Jestr

Journal of Engineering Science and Technology Review 16 (1) (2023) 109 - 114

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

An Identification Method of Power User Group Based on Big Data Analysis

Xinguo Li[1], Haiyan Zeng[1], Hongsheng Chen[1], Xiao Li[1] and Ming Liu[1,*]

[1]*State Grid Wuhan Power Supply Company, Wuhan 430000, China*

Received 15 November 2022; Accepted 22 January 2023

Abstract

Power user group is the main target of power industry management, and identifying different types of power user groups has become a research hotspot in the power industry. Power user group identification is proposed based on big data analysis to effectively protect the interests of power enterprises. Initially, the web crawler technology was used to calculate the correlation between the topics to be crawled and the keywords on the web page, and then the behavior of big data of the power user group on the power portal was extracted according to the correlation. The outliers of discrete big data were filtered, considering the consistency of big data of power user group behavior. Then, an energy field model is built to calculate the apparent energy, implicit energy, and active coefficient of power users. The active group, ordinary group, and zombie group of power users were obtained, according to the apparent energy and hidden energy of power users and the active coefficient. Results show that the method has good data crawling ability for power user groups prior to the contrast outlier filtering. It can effectively identify different types of power user groups, that is, 118 active users can be identified, with high recognition accuracy, which can be close to 100%. Conclusions provide guiding services for the management of the power industry.

Keywords: Big data, Power user group, Web crawler, Abnormal value filtering, Active coefficient

1. Introduction

Big data is a type of information asset. In the massive big data, users can gain insight into the valuable information covered by the data analysis, decision making, and other approaches to provide guidance for their industry or field. Network information is overwhelming and growing exponentially with the introduction of the cloud computing era, forming a massive amount of data, including the power industry. The development of electric power industry is related to the national economy and people's livelihood and is closely related to the national GDP. As the main modern energy, electric power plays an irreplaceable role in people's production and life. Power enterprises obtain benefits from the electric power sales. In the process of power sales, power enterprises consider the big data of electricity meter measurement as the charging standard. The identification accuracy of electric power user groups is very important to ensure the interests of power enterprises. However, in the process of identification, accuracy is difficult to ensure; thus, valuable data should be from the big data of the power industry [1]. This approach is conductive to the development and digital transformation of the power industry, which can better serve the society.

Every person or unit belongs to the power industry user given that the power gird is the foundation to support people's production, business activities, and life, leading to the diversity of power user types and the complexity of user behavior and generating diversified big data in the power industry. The big data in the power industry is large [2], thereby causing great difficulty to the management and service of the power industry.

The power industry adopts different service methods to promote its digital transformation according to different user groups, to simplify its management mode, ensure its service quality, and implement its political policies. Therefore, the identification method of power user group is greatly important. Based on the big data of power, this study investigates the identification method of power user group and provides guiding services for the management of the power industry.

2. State of the Art

In recent years, many scholars have studied the identification methods of user group, and its application has also played a significant role in different fields. For example, Zuo et al. [3] proposed network identification methods of the user group. After collecting the big data of the user group from network, the identification method used the differential cross-correlation algorithm to calculate the big data of the user group to obtain the difference value of the big data from different user groups, and then identify the user groups based on the difference. However, the accuracy of this method in identifying user groups is poor due to the interference noise and redundant data. Zhang et al. [4] proposed the identification method of user identity based on preference logic. This method collected user preference data, analyzed preferences of different users, and then divided user groups according to user preferences. This method was widely used in the field of commodity and data recommendation, but it was not universal. Liang et al. [5] put forward an online user group identification method. This method collected online big data of users' network behavior, analyzed user behavior, and constructed an evaluation index system. Then, they used fuzzy hierarchical analysis method to analyze the correlation degree between users and obtained

