Low-voltage distribution network topology verification method based on Revised Pearson correlation coefficient

Bing Lou¹, Cailong Li¹, Jun Deng¹, Linjie Zhu¹, Cong Yang¹ and Wujun Chen²,³
¹Huayun Information Science and Technology Co., Ltd, Hangzhou 310012, China
²Ningbo Tianling Technology Co., Ltd
³E-mail: 4824689@qq.com

Abstract. Due to the high cost and low real-time reliability of low-voltage distribution network topology checking by manual field verification, an online verification method based on revised Pearson correlation coefficient with error expectation values and KNN algorithm is proposed. Firstly, the Pearson correlation coefficient with error expectation values is used to judge the similarity between the user's voltage curve and the transformer substation voltage curve, then the user with incorrect household relationship is found to perform re-verification. For performing re-verification, the user sample set is generated based on the data of GIS system and the Technical Guidelines for Distribution Network Planning and Design, and the K-Nearest Neighbor (KNN) algorithm is used to find out the correct station to which area the user belongs. Finally, based on the results of the manual field verification, the correctness of the algorithm verification is judged.

1. Introduction

The correct low-voltage distribution network topology relationship, especially the correct relationship between household and transformer substation, is the basis and research hotspot of current distribution network management refinement, loss reduction and energy conservation. At present, the topology related data of the low voltage station area is basically manually entered into the computer system. Low-voltage users have complicated on-site wiring, large amount of data, changes in operating mode due to unbalanced load, resulting in computer systems errors. Therefore, it is urgent to verify the distribution network topology. At present, the relationship between household and transformer substation is the most important factor in the distribution network topology.

Based on the research and application situation at home and abroad, the topological relationship of low-voltage distribution network is mainly verified by manual site identification (manual method) and automatic system identification (online mode). The manual method needs to arrange the staff to the site for identification. The method recognizes the results accurately, but the efficiency is low and the cost is high. The online mode adopts automatic identification technology, and the recognition efficiency is high, but the success rate is limited by the application scenario. The method mainly includes power failure identification in the substation area, correlation analysis based on household frequency power zero-crossing sequence, scrambling of station feature enhancement equipment, correlation determination based on household history blackout event records, and multi-information correlation analysis [1-3]. Among them, multi-information correlation analysis method using big data has incomparable real-time advantages compared with other methods, and has become a new trend of distribution network topology verification methods.
Based on the topology analysis method [4-7] and the topology identification principle of 10kV distribution network, the multi-information correlation analysis of low-voltage distribution network topology is improved. There are two main types of distribution network topology identification methods: methods based on bus voltage measurement and methods based on injection power measurement. In [8], based on the bus voltage measurement, the outlier detection method is used to verify the distribution network topology, but the low-voltage distribution network topology in a certain area cannot be quickly and massively verified. In [9], based on the variance model of the branch voltage deviation measured by the bus injection power measurement, the topology operation structure of the distribution network is identified. However, the premise of this method is that the bus injection power of the distribution network is completely observable. In [10], the paper introduces the method of finding the abnormal electric energy meter by using the Pearson correlation coefficient algorithm from the relationship between the line loss electric quantity and the user electric quantity. However, the premise of this method is that user electric power errors are much larger than the power loss generated by other factors.

Based on the above methods, the Pearson correlation coefficient with the error expectation values [11] is improved. The latter can help improve the accuracy of topology checking by only by the former. This paper uses the revised Pearson correlation coefficient and KNN algorithm to verify the low-voltage distribution network topology, which can quickly and accurately verify the topology of the low-voltage distribution network, compared with manual verification.

2. Preliminary

2.1. False form of relationship between household and transformer substation

At present, there are two main types of false form of relationship between household and transformer substation in the low-voltage area.

1) The topological relationship of the user profile information records in the computer system does not match the actual topological relationship. This type of error can affect the correct identification of the low voltage distribution network topology. For example, when the power is changed, the user profile information is not entered in time. After the power is changed, User 1 actually belongs to Zone A, but User Profile Record User 1 is still in Zone B. If the line loss of the station area is calculated, the calculated line loss of the station area A will be small, that is, the calculated line loss is smaller than the actual line loss; the calculated line loss of the station area B is large, that is, the calculated line loss is larger than the actual line loss.

2) User profiles in the computer system miss some users. This type of error can affect the staff to properly manage the entire zone user. For example, when the industry expands the report, a new user is added to the station A, but the file information is missing the user.

Since the relationship between household and transformer substation is the most important factor in the distribution network topology, this paper verifies the distribution network topology by verifying the relationship between household and transformer substation.

2.2. Topology verification principle of low voltage distribution network

The topology verification principle of the low-voltage distribution network used in this paper is divided into the following two steps.

1) According to the similarity of the user voltage curve, verify the user of the station area and find out the users who do not belong to the station area.

