frontiers in Energy Research

BRIEF RESEARCH REPORT
published: 27 April 2021
doi: 10.3389/fenrg.2021.674865

A Power Customer Data Relational Algorithm Based on Magnanimity Fuzzy Address Matching

Peng Jin, Jing Yang, Zongwei Wang, Xiaoyang Bu and Peng Wu*

State Grid Customer Service Center, Tianjin, China

According to the short text and unstructured characteristics of customer address, a data association fusion method for address has been proposed. In this method, the address was mapped to a digital fingerprint by improved Simhash technology, which effectively reduced the dimension of massive addresses and simplified the similarity-matching process of multi-source heterogeneous addresses. Furthermore, the weight setting of the eigenvector of the simhash algorithm was improved by introducing special weight gain. A two-level index mechanism was established by the characteristics of address division and data structure of digital fingerprints; the time-consuming digital fingerprint comparison was greatly reduced. The experimental results showed that calculation efficiency was greatly optimized; accuracy and coverage of the comparison were ensured. Through address matching of different databases, information fusion can be completed and the goal which power customers' demands is connected to power grid equipment is achieved.

Keywords: improved simhash, multi-source heterogeneous data, address matching, data associations, digital fingerprint, data of electric client

OPEN ACCESS

Edited by:
Liang Chen,
Nanjing University of Information Science and Technology, China

Reviewed by:
Jie Sheng,
Soochow University, China
Jianzhong Xu,
North China Electric Power University, China

****Correspondence:***
Peng Jin
470292065@qq.com

Specialty section:
This article was submitted to Smart Grids, a section of the journal Frontiers in Energy Research

Received: *02 March 2021*
Accepted: *24 March 2021*
Published: *27 April 2021*

Citation:
Jin P, Yang J, Wang Z, Bu X and Wu P (2021) A Power Customer Data Relational Algorithm Based on Magnanimity Fuzzy Address Matching. Front. Energy Res. 9:674865. doi: 10.3389/fenrg.2021.674865

INTRODUCTION

With the deepening of electric power reform, grid enterprises have gradually begun to establish a modern customer-centered service mode in recent years. They were eager to explore the potential demands of customers to support the development of new formats and optimize the allocation of service resources. However, the characteristics of scattered and massive customer information bring great challenges to grid enterprises to carry out customer behavior mining and accurate service.

Big data technology, as a means to efficiently process massive data with complex sources, has brought revolutionary changes and has made the association analysis possible. Many cases have proved that the value will be brought into full play by integrating data in different fields, specialties, and channels (Wang et al., 2018; Shen et al., 2019; Zhou et al., 2020). Large grid enterprises, which set up big data organizations and big data platforms, have taken the initiative to carry out digital transformation (Sun, 2019) and promote the integration of grid business and customer electricity behavior information rapidly (Teeraratkul et al., 2018; Wang et al., 2019; Li et al., 2021).

Unlike banks or other financial service industries, customers rarely provide information, such as name and power electricity number, in the process of power-related services; power grid companies must carry out emergency repair services according to the address fed back by customers. Because of the fragmentation and unstructured characteristics of the information provided by the customers (Song, 2013), there are great differences between the information provided by the customers and the electric power standardized data in the same entity expression (Wang, 2012). In the process of

fusion with power grid data, it is necessary to make pair-wise comparison of all entities between the customer information database and the power marketing database. Because of the diversity and heterogeneity of linked data (Xie et al., 2015; Wu et al., 2016), there are problems of low coverage and precision in association matching. In addition, the amount of customer information is massive, and the calculation complexity, which achieves data alignment of similar address entities among different databases, is high (Shen and Feng, 2018; Kang et al., 2019).