*E-mail address: s_20202022@163.com
ISSN: 1791-2377
doi:10.25103/jestr.161.14

the same type of user groups according correlation degree. However, this method was affected by the parameter selection of fuzzy hierarchy analysis method in the application process, thereby leading to poor accuracy of user group identification. Guo et al. [6] predicted the total supply curve of the power market based on the LSTM model, constructed the LSTM model, conducted parallel training on the temporal and non-temporal characteristics of the power market, and then used the power load data for sample analysis. This method could improve the quality of power service. Zhou et al. [7] implemented a study on the bidding strategy of integrated energy selling e-commerce under the influence of green power certificate trading and transmission obstruction. First, according to the quota system and the relevant policies of green power certificate, the optimal mode of two markets in which e-commerce sellers participate in bidding was constructed. The random forest regression algorithm and multi-scenario reduction technique were used to combine the wind power output prediction results with the error scenarios, and then stimulate the multi-scenario actual wind power output. On this basis, it established a mathematical model to minimize the blocking cost of power grid based on node price. The mathematical model and calculation method established in this study are proven accurate and effective through calculation by examples. Although the above studies have made certain progress, studying more accurate user group identification methods in the face of the defects of current user group identification technology is imperative.

3. Methodology

3.1 Big data collection of power user behavior based on topic correlation

Web Crawler is a technology that extracts a script or program of Internet data according to certain rules. Web Crawler can connect the entire Internet and read and save the content of each web page. In this study, Web Crawler is used to crawl the big data of power user behavior on the power portal website. The detailed process is discussed in the following section.

The power portal website has many functions and web pages. Thus, the topic of the web crawler technology should be set initially when using the technology to crawl the big data of power user group behavior to avoid wasting the resources and network bandwidth of the power portal website and reduce the time to crawl the big data of the power user group behavior. The web crawler obtains the optimal Uniform Resource Locator (URL) [8, 9] based on the topic. URL is describes the location of information on the network service program, similar to the function of coordinates. *TF·IDF* refers to the weight of the topic key works to be crawled by the web crawler, among which *TF* refers to the frequency of topic keywords in the web crawler process, and the calculation formula of this frequency is as follows:

$$TF_i = d_j \cdot t_i \frac{Q_{i,j}}{\sum_k Q_{k,j}} \tag{1}$$

where $Q_{i,j}$ refers to the occurrence times of topic words; $\sum_k Q_{k,j}$ refers to the total number of all topic words in the behavior data of the power user group; d_j and t_i refer to the webpages, where the behavior data of the power user group and topic words, respectively, are located; i refers to the i^{th} key words; j refers to the j^{th} webpage; k refers to the total number of subject terms in the k^{th} page.

In the weight of the topic keywords to be crawled by the web crawler, *TF·IDF* refers to the frequency of the web crawler crawling the behavior data of the power user group, and its calculation formula is as follows:

$$IDF_i = \log \frac{|\varphi|}{\left|\{j : t_i \in d_j\}\right| + 1} \tag{2}$$

Where $|\varphi|$ refers to the total number of behavior data text of power user group crawled by the web crawler, and $\left|\{j : t_i \in d_j\}\right|$ refers to the total number of texts containing topic words.

On the basis of the results of Formula (1) and Formula (2), the weight of the behavior big data of the power user group to be crawled is calculated [10], and the expression formula is as follows:

$$w_i = TF_i \cdot ID_i \tag{3}$$

where w_i refers to the weight of the behavior big data of the power user group of the i^{th} keyword to be crawled.

The calculation results in Formula (3) are arranged in descending order, and the first N weights in the big data weight sequence of power user group behavior are considered big data keywords of the power user group behavior.

For the big data keyword of the i^{th} power user group behavior, the spatial vector model is used to calculate the topic correlation when the web crawler crawls the big data of the power user group behavior. n is the dimension of the spatial vector model. After calculating the behavior big data keyword weight of the i^{th} power user group w_i by using Formula (3), the weight is the vector space dimension value in the spatial vector model [11-13]. The crawling topic of the power user group behavior big data is set as q, and the keywork weight in the setting topic is represented by $f_{i,q}$. Thus, the topic expression formula when the web crawler crawls the behavior big data of the power user group is as follows:

$$\partial = \sum_{k \in q} f_{i,q} = (w_1, w_2, \cdots, w_n) \tag{4}$$

Where ∂ refers to the topic when the web crawler crawls the big data of the power user group behavior.