2) For users who do not belong to the station area, find the correct station area to which the user belongs according to the Technical Guidelines for Distribution Network Planning and Design and the K-Nearest Neighbor (KNN) algorithm.

In low-voltage distribution networks, the voltage fluctuates frequently due to the uncertainty of the load everywhere. The load with relatively close electrical distance has similar voltage fluctuation curves (high correlation), while the load with relatively long electrical distance has low similarity of
voltage fluctuation curve (low correlation) [12-13]. Therefore, the user voltage curve is selected as the verification basis.

3. Proposed solution

3.1. User voltage curve similarity measure

The measurement of voltage similarity of smart meters in this paper uses the Pearson correlation coefficient (PCC).

The PCC, also known as the Pearson product moment correlation coefficient, is used to measure the linear correlation between two variables X and Y, with values between -1 and 1.

The PCC is defined as shown in Equation (1).

\[ P = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}} \]  

where \( \bar{X} \) and \( \bar{Y} \) represent the mean of X and Y, respectively. The value of P is between -1 and 1. The greater the absolute value of P, the higher the correlation between variable X and variable Y. The smaller the absolute value of P, the lower the correlation between variable X and variable Y. For the evaluation of PCC as shown in Table 1.

| Range of P | Relevance                      |
|-----------|-------------------------------|
| 0~0.2     | Very weakly related or unrelated |
| 0.2~0.4   | Weak correlation               |
| 0.4~0.6   | Moderate correlation           |
| 0.6~0.8   | Strong correlation             |
| 0.8~1     | Extremely strong               |

PCC evaluation method has an important mathematical characteristic, which will not change the coefficient after the position translation of the two vectors. There is a certain degree of error in the application of this characteristic to the distribution network topology verification. If the trend of the function image drawn by two vectors is the same, the PCC will not change, such as the vector X is translated to A + BX and vector Y translated to C + DY, where parameters A, B, C and D are constants, does not change the PCC between them.

Based on this, by improving the PCC method, the error expectation values (EEV) are introduced to avoid the above situation, which greatly improves the accuracy of verification and identification. The calculation formula of EEV is as follows [11].

\[ EEV = \frac{\sum_{i=1}^{n} |X_i - Y_i|}{n} \]  

where \( n \) is the number of elements of vector X and Y.

The closer the EEV of two vectors is to 0, the closer the two vectors are. The improved method can judge the correlation of two vectors together in PCC and EEV, which greatly improves the accuracy of topology verification.

The specific steps of the algorithm (shown in Figure 1) are:

1) Calculate the PCC between users to form a Pearson correlation coefficient matrix.

2) Let the initial value of the counting variable Z be 0. If the PCC of a user with the transformer meter is greater than or equal to 0.8, and the EEV is smaller than 0.2, then the value of Z is increased by 1. When this user is compared with all other users, the final Z is obtained.

3) According to Z and m*n (where m is the fractional threshold, when the m=0.2 is obtained according to a large number of experiments, the algorithm works best, n is the number of user tables),
it is judged whether the user belongs to the same station. If $Z > m \cdot n$, the user belongs to this zone; otherwise, user $i$ does not belong to this zone.

4) Repeat step 2) until all user verifications are complete.

5) The algorithm ends.

The proposed algorithm here is composed by the method used in [14] where the PCC between the voltages of the station and each user, instead of those between users was applied, and the revised PCC method with EEV. The proposed method could overcome the shortcomings that the PCC without EEV cannot check the voltage curves which are transformed by linear transformation.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{flowchart.png}
\caption{The flow chart of the revised PCC algorithm.}
\end{figure}

3.2. Principles and sample sets of the KNN algorithm

The K-Nearest Neighbor (KNN) algorithm is one of the simplest methods in data mining classification technology, and has been widely used in many fields. The core idea of the KNN algorithm is that if a sample has most of the k nearest neighbors in the feature space belonging to a certain category, the sample also belongs to this category and has the characteristics of the samples on this category [15]. The method determines the category to which the sample to be classified belongs based on only the category of the nearest one or several samples in determining the classification decision. The KNN method is only relevant to a very small number of adjacent samples in class decision making. Since the KNN method mainly relies on the surrounding contiguous samples, rather than relying on the discriminant domain method to determine the category, the KNN method is more suitable than the other methods for the crossover or overlapping sample set of the domain.

