Real-time Multidimensional Data Mining and Analysis Technology Based on Big Data

Yuan Tian*
Information Center
Yunnan Power Grid Co., Ltd
Kunming, China
e-mail: 228478554@qq.com

Wen Ma
Information Center
Yunnan Power Grid Co., Ltd
Kunming, China
e-mail: 22040278@qq.com

***Abstract*—With the wide application of modern information technology in various industries, more and more data is generating and storing. The processing of these massive data has become an important challenge in the development of the industry now, and the processing and deep mining of these data are also beneficial to the development and survival of enterprises. In this paper, the data mining technology of power industry is deeply analyzed to realize the real-time multidimensional linkage data mining of power grid system based on information visualization information interaction technology and data mining technology. So that the interrelation and internal connection in the massive data of power can be deeply excavated, and the visualization is carried out by means of charts and other ways. The real-time load monitoring, heavy overload monitoring and power imbalance monitoring in 2020 are carried out as research examples. The multidimensional linkage of various indexes of power data is demonstrated through the case study and analysis.**

Keywords—Big data, Real-time, Multidimensional linkage, Data mining, Power

I. INTRODUCTION

At present, with the continuous upgrading and reform of China's power industry, the construction of digital grid is gradually promoted. In the process of operation and management of digital grid, a large number of grid data has been accumulated. How to use these data has become an important challenge for the current grid operation. It is a new application field of data mining technology to mine useful information for power grid enterprises from massive power grid data. Data mining technology is used to deeply mine the massive grid data generated and accumulated by the operation of digital grid, from which we can find the relevance between the data and guide the business development of enterprises. At the same time, the use of big data technology is also conducive to the construction of grid smart cloud platform and the modernization of the power grid industry. [1]In power grid enterprises, the data visualization technology combined with data mining technology can not only deeply analyze a large number of data, but also comprehensively improve the business level of power grid management. Although the technology is still in its infancy, it can be predicted that it will have a very wide range of applications and a very large application prospect. At the national level, China has fully launched and real-time national big data strategy, which is also an important part of the national industrial upgrading and reform strategy, to create an independent industrial ecosystem of data and technology. [2]For enterprises, after a long period of information application education and construction preparation, many enterprises have been more firmly supporting the application of big data, and hope to improve business efficiency and enhance their core competitiveness. [3-4] For the management of power grid, it is also necessary to strengthen the application of new technologies such as big data and artificial intelligence. In order to realize real-time calculation of massive data in power grid data, multidimensional association analysis must be carried out to meet the needs of visualization. Applying the results of big data technology to regional power analysis can improve the construction planning of distribution network, optimize the business environment of power, and improve the service level to customers.

II. REAL TIME MULTIDIMENSIONAL LINKAGE DATA MINING TECHNOLOGY BASED ON BIG DATA TECHNOLOGY

A. Information Visualization and Interaction Technology

Information technology is the visualization of nonspatial data or spatial data. In order to realize the storage, retrieval, classification and transmission of massive data at the same time, it is also necessary to understand the internal relationship and development trend of different data through in-depth analysis. [5]In this way, the deep meaning of data can be mined, so as to improve the data perception, enhance people's cognitive ability and level. Information visualization is a compression of data knowledge, which compresses more information and knowledge into a very small space. [6] For compressed data knowledge, we can use the integration of graphic design and information design to achieve the induction and collation of data, which makes these data have strong artistry and certain applicability. Using information visualization can better solve and deal with information problems, meet the different data use needs of more people, help people to analyze the hidden problems of information data from the data, visualization can make the data "live" in the form of graphics. Information visualization technology mainly consists of two processes. [7]The first process is to transform data into visual graphics that are easy to understand and accepted by people. The second is to use human-computer interaction to enable users to obtain information at all stages of information data transformation. Therefore, information visualization design runs through the entire user's operation process.