j refers to the web page of the power portal website, x_i=1 refers to the frequency of high-frequency keywords, and $x_i w_i$ is the vector dimension value of the spatial vector model. The topic expression formula of the webpage of the power portal website is as follows:

$$\beta = \sum_{k \in j} f_{k,j} = (x_1 w_1, x_2 w_2, \cdots, x_n w_n) \tag{5}$$

According to the results of Formula (4) and Formula (5), the correlation between webpage j and topic q is calculated, and its expression formula is as follows:

$$Sim(j,q) = \cos < \partial, \beta > \tag{6}$$

Where *Sim(j,q)* refers to the correlation of webpage *j* and topic *q*.

The threshold of the webpage and topic is set to 0.7; when *Sim(j,q)* is ≥0.7, a correlation is found between the web page and the topic [14].

Based on the correlation of the webpage and topic, the process of using web crawler technology to crawl the behavior big data of power user group is as shown in Figure 1. When the web crawler program is started, it initially crawls the topic of the big data behavior of the power user group to be crawled by the web crawler, and then the correlation between the topic and keywords in the web pages of the power portal website is analyzed. If the topic is unrelated to the keyword, then it changes the webpage and then recalculates the correlation; otherwise, it reads the RUL link, and then determines whether the link has been accessed. If the link has been accessed, then the RUL link is stored in the accessed queue; otherwise, it is stored in the access queue. After reading the RUL in the RUL link in the queue to be accessed, the web page shall be parsed and processed. Then, the HTTP request is sent to the user, and the web page parsing content is stored to complete the big data collection of the power user group behavior.

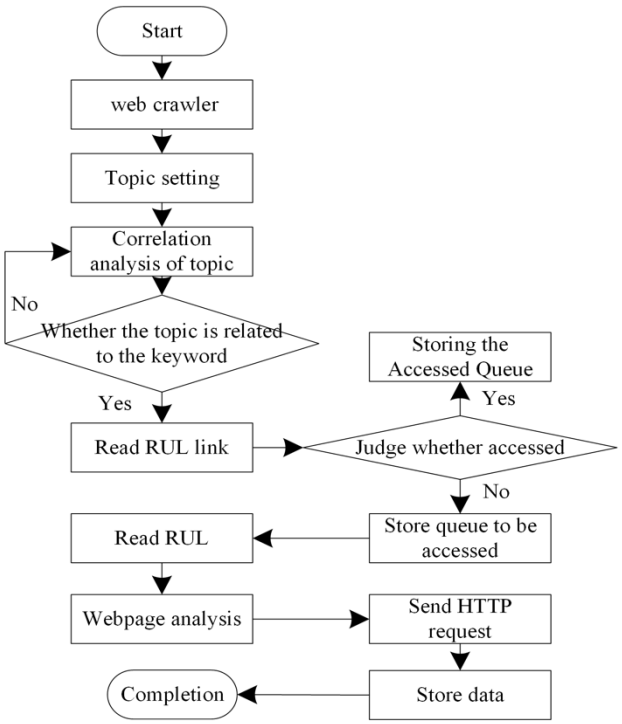


Fig.1. Crawling process of power user group behavior data by web crawler

3.2 Outlier filtering of the big data of the power user behavior

The big data of the power user group behavior cover a large amount, and some are missing and redundant [15]. Thus, the collected big data of the power user group behavior should be processed by filtering outliers. The outlier detection method [16,17] is a big data analysis algorithm, which analyzes the distribution position of data to determine whether the data deviate from other data and whether they are redundant or missing. Here, the outlier detection method is used to filter outliers in the big data of the power user behavior. The detailed process is discussed in the following section.

This study describes the big data of the power user behavior in the form of field values to obtain the discrete outlier data in the big data of power user behavior. The discrete and isolated big data of the power user behavior are obtained by calculating the consistency of the field values.