In order to realize the information fusion between two databases, it is generally necessary to rely on the reliable similarity function (Liu et al., 2017). The comparison texts were decomposed into a set of tokens, and then, the token set was transformed into an n-dimensional vector; the similarity of comparison texts was evaluated by calculating the cosine similarity between vectors. The order of the token was not considered in this method, but the quality of tokens seriously affected the accuracy of comparison (Ye, 2011). At present, the mainstream algorithms mainly use word segmentation technology to form tokens, but segmentation ambiguity has a great impact on the results. In order to solve the problems caused by word segmentation, the q-gram similarity function was proposed (Sun and Wang, 2014). Since the length of the substring was fixed, there was overlap between the tokens, which could effectively avoid the problem caused by segmentation divergence. However, the q-gram method will greatly increase the amount of calculation, where q is generally <4. In the minimum edit distance method (Belazzougui and Venturini, 2016), it is not necessary to segment the word vector; this method transformed the target text into the comparison text by inserting, deleting, and exchanging operations. The operation cost was used as the similarity function, and the difference between matched texts could be better quantified. In addition, this method was not sensitive to local missing characters. Unstructured text entities were more likely to be converted into structured data for comparison, so the local sensitive hashing algorithm was proposed (Can et al., 2017), which could map the text data to a fixed-length fingerprint set through a special hash function. Compared with matching the text entity, the comparison of a fingerprint could greatly reduce the complexity of data storage and calculation and provide an effective choice for mass data comparison. In addition, the computational complexity of matching text entity was proportional to the quadratic power of the entity size in the database (Zhuang et al., 2016). For the entity alignment of massive data, the computing resources and computational efficiency were unbearable, and further dimensionality reduction must be carried out. Partitioned indexes (Qu et al., 2018) were used to reduce computational complexity, and the entities with the same key value were placed in the same block. The similarity comparison was only carried out in the same block, and it was no comparison between blocks, which should greatly reduce the amount of calculation.

At present, there are few cases of data fusion through the address in the field of the electric power industry. Therefore, this study proposes a data fusion method based on the improved simhash algorithm, which considers the short text characteristics of the address comprehensively. Through the fuzzy address-matching technology, customer behavior data can be associated with data of power grid equipment. Furthermore, through the two-level partition index mechanism, the dimensionality reduction of comparison is realized, which can support the massive data matching of more than 100 million levels.

ADDRESS FUZZY MATCHING ASSOCIATION MODEL BASED ON IMPROVED SIMHASH ALGORITHM

Association Model Based on Address Data

State Grid Corporation of China (SGCC) has constructed complete file information of customers, metering devices, and power grid equipment, which can be linked through a unique power consumption number. Customers only provide the detailed address of the fault when calling the customer service hotline but cannot provide valuable information such as the power consumption number. Therefore, the information provided by power customers is difficult to directly relate to the power supply point, which has a negative impact on fault research and data analysis.

Figure 1 shows the data model structure. The work order of the customer service can be associated with power grid equipment only through customer file information and metering point information. Besides the customer electricity number, the site address of the work order can also be associated with the customer file information reservation. However, the above data are distributed in different data fields. In order to associate the data through the address, the address data must be standardized to improve the coverage of data matching.

Standardization of Customer Address

At present, the site address of SGCC repair work order is divided into two parts: one is structured data, including the information of provincial-, municipal- and county-level power supply units; the other part is unstructured data, covering the string information of village, town, street, road, community, and door combination.

Affected by the accuracy of the expression and local dialects of customers, some addresses may have some deviation. The first step is to remove invalid information, such as separators, spaces, and English letters in the address, and convert the Arabic numerals information into Chinese characters; the second step is to filter and remove the house number and other information in the text to achieve the comparison of address dimensions, such as community, village, and street; and the third step is to convert the cleaned address into Chinese pinyin.

Principle of Simhash Algorithm

The simhash algorithm was proposed by Charikar in 2002 (Naumann, 2010). As a digital fingerprint algorithm, it aimed to solve the problem of removing the duplicate of massive web pages. Due to digit limitation of a fingerprint, the matching objects could be detected quickly in a large-scale database. So the data dimension was compressed, and the number of similarity comparison was greatly reduced. In addition, local sensitivity