At the beginning of information visualization technology, it was mainly to filter the massive raw data, but considering the large scale of the data itself, it is impossible

to fully present it to users. In this case, the typical representative data can be selected. When users select data, the system will use more prominent colors or other ways to display, so as to facilitate communication between people. If it doesn't matter, keep it light gray. In order to find a visual presentation that satisfies the user's mental model, we can obtain a more novel visual form by using the algorithm. The choice of this visual form must be able to meet the user's mental needs. For information elements, information is abstract, and each information needs a visual element that can represent its meaning, which can be graphics, icons, photos, etc. In order to realize the interaction of visual interface, users need to use certain tools or software operations. These tools are easy to get and operate, and they are more natural ways of operation. A good form of information visualization can help users to find problems. Users can find problems and rules in data information from the interface. The corresponding system can take some tips, warnings and other positive help to users. The chart component based on visualization technology includes pie chart, scatter chart, funnel chart, bubble chart, thermodynamic chart, etc.

Because the monitoring of power system is a real-time monitoring process, most of the scenario analysis is based on the data comparative analysis of time series, so it is necessary to count the results of time dimension in advance, and then use the chart component to display. However, in the case of dealing with massive data, it may be manifested as the lack of panoramic display ability, and there will be some problems in data loading and implementation display. In order to meet the processing requirements of massive data, this paper proposes the panoramic real-time playback technology of massive data, which can achieve high-performance storage of massive data, and achieve index mapping through GPU parallel computing at the service end, and complete the panoramic scale data visualization display at the millisecond level. [8-10]It takes less time to load panoramic data, which can be used to play panoramic thermal map of time series, and display the thermal effect changes as animation effect. In addition, geographic location information can also be used to realize map visualization display, including GIS map, Baidu Gaode, offline map and other map data or location coordinate information, using the results of data analysis to achieve visual rendering display.

The use of visualization components can help the front-line personnel and management decision-makers in the power industry to understand the power supply situation and equipment efficiency level in a specific area, and help them to carry out the operation and management of the power system. This time, the virtual data cube modeling technology is introduced to realize the data information visualization of the power system. Virtual data cube Modeling Technology (hereafter referred to as data cube) doesn't care about the real structure of data storage, but it needs to associate all dimensions and indicators into a dataset through view Association, and specify the column fields in the view as indicators or dimensions. [11]The virtual data cube modeling based on this idea can be realized only by changing the view and redefining the index and dimension column, and does not need to recalculate the physical table structure. By using the data cube modeling method, the view can be associated into a virtual data cube by using the fact table, and the field types in all data cube views can be given different dimension indicators. Then, we use the unified interface to aggregate the indicators in multidimensional, so that the final results can be unified and rendered by the front component.[12] The front-end components include geographic hot spot map, multidimensional data Kanban, etc. the multidimensional data indicators rendered by these front-end components can be applied to specific data mining analysis. Therefore, we can build a multi angle cube model for different analysis scenarios. When we do data mining analysis, we have a more flexible combination mode. We can also realize interactive exploration and analysis of the system through multidimensional data linkage analysis and other ways. The flow of virtual data cube modeling technology is shown in Figure 1.

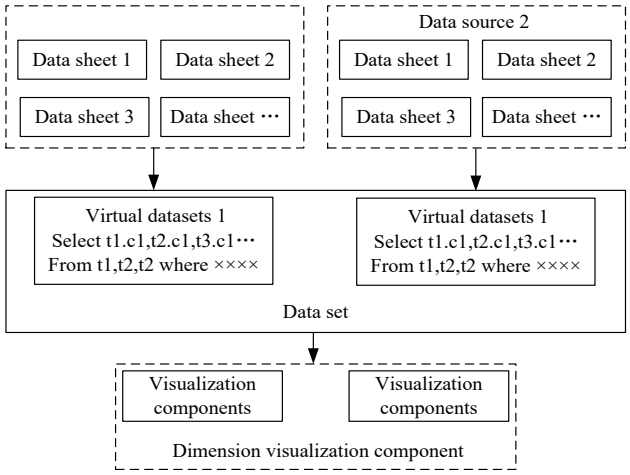

Fig. 1. Multidimensional visualization data cube modeling technology

B. ARIMA Model

Further introduce the ARIMA model, which can be used to convert non-stationary time series into stationary time series, and then use the lag value in the dependent variable and the present value of the random error term to perform regression analysis. What's more, it can use the results of regression analysis to construct a stationary time series model. The ARIMA model can be used to analyze and judge whether the time series in the regression analysis is stable and judges different parts of it. The construction process of the model is shown in Figure 2.