O refers to the big data set of the power user behavior obtained using web crawler technology, A_k refers to the big data field of the power user behavior in the big data set, and the k^{th} field value is represented by V_{kk}; thus, the consistency of the big data field of power user behavior is obtained using Formula (7), as follows:

$$u_R(V_{kk'}) = O \cdot \frac{2}{1+\frac{D_k}{TV_k \cdot N_{kk'}}} \tag{7}$$

Where $u_R(V_{kk'})$ refers to the consistency of big data field of discrete power user behavior, D_k refers to the total number of records when the field A_k is not null, $N_{kk'}$ indicates the occurrences of $V_{kk'}$, and TV_k is the total number of different values of A_k. In this formula, when the number of occurrences of the big data field value of a certain power user behavior is much lower than the average of all big data, the discrete consistency of the big data of the power user behavior approaches 0. After obtaining the consistency of big data field values of all power user behavior by using Formula (7), whether the field value is an outlier α is confirmed. When $u_R(V_{ij})$ value is less than α, the data corresponding to the field value of the current power user behavior big data is an outlier. After filtering the outlier data, the big data of the power user behavior without missing or redundant data are obtained; they are represented by *O'*. Thus, the outlier filtering of big data of power user behavior is completed.

3.3 Identification of power user group based on energy field model

The big data of the power user group behavior contain three data types: user attributes, behavior, and network topology of the power portal website. When users interact on the power portal website, the active user group can be found by analyzing the degree of user activity. When power users interact on power portal websites, their behavior information spreads outward through network nodes, similar to the energy field transmission in physics. Therefore, big data set *O'* of the power user behavior is inputted without missing or redundant data, and the power user group is identified using the energy field model. When a power portal user node V_a releases an interactive information, the customer service node $V_{a'}$ of the power portal website feedbacks the interactive information, which are likely to be seen by other users. Meanwhile, the network node energy of the power portal users and customer service nodes are enhanced when other users see the interactive information. Based on above analysis, $U_1=\{V_1, V_2, \cdots, V_m\}$ refers to the network node collection of the power portal website, among which V_a is the arbitrary network node of the power portal website, and $\alpha \in m$, *m* is the total amount of network nodes of the power portal website. When customer service feedback $V_{a'}$ information directly, the energy injected is the apparent energy, which is marked as *E(r=1)*. Therefore, the apparent energy of nodes is directly affected by user attributes and user behavior. The apparent energy V_a is obtained using Formula (8), as follows:

$$O': U_{property}, UF_{action} \rightarrow E(r=1) \tag{8}$$

Where $U_{property}$ is the attribute of V_a, and UF_{action} is the behavior of the user node in the power portal website.

$U_2=\{B_1,B_2,\cdots,B_n\}$ refers to the combination of users who can view the behavior of the user node V_a, $B_{a'}$ is any user who can view the behavior of the user node V_a, and $\alpha' \in n$, n is the sum of the users who can view the behavior of user node V_a. The hidden energy indicates that $B_{a'}$ can view the input energy of user node V_a directly; it is represented by $E(r>1)$. The calculation format is shown in Formula (9), as follows:

$$O':U_{property},E_{father},NTS\xrightarrow{d_{aa'}}E_a(r>1) \tag{9}$$

Where E_{father} is the energy of V_a parent node, NTS is the network topology structure of the power portal website, $d_{aa'}$ is the number of hops between user nodes V_a and $B_{a'}$, and $E_a(r>1)$ is the hidden energy of V_a.

The activity index is used to measure the active degree of the power user group on the power portal website, which is the dimensionless behavior of the power user group on the power portal website. Δt refers to the unit time period of the power user group behavior big data, AI is the active index of each network node within this unit time period, and the active index sequence of the network nodes composed of H nodes within this time period is $Rank(H,\Delta t)=\{AI_\alpha,\alpha=1,2,\cdots,H\}$ Thus, the active index $AI(Va)$ of user node V_a is shown in Formula (10), as follows:

$$AI(V_a)=\frac{\sum_{c=1}^{N}C_{retweet}(V_{fans},\Delta t)}{\sum_{c=1}^{N}C_{post}(V_{fans},\Delta t)} \tag{10}$$

Where $C_{retweet}$ refers to the number of times the behavior of user node V_a can be seen by other user nodes, C_{post} is the number of times that the behavior of V_a has been propagated by other user nodes to other nodes, and V_{fans} is the interaction times between the user node and the power portal network node.