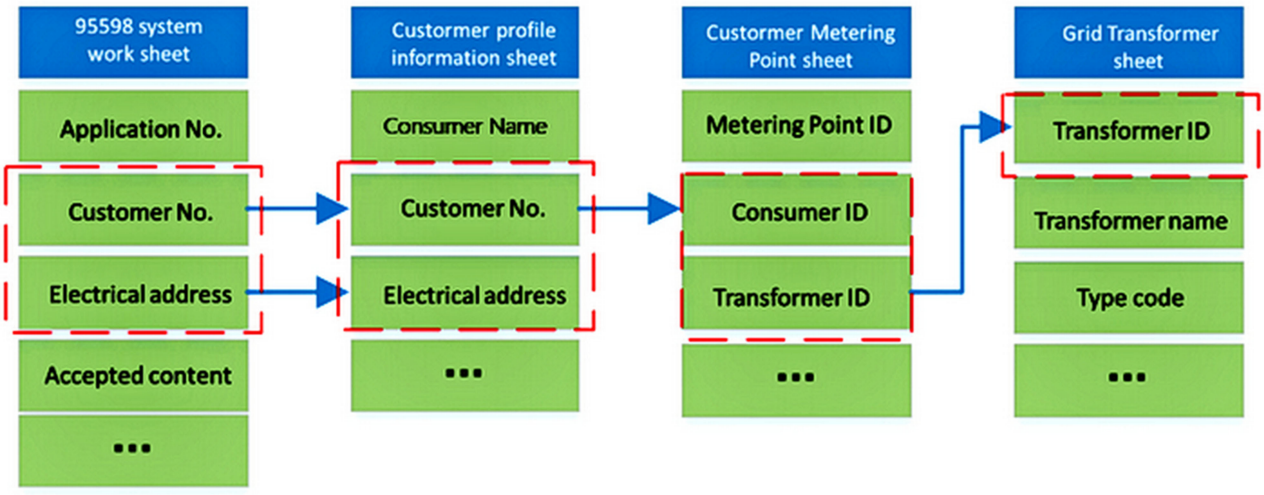

FIGURE 1 | Model of customer demands associated with power grid equipment.

was demonstrated in this method, and the similarity of any two entities was proportional to the similarity of the generated fingerprint, which helped to quantify the differences between the contrast entities.

The core idea of the simhash algorithm is to map the unstructured text set to the unique signature value generated from the original data. The generated digital fingerprint is a series of fixed-length binary codes, as shown in **Figure 2**. The whole process can be divided into four steps: text feature extraction, fingerprint generation, fingerprint index, and matching calculation.

By using the traditional MD5 hash technology, the word vector is mapped to the digital fingerprint and weighted according to the given word vector.

$$f(hash_{pv}) = \begin{cases} weight_p, & hash_{pv} = 1 \\ -weight_p, & hash_{pv} = 0 \end{cases} \quad (1)$$

where $weight_p$ is the weight of the p-th word vector; $Hash_{pv}$ is the value of the v-th bit of the hash fingerprint that the p-th word vector is mapped.

Then, the weighted values of all word vectors in the text are accumulated and merged to form a new sequence T_j. The dimension is reduced to form the final simhash digital fingerprint according to Equation (2).

$$Simhash_j = \begin{cases} 1, & T_j > 0 \\ 0, & T_j < 0 \end{cases} \quad (2)$$

After the batch formation of digital fingerprints, the similarity between fingerprints needs to be further judged. Generally, the similarity between digital fingerprints is measured by hamming distance. There are two binary strings, **x** and **y,** of length n. The hamming distance between them can be calculated as follows:

$$Hamming(\mathbf{x}, \mathbf{y}) = \sum_{1}^{n} (x_i \oplus y_i) \quad (3)$$

where $\oplus$ is XOR operation. Different digits of digital fingerprints can be calculated by hamming distance. Generally, the higher the coincidence degree of two text sets, the higher the similarity between digital fingerprints. If two 64-bit binary strings have less than three different characters, they can be regarded as similar text sets.

Improved Simhash Algorithm Considering Address Characteristics

There are no predicate, attribute, adverbial, and complement in Chinese addresses. The average length of the addresses of 1 million customers is 19.7 characters. Generally, the length of an address is <45 characters, which basically follows the normal distribution. The number of address characters is relatively small, and the traditional word vector is not used as the simhash feature vector. Instead, the character is used as the feature vector to avoid the misjudgment and the time cost caused by word segmentation.