The methods of converting non-stationary time series to stationary time series using ARIMA model mainly include difference method and logarithm method. Firstly, unit root test and DF unilateral test are carried out. Let the non-stationary state of the sequence be H_0, the expression is $|\phi_1| \geq 1$, the stationary state of the sequence is H_1, and the expression is $|\phi_1| < 1$. From this, the test statistical measurement formula can be obtained as follow.

$$\tau = \frac{|\hat{\phi}_1| - 1}{S(\phi_1)} \sim t \tag{1}$$

Further use AFC and PACF to determine the form of ARIMA model, the calculation formula is shown as follow.

$$ACF_{x_t,x_{t-k}} = \frac{E[(x_t - Ex_t)*(x_{t-k} - Ex_{t-k})]}{E(x_{t-k} - E(x_{t-k}))^2} \quad (2)$$

$$PACF_j = \frac{E[(x_t - Ex_t)*(x_{t-j} - Ex_{t-j})]}{\sigma_t * \sigma_{t-j}} \quad (3)$$

For the non-seasonal ARIMA model's judgment formula, it is determined whether the model belongs to the AR model or the MA model. The AR model is shown below.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t \quad (4)$$

Also, the MA model is shown below.

$$y_t = c + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-q} \quad (5)$$

Finally ARIMA model formula is obtained by combining formulas above.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_p e_{t-q} \quad (6)$$

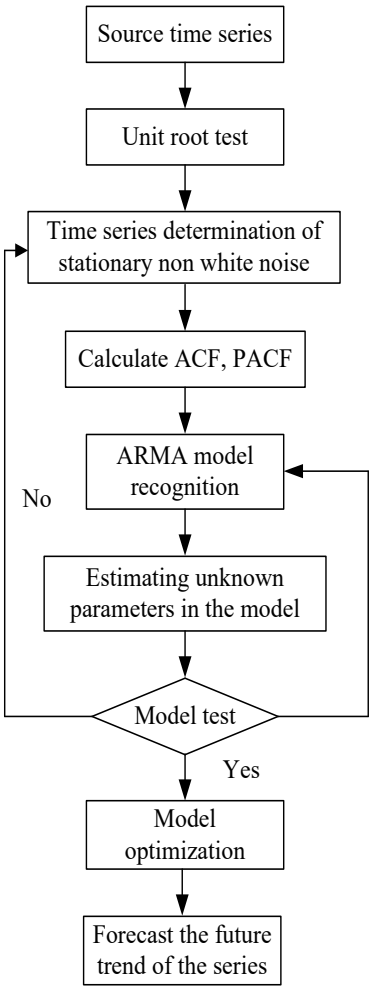

Fig. 2. ARIMA model building process

The structure judgement of the model is showing in Table I.

TABLE I. MODEL STRUCTURE JUDGMENT

Model structure	AR Model	MA Model	ARIMA Model
ACF	Trailer	Truncation	Trailer
PACF	Truncation	Trailer	Trailer

C. Neural Network Model

Neural networks are widely connected by multiple processing units to form a complex network system, which can reflect the characteristic characteristics equivalent to human brain functions. It is a highly complex nonlinear dynamic learning system. The mathematical model of the neural network is a description of the characteristics of the brain system. The model has very good parallel distributed processing characteristics, has strong distributed storage and learning capabilities with a high degree of robustness and fault tolerance. Taking the neuron in Figure 3 as an example, the neuron in the structure is connected to the input signals of the three external neurons, and is transmitted through the weight connection. w_1, w_2, w_3 in the figure is the weight of the neuron, x_1, x_2, x_3 is the input neuron, and y represents the output neuron. Assume that the activation function of the neuron is f, and the expression of the output neuron is as follows.

$$y = f(\sum_{i=1}^{3} w_i x_i - \theta) \quad (7)$$

After entering the input layer neurons, the original data will be transmitted layer by layer to the final output layer result, and the output layer error is calculated. After the error is propagated back to the hidden neural layer, the connection weight and threshold are adjusted continuously, and the expected error convergence is achieved after repeated iterations. The process is shown in Figure 3.