Formula (8) and Formula (9) are used to calculate the apparent energy and hidden energy of the current power user group, respectively, and Formula (10) is used to calculate the active index of the apparent energy and hidden energy of the power user group. When the active index of the apparent energy and hidden energy of the power user group is higher than 0.7, the power user group corresponding to the apparent energy is important. When the active index of the apparent energy and hidden energy of the power user group is between 0.4 and 0.7, the power user group corresponding to the apparent energy is ordinary. When the active index of the apparent energy and hidden energy of the power user group is less than 0.4, the power user group corresponding to the apparent energy is a zombie user group.

4. Results Analysis

The power users of a city's power portal website is considered the experimental object, and the proposed method is used to identify the user group of this power portal website. The city's power portal website provides professional knowledge and technology in energy, electricity, and other related fields for the personnel in the electric power industry. At the same time, it sets up exhibitions and training projects in the electric power industry and provides basic payment and business consulting services for users.

The proposed method is used to crawl the input big data of the user group on the city's power portal website during a certain period of time, and part of the crawling results are shown in Table1. Table1 shows that the paper method can effectively crawl the number of password entered by users in the login operation, the number of clicks on the top button when they view the top articles in the forum, and the number of clicks on the dialog box when they communicate with the customer service. The results indicated that the paper method has a strong ability to collect data from power users.

Table 1. Part crawling results of input big data of user group

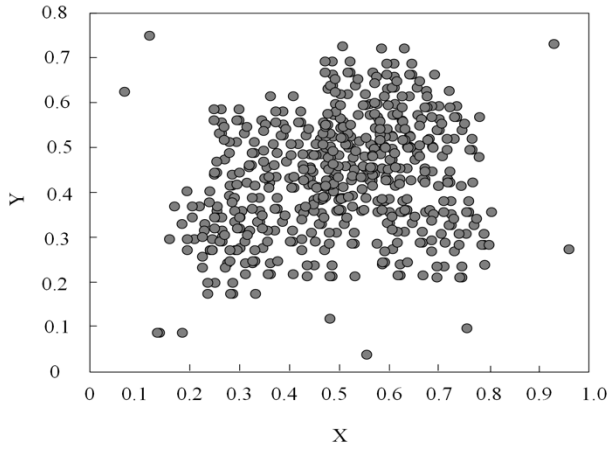
Functions	Data/Time	Description
Login	59	Enter password
View top article	338	Click the top button
Like the article	1023	Click the like button
Collect the article	559	Click the collect button
View forum comments	305	Comment sort button
View training video	119	Retrieve training video keywords
Communicate with customer service	582	Click the dialog box

A large number of power user group data collected by the proposed method is considered the experimental object, and this method is used to filter outliers. Then, the distribution state of a large number of power user group data is presented by undirected graph, and the filtering effect of the proposed method on the outliers of a large number of power user group data is analyzed; the results are shown in Figure 2. The figure shows that the initial distribution state of a large number of power user group data collected by the proposed system has the most clustered data, but few scattered big data points are discrete. After outlier filtering and processing of a large number of power user group data with the proposed method, discrete data points around the aggregated big data are removed, and the consistency of the power user group data is guaranteed. The results indicated that the proposed method has good ability to filter outliers from the user group data, thereby reflecting the ability of the proposed method to identify the power user group.

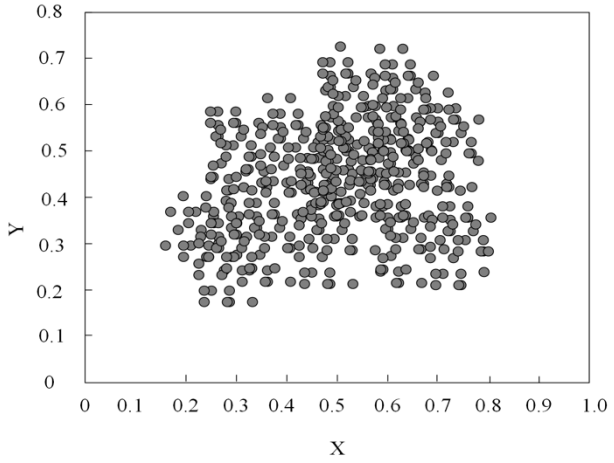
A total of 400 power users are considered experimental objects, which were divided into eight groups, with 50 power users in each group. The proposed method is used to calculate the energy field of all power users. In addition, the average energy field value of power users in each group is calculated according to the group division. The average energy field value of each power user is considered the measurement index to analyze the energy field capability of power users calculated by the proposed method; the results are shown in Figure 3. The figure shows that the average energy value calculated by the proposed method is almost exactly consistent with the actual energy value, indicating that the proposed method is sufficiently accurate to calculate the energy value of the power users. It also indicates that the proposed method can effectively distinguish the energy values of different power users and has a relatively accurate ability to identify power user groups.