The structure of an address is basically extended from large administrative divisions to fine addresses with strong regularity. The accuracy of fine address matching often determines the matching of the whole address. In order to improve the resolution accuracy of a terminal fine address and the weight component of a fine address, the weight of the simhash feature vector is adjusted according to the following Equation (4):

$$K_{\mathrm{i}} = \lfloor i^k \rfloor (i \in [1, addr_{len}]) \quad (4)$$

where K_i is the weight gain of the improved eigenvector, $addr_{\mathrm{len}}$ is the word length of the address text, i is the character bit of the eigenvector, k is the power exponent parameter, and $\lfloor\rfloor$ represents the down rounding function.

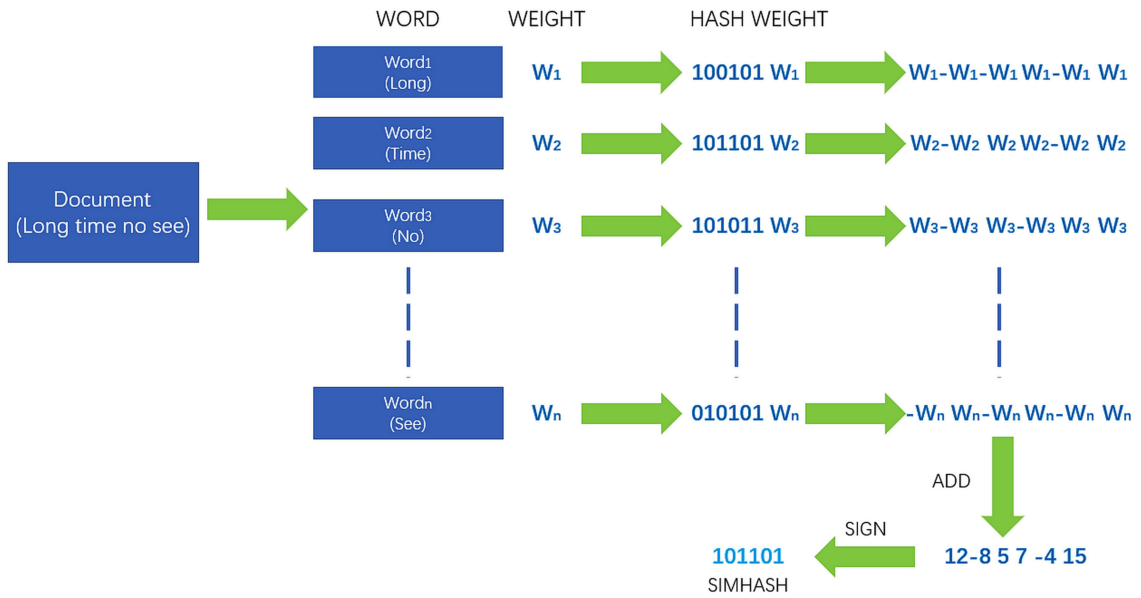

FIGURE 2 | Simhash algorithm flow.