Due to the large number of users in the power grid, it is very expensive and inefficient to search for K neighbors in the entire sample set; therefore, it is necessary to narrow the search range and improve
the verification efficiency. According to the provisions of 9.6.4 in the Technical Guidelines for Distribution Network Planning and Design [16], 220/380 V lines should have a clear power supply range, and the power supply radius should meet the requirements of the terminal voltage quality. In principle, the power supply radius of A+ and A power supply areas should not exceed 150 m, Class B should not exceed 250 m, Class C should not exceed 400 m, Class D should not exceed 500 m, and the power supply radius of Class E power supply area should be determined according to the requirements (Classes (+A, A, B, C, D, and E) are the power supply areas defined in the “Technical Guideline for Distribution Network Planning and Design”). Therefore, the sample set of the KNN algorithm consists of all users in the zone where the user is located and all users in the adjacent zone. Equation (2) is a formula for calculating the distance between two mesas.

\[
d_{AB} = R \cos^{-1} \left[ \cos \omega A \cos \omega B \times \cos (jA - jB) + \sin \omega A \times \sin \omega B \right]
\]

where the latitude and longitude of transformers A and B are \((\omega A, \omega B)\), \((jA, jB)\), respectively; the radius of the earth is \(R\).

3.3. Algorithm step of the proposed solution

Specific steps are as follows:

1) According to the user voltage data collected by the total meter and the user’s smart meter, select the voltage data (the time is consistent) of the user in the calibration station to form the voltage sequence. The acquisition time of the smart meter is 15 min, and the day is 96 data points.

2) Determine the T value in the Pearson correlation coefficient algorithm based on the selected voltage data amount. If it is 7 days of data, then \(T=7\). Then select a reasonable similarity determination threshold \(P\). When the correlation coefficient is less than \(P\), it can be determined that the voltage curve of the user is not similar to the total table of the station; when the correlation coefficient is greater than or equal to \(P\), it can be determined that the voltage curve of the user is similar to the total table of the station. Compare with the user profile information of the computer system, so that you can find out whether the topology relationship of the station is correct.

3) According to the GIS system data and the “Technical Guidelines for Distribution Network Planning and Design”, use Equation (2) to calculate the adjacent stations of the user to be verified, and then obtain all users in the station area and adjacent stations in the verification user. A set of algorithm samples built by all users.

4) Selecting the \(K\) value of the KNN algorithm to make the algorithm work best (through experiments, the algorithm works best when \(K=7\)); then use the KNN algorithm to analyze the algorithm sample set to find out the correct partition to which the user belongs.

5) The manual field check verifies the household relationship to verify the correctness of the algorithm.

4. Case studies

According to the above-mentioned household relationship verification principle, a 7-day data of the user’s smart meter voltage in the area B to be verified is selected, and then the Pearson correlation coefficient algorithm is used to obtain the correlation coefficient between the user’s smart meter voltages.

The calculation results of the unmodified user voltage similarity algorithm are shown in Figure 2. It can be seen from Table 2 that the correlation coefficient between user 12 and other users is significantly smaller, and the correlation coefficient between other users is relatively high; therefore, it can be determined that user 12 and other users do not belong to the same zone.

The improved user total table voltage similarity algorithm is shown in Table 3. The correlation coefficient between the user 12 and the station total table is obviously low, and it can be judged that the user 12 does not belong to the station area.

After comparative analysis, it can be found that the number of calculations of the unmodified algorithm is \(12 \times 12\) times, and the number of times of the improved algorithm is \(12 \times 1\) times. When the
number of users in a station is large, the improved algorithm is significantly more efficient than the unimproved algorithm.

![Correlation Coefficient Image]

**Figure 2.** Inter user voltage Pearson correlation coefficient.

**Table 2.** Improved algorithm effect diagram.

| User Number | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | 11  | 12  |
|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Area B      | 0.91| 0.93| 0.85| 0.88| 0.81| 0.84| 0.85| 0.92| 0.9  | 0.91| 0.87| 0.53|
| total meter |     |     |     |     |     |     |     |     |     |     |     |     |

It can be judged from Figure 2 or Table 2 that the user 12's household relationship is wrong, and it is necessary to find out the area to which the user 12 belongs.

Firstly, according to the GIS data and Formula (2), the adjacent station area of station B is calculated, and the KNN algorithm sample set of user 12 is constructed. The calculated area of the station area B is 3, which are named as area A and area C respectively. And zone D; there are 10 users in station A, user numbers are A1 to A10, there are 8 users in station C, user numbers are C11 to C18, and there are 5 users in area D, the user number is D19 to D24; these user numbers are the sample set of the KNN algorithm. Then, the KNN sample set composed by the user 12 is analyzed by the KNN algorithm to match the zone that the user 12 is most likely to belong to. The effect chart of the KNN algorithm is organized as shown in Table 3. The first column of the table is the correlation coefficient between the user 12 and the user in the KNN algorithm sample set, and the second column is the user number. Since K=7, only the first seven correlation coefficients are seen. It can be found from Table 3 that among the top 7 correlation coefficients, 4 users belong to the station area D, 1 user belongs to the station area C, and 2 users belong to the station area A. It can be concluded that the user table 12 is most likely to belong to the station area D.