A total of 300 power users from the city's power portal website are considered experimental objects, including 118 active power users, 93 ordinary users, and 25 zombie users. The proposed method is used to identify the 300 power users, and the identification results are shown in Table 2. Table 2 shows that the number of users in the identified

cumulative power user groups gradually increased with the increase in identification time in the process of identifying power user groups. Among them, at 0.8 s, the proposed method identified 25 zombie power users among the 300 power users; at 1.6 s, it identified 93 ordinary users; at 1.8 s, it identified 118 active users among the 300 power users. The identification result is exactly the same as the actual result. The finding shows that the accuracy of the proposed method is close to 100%, and it has a relatively excellent ability of identifying power user groups.



(a) Prior to outlier filtering



(b) After outlier filtering

Fig. 2. Outlier filtering test results of power user group data

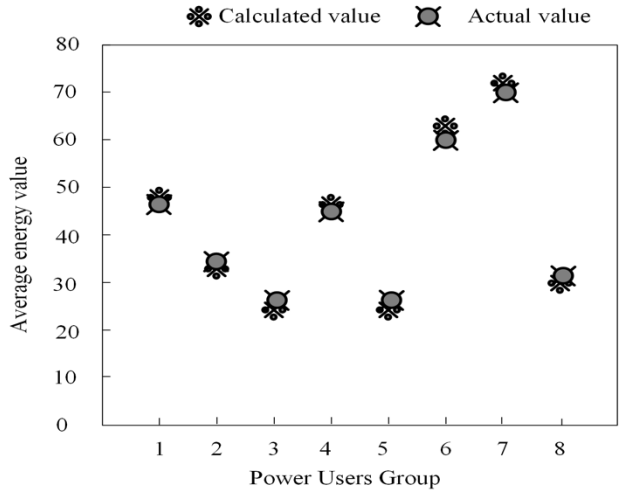


Fig. 3. Calculation results of average energy field of power user group

Table 2. Identification result of power user groups

Time/s	Total quantity		
	Active group	Ordinary group	Zombie group
0.2	29	19	3
0.4	33	36	9
0.6	49	44	18
0.8	51	59	25
1.0	68	62	25
1.2	76	75	25
1.4	93	88	25
1.6	109	93	25
1.8	118	93	25
2.0	118	93	25

5. Conclusion

This study investigates the identification method of power user group based on big data. In this method, web crawler technology is used to crawl the big data of power user group behavior and obtain the active degree of power users on the power portal website as well as the apparent energy and hidden energy from the big data, and then determine the type of power user group. After experimental verification, the following conclusions are drawn:

(1) After the outlier filtering process of a large number of power user group data with the proposed method, the discrete data points around the aggregated big data can be removed, resulting in better outlier filtering ability of the user group data.

(2) The average energy value calculated by the proposed method is almost exactly consistent with the actual energy value, thereby effectively distinguishing the energy values of different power users and obtaining a relatively accurate ability to identify power user groups.

(3) The proposed method identified 118 active users. The identification result is exactly the same as the actual result, and the accuracy of identifying the power user group is close to 100%, indicating a relatively excellent identification ability of power user group.

With respective to the identification of power user groups, the next work is to build a user similarity identification model, which can deeply understand the characteristics of power user groups and provide accurate implementation of demand response strategies to improve the energy efficiency of the power grid, which can be further studied in the future.