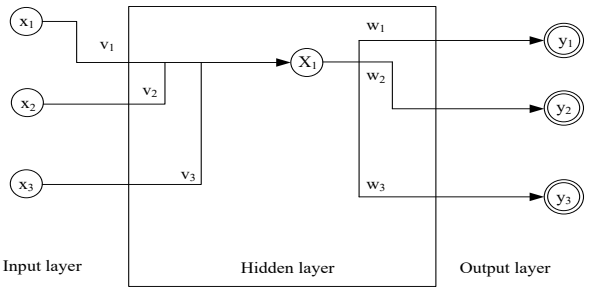

Fig. 3. Neural network structure diagram

As can be seen from Figure 3, x_1, x_2, x_3 is the input layer, v_1, v_2, v_3 is the hidden layer weights of input layer value, w_1, w_2, w_3 is the output layer weights, y_1, y_2, y_3 is the output layer neurons, and $\theta_1, \theta_2, \theta_3$ is the threshold of output layer. For the hidden layer b_n, its input expression and output expression are as follows.

$$a_h = \sum_{i=1}^{3} v_{ih} * x_i \tag{8}$$

$$\hat{y}_k = f(\beta_k - \theta_k) \tag{9}$$

For the output neurons k , there is the following formula.

$$\min E_k = \sum_{j=1}^{k} (\hat{y}_j - y_j)^2 \tag{10}$$

III. Data Processing Before Multidimensional Linkage Data Analysis

Power grid system management is developing towards automation and intelligence. Data mining technology in digital power grid has very practical application value in regional power consumption. It can be specifically applied to test load monitoring, power supply quality, and power consumption service topics. The analysis and mining of power consumption data need to solve users' needs from the perspective of project objectives. Take the heavy overload monitoring application as an example, the business understanding is shown in Figure 4.

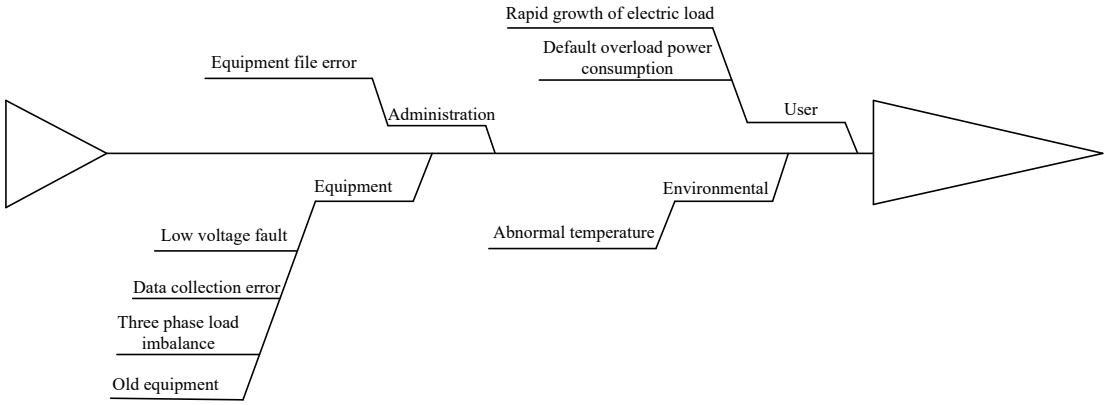

Fig. 4. Business understanding of heavy overload monitoring application

Data exploration is the empirical analysis basis of regional electricity management based on big data technology, including data collection, sorting, data exploration and descriptive analysis. Among them, data collection refers to the relevant process of collecting and acquiring data mining, and these data are obtained from outside the public channels. Data understanding is to understand the field meaning of the collected data and confirm it with the business understanding. The exploratory analysis of data refers to the analysis of collected data and retrograde exploratory data by using the gray correlation analysis method, such as obtaining the correlation of temperature, humidity, air pressure and other relevant data parameters. The exploration and analysis of the research variable relationship indicators include the linear relationship, the degree of collaborative change, the correlation between indicators, etc., as well as the distribution hypothesis test of variables. Descriptive analysis of data is to show the relationship between variables by means of graphs and other ways, which can be displayed by means of quartiles, extreme values, box line graphs, trend graphs, etc., or by means of histograms to carry out classified statistics of samples.