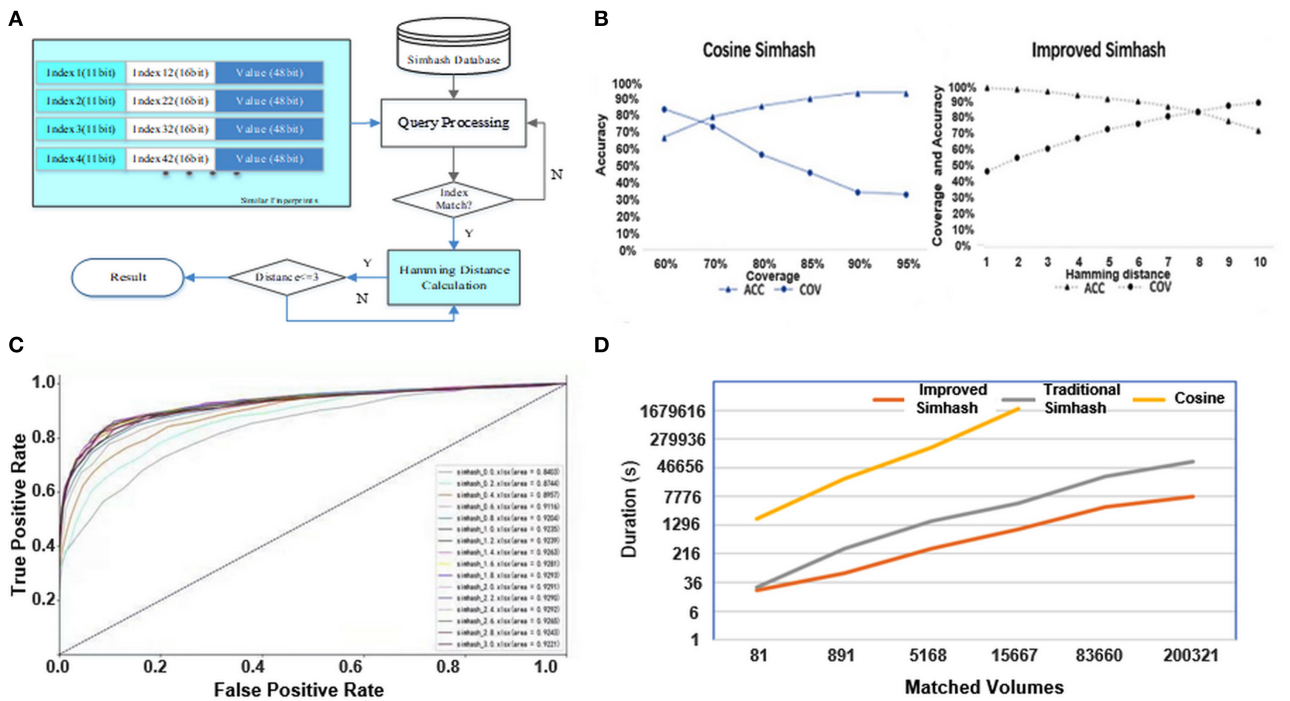

FIGURE 3 | (A) Hierarchical index architecture. **(B)** Algorithm performance evaluation**. (C)** Influence of parameter change on ROC curve. **(D)** Efficiency comparison of different methods.

Segmented Index Method Based on Improved Simhash Algorithm

Text information dimensionality reduction can be achieved by Simhash technology, but the workload of text comparison between the two types of data is very huge. It is necessary to compress the comparison amount through the segmented index method to support the ability of massive data processing. In the address structure, the detailed address text is unstructured, while the addresses of provinces, cities, districts, and counties are usually structured.

TABLE 1 | Comparison of data classifying quality.

Indicators / Hamming distance		1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)
The proposed method	Accuracy	99.07	98.05	96.86	94.29	92.34	90.17
	Coverage	47.35	55.53	61.39	67.59	73.34	75.01
The traditional simhash method	Accuracy	98.62	95.67	89.95	85.41	77.89	70.76
	Coverage	15.82	24.45	37.61	52.43	66.26	72.43

Therefore, the index mechanism is established based on the dimension of a district and a county, and the simhash comparison is limited to the same district and the county, which greatly reduces the amount of unnecessary comparison. At present, there are more than 1,400 county-level power supply units within the scope of SGCC. The resolution of the 11-bit binary code is 2,048, so all district- and county-level units can be indexed by the 11-bit binary code.

If the hamming distance is <3, two strings, which are converted to the 64-bit digital fingerprint, are considered to be similar. According to the drawer principle, if two similar texts are serialized into 64-bit digital fingerprints, there must be 16 bits that are exactly the same. Therefore, the same 16-bit binary code can be used as the index; the actual comparison bit will be reduced to 48 bits.

By combining the above two-tier index construction methods, the matching workload will be greatly reduced. In addition, the repeated calculation can be avoided effectively when the new address is added to the database, as shown in **Figure 3**.

EXAMPLE ANALYSIS

Sample Selection and Experimental Environment

In order to prove the address matching verification, the standard addresses of power customer file information and the address of customer demand work order are processed centrally as the basic data for comparison. Address processing includes eliminating a house number and other detailed information.