**Table 3.** KNN algorithm sorting diagram.

| correlation coefficient | 0.97 | 0.96 | 0.95 | 0.93 | 0.91 | 0.92 | 0.91 | 0.9 | 0.86 | 0.84 | 0.86 | 0.83 |
|-------------------------|------|------|------|------|------|------|------|-----|------|------|------|------|
| User Number             | A1   | C18  | D20  | A7   | D23  | D24  | D22  | A5  | A3   | D19  | A2   | A6   |
| correlation coefficient | 0.82 | 0.79 | 0.4  | 0.42 | 0.41 | 0.38 | 0.38 | 0.35| 0.33 | 0.32 | 0.31 |      |
| User Number             | D21  | A4   | C15  | A9   | C16  | C12  | A10  | C14 | C11  | C13  | A8   | C17  |

5. Conclusions

This paper proposes a low-voltage distribution network topology verification method based on Pearson correlation coefficient and proximity algorithm for the problem of household relationship and inaccurate file information involved in low-voltage distribution network topology. Firstly, the station
area to be verified is selected, and the Pearson algorithm is analyzed for the user's smart meter voltage in the station area, and the correctness of the household relationship in the area is analyzed. Then, based on the latitude and longitude coordinates of the GIS system area, calculate the neighboring area of the user to be verified, and then complete all the user groups of the neighboring station to complete the training sample set, and finally use the KNN algorithm to analyze the training sample set. Find the best match combination as the correct zoning affiliation. Finally, the operation and maintenance personnel are arranged to go to the site for confirmation and determine the correctness of the results to be verified. The verification method requires low labor cost, fast calculation speed, small calculation amount, strong real-time performance, strong operability, and high application value.

References
[1] Neubauer A and Niederholz J 2009 Method and device for calculating zero-crossing reference sequences for signal detection of angle-modulated signals based on zero crossings of the received signal: U.S. Patent 7, 564, 926[P]. 2009-7-21
[2] Matthewman S and Byrd H 2014 Blackouts: a sociology of electrical power failure[J].
[3] Slonim N, Atwal G S, Tkacik G, et al. 2005 Estimating mutual information and multi-information in large networks[J]. arXiv preprint cs/0502017
[4] Lehr A, Atias I, Korcharz D, et al. 2006 Data communication network providing power over network connections with node identification functionality: U.S. Patent 6, 985, 713[P]. 2006-1-10
[5] Ying J, Chen L and Liu M 2018 Research on Check Method of Feeder Topology Model for Distribution Main Station[C]. 2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2) IEEE 1-5
[6] Zhou D, Zhang B, Yao F, et al. 2006 Fast search for transmission section based on graph theory[C]. Zhongguo Dianji Gongcheng Xuebao (Proceedings of the Chinese Society of Electrical Engineering) 26(12) 32-38
[7] Arghandeh R, Gahr M, von Meier A, et al. 2015 Topology detection in microgrids with micro-synchrophasors[C]. 2015 IEEE Power & Energy Society General Meeting IEEE 1-5
[8] Wang J, Ji X, Li K, et al. 2019 Topology Verification of Low Voltage Distribution Network Based on k-means Clustering Algorithm[C]. IOP Conference Series: Materials Science and Engineering IOP Publishing 569(5) 052096
[9] Deka D, Backhaus S and Cherikov M 2017 Structure learning in power distribution networks[J]. IEEE Transactions on Control of Network Systems 5(3) 1061-1074
[10] Sheugh L and Alizadeh S H 2015 A note on pearson correlation coefficient as a metric of similarity in recommender system[C]. 2015 AI & Robotics (IRANOPEN) IEEE 1-6
[11] Jiang Haiyang, Li Qian, Li Tianqi, Shi Xiuping, Wei Ping, Yang Cheng and Huang Yuanchao 2019 Identification of Key Parameters of Primary Frequency Modulation Based on Improved Pearson Product Moment Correlation Coefficient[J]. M&E Engineering Technology 48(07) 180-184
[12] Luan Wenpeng, Yu Yixin and Gong Bing 2015 AMI data analytics[J]. Proceedings of the CSEE 35(1) 29-36
[13] Luan W, Peng J, Maras M, et al. 2015 Smart meter data analytics for distribution network connectivity verification[J]. IEEE Transactions on Smart Grid 6(4) 1964-1971
[14] Xiao Yong, Zhao Yun, Tu Zhidong, Qian Bin and Chang Runmian 2019 Topology checking method for low voltage distribution network based on improved Pearson correlation coefficient[J]. Power System Protection and Control 47(11) 37-43
[15] Bi Xiaomei and Bi Ruihua 2009 KNN algorithm summary[J]. Science and Technology Innovation Guide 2009(14) 31
[16] Zhao Mingxin, Liu Wei, Chen Hai, et al. 2016 Interpretation of technical guidelines for distribution network planning and design[J]. Power supply 33(2) 2-7