In data mining and analysis, the database is very easy to receive noise, omission, inconsistent data and other interferences, which will cause great trouble to data mining and analysis. Therefore, it is necessary to preprocess these data. After preprocessing, data redundancy and low efficiency of data mining can be avoided. Data can be processed through data preprocessing technology, Identify, analyze and merge data from different sources, inherit and merge multiple data into a unified data storage, forming a data warehouse or data party. By clustering, clustering, deleting redundant feature data and other methods to achieve data compression, preprocessing before data use can greatly improve the quality of data mining mode, improve the efficiency of data mining, and reduce the time required for data mining.

By filling in the missing values, identifying and deleting the outsider, the data can be cleaned up. Because the source data sources may be different, these different data may be composed of the same concept attribute data, resulting in data redundancy or even inconsistency. Using data inheritance can solve and deal with the problems of different attributes of different source data, make the data more precise and clear, and make it have a very good structure. After observing and analyzing the characteristics of the data, the original data may not meet the requirements of the model. In this case, it is necessary to realize the transformation of the data, so that they can map to the smooth performance in a specific interval to meet the requirements of the model. For massive data, if we want to process these data at the same time, it will inevitably cause a waste of computer resources, generate a lot of demand, and the speed of data mining will also decline rapidly. In this case, we need to compress the data, but can not damage the data mining results. There are many data reduction strategies, such as dimension reduction, data compression, data aggregation, etc., which should be selected according to the specific situation.

IV. Real-Time Multidimensional Linkage Data Analysis of Regional Power Consumption Based on Big Data Technology

A. Real Time Load Monitoring

The data mining technology and visualization technology are applied to the real-time monitoring of the

power consumption in a certain area, and the real-time load monitoring of the power consumption in a certain area is carried out. The monitoring results are displayed in the way of curve chart, which supports historical playback and playback according to the time axis. The multidimensional statistics of distribution load can be carried out and displayed to users in the way of chart. Users can also zoom in, zoom out, translate and other operations on the thermal diagram, etc., and realize the linkage of thermal diagram, statistical diagram and total load curve in the monitoring process. The thermodynamic diagram in the system can show the distribution of load in the whole area, view the historical load data, and play the historical load data through the adjustment of the time axis. This time, the load brightness value of distribution transformer measurement point is calculated at a certain time point in 2020, and shows a specific change trend according to 15min, as shown in Figure 5 and Figure 6. With these icons, you can monitor different voltage levels and view the load of a single distribution transformer.