References

1. Ding, H. J., Pinson, P., Hu, Z. C., Wang, J. H., Song, Y. H. "Optimal offering and operating strategy for a large wind-storage system as a price maker". *IEEE Transactions on Power Systems*, 32(6), 2017, pp. 4904-4913.
2. Guo, H. Y., Chen, Q. X., Xia, Q., Kang, C. Q. "Electricity wholesale market equilibrium analysis integrating individual risk-averse features of generation companies". *Applied Energy*, 252, 2019, pp. 113443.
3. Nasrolahpour, E., Kazempour, J., Zareipour, H., Rosehart, W. "A bilevel model for participation of a storage system in energy and reserve markets". *IEEE Transactions on Sustainable Energy*, 9(2), 2018, pp. 582-598.
4. Chen, R. D., Paschalidis, I. C., Caramanis, M. C., Andrianesis, P. "Learning from past bids to participate strategically in day-ahead electricity markets". *IEEE Transactions on Smart Grid*, 10(5), 2019, pp. 5794-5806.
5. Guo, H. Y., Chen, Q. X., Gu, Y. X., Shahidehpour, M., Xia, Q., Kang, C. Q. "A data-driven pattern extraction method for analyzing bidding behaviors in power markets". *IEEE Transactions on Smart Grid*, 11(4), 2020, pp. 3509-3521.
6. Guo, H. Y., Chen, Q. X., Zheng, K. D,, Xia, Q., Kang, C. Q. "Forecast aggregated supply curves in power markets based on LSTM model". *IEEE Transactions on Power Systems*, 36(6), 2021, pp. 5767-5779.
7. Zhou, X. J., Peng, Q., Yang, R., Han, Z. Y., Wang, M. "Power price marketing strategy of comprehensive energy-based electricity sales company participating in electricity market competition under ubiquitous environment of Internet of Things". *Power System Technology*, 44(4), 2020, pp. 1317-1324.
8. Jiang, Z., Lin, R. H., Yang, F. C. "A hybrid machine learning model for electricity consumer categorization using smart meter data". *Energies*, 11(9), 2018, pp. 110-117.
9. Lindén, M., Helbrink, J., Nilsson, M., Pogosjan, D., Ridenour, J., Badano, A. "Categorisation of electricity customers based upon their demand patterns". *CIRED-Open Access Proceedings Journal*, (1), 2017, pp. 9-15.
10. Minghao, P., Ryu, K. H. "Local characterization-based load shape factor definition for electricity customer classification". *IEEE Transactions on Electrical and Electronic Engineering*, 12(9), 2017, pp. 22-30.
11. Sharma, P., Singh, R., Foropon, C., Belal, H. M. "The role of big data and predictive analytics in the employee retention: a resource-based view". *International Journal of Manpower*, 43(2), 2022, pp. 411-447.
12. Yang, Z., Chen, X., Zhang, J. Jin, M., Zhou B. "Construction and application of information ecology model in power industry under big data environment". *Information Science*, 38(5), 2020, pp. 88-92.
13. Zuo, Y., Cui, S., Zhang, Q., Zhu, L. "Research and analysis on the identification technology of user group network in distribution station area". *IOP Conference Series: Earth and Environmental Science,* 714(4), 2021, pp. 29-35.
14. Zhang, J., Guo, Y. "Social network user identity recognition method based on preference logic". *Computer Simulation*, 39(4), 2022, pp. 450-453.
15. Liang, R., Zhang, L., Wang, L. "Study on the identification of lead users in the Internet-based open and innovative communities". *Journal of Machine Design*, 36(10), 2019, pp. 24-30.
16. Kiannejad, M., Salehizadeh, M. R., Oloomi-Buygi, M., Shafie-khah, M. "Artificial neural network approach for revealing market competitors' behavior". *IET Generation, Transmission & Distribution*, 14(7), 2020, pp. 1292-1297.
17. Gu, Y. X., Zhang, K. D, Wang, Y., Zhang, X., Chen, Q. X. "An online approach for partial topology recovery in LMP markets". *International Journal of Electrical Power & Energy Systems*, 134, 2022, pp. 107-115.