The algorithm test is based on the big data platform, with 10 physical machines, which are dual-channel 12 core Xeon e5-2650 V4 processors (2.2 GHz). The Hadoop cluster of this platform has 24 nodes, 1.6T memory, and 235T physical storage.

Experimental Results

In the simhash algorithm, the more the same number of the digital fingerprint, the more similar. But the higher the similarity, the smaller the hamming distance, which leads to lower coverage in similar matching. The performance of the proposed method is further analyzed by comparing the traditional simhash algorithm. The test results are as shown in **Table 1**. The test results are as follows:

With the same accuracy, the coverage of the proposed method is better than the traditional simhash method. In the case of short text, the improvement is more obvious, and the accuracy and the coverage are more than 80%.

In addition, compared with the cosine similarity comparison method proposed in (Ye, 2011), the improved simhash method has obvious advantages in accuracy and coverage, as shown in **Figure 3B**.

In the cosine similarity comparison method, the feature vector is determined by the word segmentation technology. Once the unknown words appear in the address, it is easy to produce segmentation divergence, which greatly affects the accuracy and success rate. In the improved simhash method, the feature vectors are determined by characters to avoid the problems that may be caused by word segmentation, so the matching accuracy and the coverage are improved.

The concept of weight gain is proposed in (Sun, 2019). In order to get a more suitable weight gain, we traversed the k value between [0.2, 3] and further analyzed the influence of k value change on accuracy and coverage. As can be seen from the **receiver operating characteristic** (ROC) curve in **Figure 3C**, with the increase of k value, the weight of the terminal address, such as village, town, and cell name, also increased. In addition, the area under the ROC curve continued to increase, and the generalization ability gradually improved. When $k = 1.8$, the area under the ROC curve reached the maximum (AUC = 0.9294). Then, the generalization ability of the model decreased with the increase of k.

Cosine similarity, traditional simhash, and improved simhash are tested by matching the address of the village and the county level, and the specific time consumption is shown in **Figure 3D**. The cosine similarity method contains word segmentation, cosine calculation, and other steps, and dimensionality reduction cannot be realized. When the amount of calculation increases greatly, the results may not be calculated.

When the number of matching jobs is low, the efficiency of traditional simhash and improved simhash is almost the same. With the increase of the number of matching, the efficiency advantage of the improved simhash method appears. The main reason is that the traditional simhash method needs to segment the address, and the time consumption of text-dimension reduction is higher than the improved simhash method.

MAIN CONCLUSIONS

An address-matching method based on the improved simhash algorithm is proposed to realize the association between

unstructured addresses. There are two innovations in this method: first, the traditional simhash algorithm is optimized by improving the weight gain of simhash and adjusting the text vector composition so as to ensure the coverage and accuracy of address matching; second, according to the characteristics of the hash fingerprint and address text, a two-level index mechanism is constructed to reduce the complexity of the address-matching algorithm and improve the efficiency of data fusion. The mechanism solves massive address-matching problems and helps to link up the non-standard addresses reflected by power customers to power grid equipment so as to enhance the application value of power customer data.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

PJ led the analysis and wrote the manuscript. JY performed the experiment. ZW carried out the study and collected important background information. XB contributed to the research of algorithm concept. PW helped perform the analysis with constructive discussions. All authors contributed to the article and approved the submitted version.

REFERENCES

Belazzougui, D., and Venturini, R. (2016). Compressed string dictionary search with edit distance one. *Algorithmica* 74, 1099–1122. doi: 10.1007/s00453-015-9990-0

Can, L., Qian, J., and Dong, Y. (2017). M2LSH:an LSH based technique for approximate nearset neighbor searching on high dimensional data. *Acta Electron. Sin.* 45, 1431–1442. doi: 10.3969/j.issn.0372-2112.2017.06.022

Kang, S., Ji, L., Liu, S., and Ding, Y. (2019). Cross-lingual entity alignment model based on the similarities of entity descriptions and knowledge embeddings. *Acta Electron. Sin*. 47, 1841–1847.

