching algorithm in curve comparison, Computer Engineering and Application, 2007, pp. 8-9+17.
[8] Jin P, Yang J, Wang Z W and Liu K P, A fuzzy address recognition model based on minimum distance editing, Electric Power Big Data, 2019, pp. 9-15.