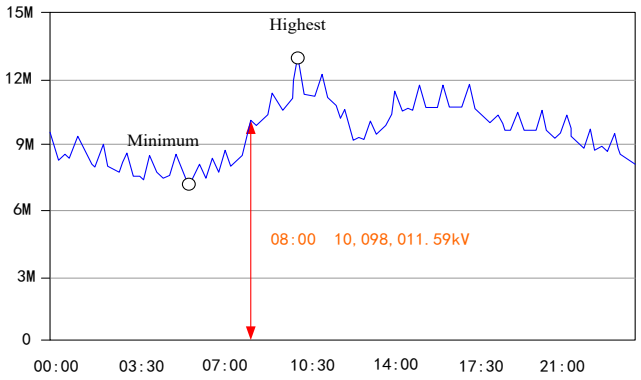

Fig.5 Load trend at distribution and transformation time point

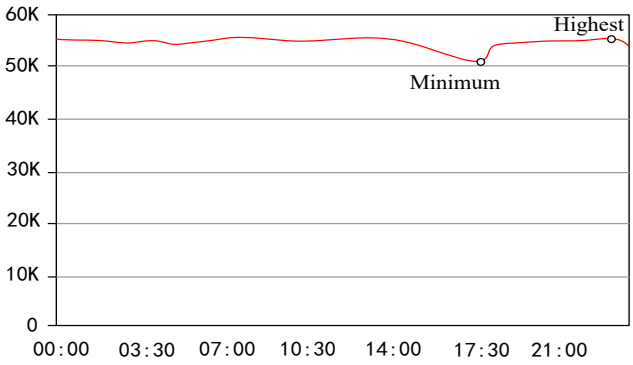

Fig. 6. Time point load data change chart

B. Heavy Overload Monitoring

After the load data of distribution and transformation are accessed, the final results will be displayed in the form of visualized thermodynamic diagram, covering the distribution of variable weight and overload in the whole region, and can be zoomed and historical playback based on the time axis. The distribution of heavy overload in all distribution and transformation loads is shown in chart form after multidimensional statistics. Each chart can realize real-time linkage, and support to locate and view the delay distribution variable overload indicators. On the map, the distribution of regional heavy overload at different time points can be calculated by using the geographic data of longitude and latitude. The judging conditions of heavy overload distribution transformer mainly include total times, total time, maximum load rate, average load rate, total time, etc. the multidimensional statistics of distribution transformer load and real-time linkage of multiple charts are used to display the information that needs to remind users with different colors. Figure 7 shows the distribution of time points and the week of heavy overload.

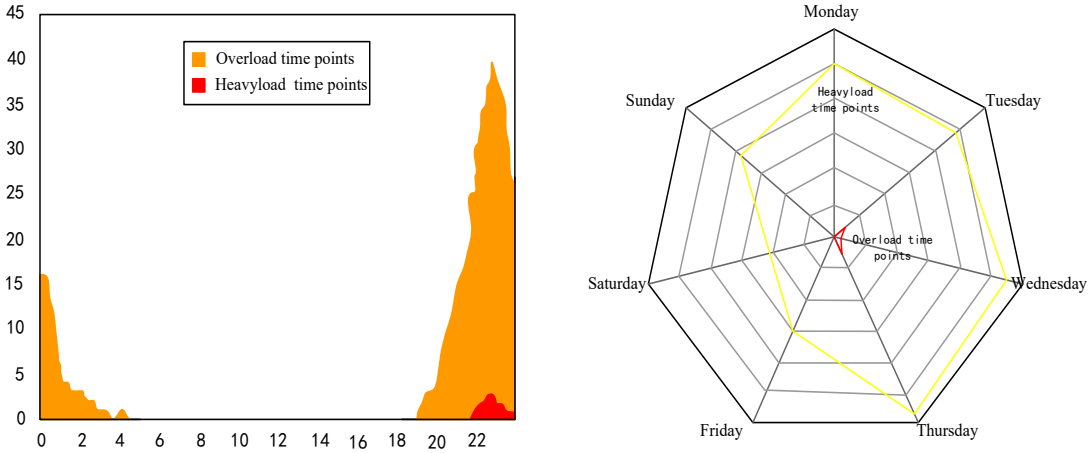

Fig. 7. Time point distribution and weekly distribution of heavy overload

C. Current Imbalance Monitoring

When monitoring the current imbalance, we can combine the power factor data of distribution transformer with the geographical coordinates to calculate the current imbalance distribution and change trend of each distribution transformer measurement point. It can be monitored according to the preset administrative area according to the spatial dimension, and the unbalanced distribution of distribution and transformation flow can be displayed in the form of charts, which can realize real-time linkage, real-time analysis of user-defined areas, power factor curve, statistical chart and support real-time linkage of thermal chart. For the current imbalance monitoring, you can view the current imbalance of a single distribution transformer, and play back the history of the thermodynamic diagram. The time point power diagram is shown in Figure 8.