Li, Y., Wang, C., Li, G., and Chen, C. (2021). Optimal scheduling of integrated demand response-enabled integrated energy systems with uncertain renewable generations: a Stackelberg game approach. *Energy Convers. Manage.* 235:113996. doi: 10.1016/j.enconman.2021.113996

Liu, Z., Chen, J., Zheng, J., Hua, J., and Xiao, L. (2017). Research on aggregation model for chinese short texts. *J. Softw*. 28, 2674–2692. doi: 10.13328/j.cnki.jos.005147

Naumann, F. (2010). *An Introduction to Duplicate Detection*. San Rafael, CA: Morgan and Claypool.

Qu, Z., Fan, M., Zhou, R., Wang, H., and Zhu, D. (2018). Inverted index query technique of non-primary key for mass dispatch and monitoring information of distribution network. *Power Syst. Protection Control.* 46, 162–168. doi: 10.7667/PSPC171742

Shen, B., and Feng, J. (2018). Crowdsourcing knowledge base index alignment. *Chin. J. Comput*. 41, 1814–1826. doi: 10.11897/SP.J.1016.2018.01814

Shen, J., Cao, R., Su, C., Cheng, C., Li, X., Wu, Y., et al. (2019). Big data platform architecture and key techniques of power generation scheduling for hydro-thermal-wind-solar hybrid system. *Proc. CSEE* 39, 43–55+319.

Song, Z. (2013). Address matching algorithm based on chinese natural language understanding. *J. Remote Sens*. 17, 788–801. doi: 10.11834/jrs.20132164

Sun, D., and Wang, X. (2014). Q-gram index for approximate string matching with multi-seeds. *Comput. Sci*. 41, 279–284. doi: 10.11896/j.issn.1002-137X.2014.09.053

Sun, Y. (2019). Accelerate digital transformation to create a new pattern of coordinated development. *Energy Res. Utilization* 1, 4–5+7. doi: 10.16404/j.cnki.issn1001-5523.2019.03.001

Teeraratkul, T., O'Neill, D., and Lall, S. (2018). Shape-based approach to household electric load curve clustering and prediction. *IEEE Trans. Smart Grid* 9, 5196–5206. doi: 10.1109/TSG.2017.2683461

Wang, D. (2012). Power data center infrastructure based on cloud computing and its key technologies. *Automation Electric Power Syst*. 36, 67–71+107.

Wang, Q., Li, F., Tang, Y., and Xue, Y. (2018). On-line prediction method of transient frequency characteristics for power grid based on physical-statistical model. *Automation Electric Power Systems*. 42, 1–11. doi: 10.7500/AEPS20171001001

Wang, Y., Zhang, N., Kang, C., Xi, W., and Huo, M. (2019). Electrical consumer behavior model:basic concept and research framework. *Trans. China Electrotech. Soc*. 34, 2056–2068. doi: 10.19595/j.cnki.1000-6753.tces.190073

Wu, Q., Gao, J., Hou, G., Han, B., Wang, K., and Li, G. (2016). Short-term load forecasting support vector machine algorithm based on multi-source heterogeneous fusion of load factors. *Automation Electric Power Syst*. 40, 67–72+92. doi: 10.7500/AEPS20160229012

Xie, G., Hu, Y., Chen, J., Yu, N., and Zhou, H. (2015). A fusion method for multi-source and heterogeneous parameters of power grid and its engineering application. *Automation Electric Power Syst.* 39, 121–127. doi: 10.7500/AEPS20140424004

Ye, J. (2011). Cosine similarity measures for intuitionistic fuzzy sets and their applications. *Math. Comput. Model.* 53:91–97. doi: 10.1016/j.mcm.2010.07.022

Zhou, F., Zhou, H., and Diao, Y. (2020). Development of intelligent perception key technology in the ubiquitous internet of things in electricity. *Proc. CSEE* 40, 70–82+375. doi: 10.13334/j.0258-8013.pcsee.191198

Zhuang, Y., Li, G., and Feng, J. (2016). A survey on entity alignment of knowledge base. *Comput. Res. Dev*. 53, 165–192. doi: 10.7544/issn1000-1239.2016.20150661

Conflict of Interest: PJ, JY, PW, ZW, and XB are employed by the same company State Grid Customer Service Center.