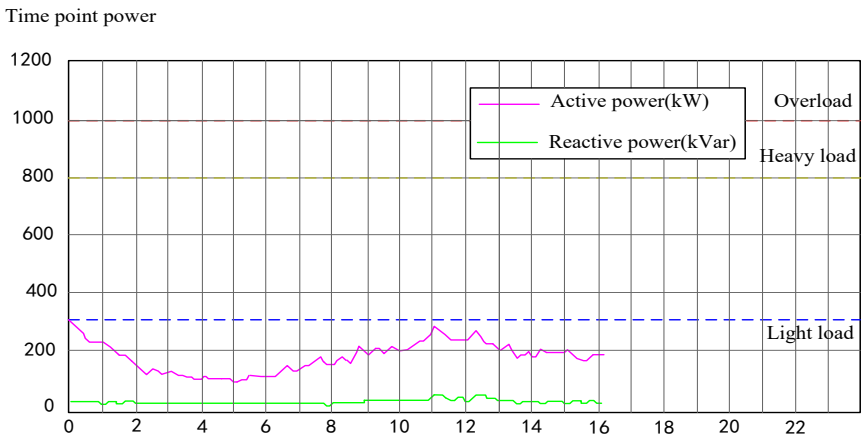

Fig. 8. Time point power diagram

V. Conclusion

In order to realize the multidimensional linkage data mining technology of power system, firstly, information visualization interaction technology is introduced. The data can be displayed in the form of charts based on the visualization technology, and then the ARIMA model is introduced. By using this model, the non-stationary time series can be transformed into stable time series. In order to better realize the data mining and processing, the neural network model can be used to repeatedly optimize the original data, so as to obtain the optimization results that meet the preset objectives. Before data mining and visualization, all the source data must be preprocessed to realize the business understanding of the data. After the introduction of the regional distribution case diagram at a certain time point in 2020, the real-time load monitoring, heavy overload monitoring, and current imbalance monitoring of the power grid are realized respectively. In the monitoring process, the real-time linkage of various indicators can be achieved. The final results are shown in the form of charts, which is convenient for users to monitor and manage the power grid in real time.

References

[1] D. Dahmani, S. A. Rahal, G. Belalem, “Improving the performance of data mining by using big data in cloud environment.” Journal of Information & Knowledge Management, vol. 15, no. 4, pp. 1650038, 2016.

[2] Z. Y. Qu, J. Y. Huang, A. “Fast scene constructing method for 3d power big data visualization.” Journal of Communications, vol. 10, no. 10, pp. 773-777, 2015.

[3] S. Sanga, A. Vladimirova, D. Richard, Goold, et. Abstract 4870: GenePool: “A cloud-based technology for rapidly data mining large-scale, patient-derived cancer genomic cohorts including The Cancer Genome Atlas.” Cancer Research, vol. 75, no. 15 Supplement pp. 4870-4870, 2015.

[4] Abraham, Yosipof, Omer, et al. “Visualization based data mining for comparison between two solar cell libraries.” Molecular Informatics, 2016.

[5] S. A. S. Alhamdy, A. Pourghassem, M. G. Ahmadi, et al. “Electric power load forecasting of babol city based on BP neural network.” Life Science Journal, vol. 10, no. 2, pp. 950-953, 2013.

[6] OLAGUNJU, MUKAILA, JIMOH, et. “Electric power energy distribution estimate using visualization technique.” International Journal of Innovative Research in Computer & Communication Engineering, vol. 1, no. 4, 2013.

[7] S. Takeshi, H. Yoichi. “Visualization of electric power at both campus and a measurement of through restrictions period on the use of electricity.” ieice technical report information & communication management, pp. 112, 2012.

[8] D. Zhang, S. Li, “Optimal dispatch of competitive power markets by using powerworld simulator.” International Journal of Emerging Electric Power Systems, vol. 14, no. 6, pp. 535-547, 2013.

[9] G. Sassenrath, “Development of synthetic imagery for visualization of cotton reflectance.” Electric Power Systems Research, vol. 80, no. 2, pp. 230-239, 2010.

[10] X. Peng, D. Deng, S. Cheng, et al. “Key technologies of electric power big data and its application prospects in smart grid.” Zhongguo Dianji Gongcheng Xuebao/proceedings of the Chinese Society of Electrical Engineering, vol. 35, no. 3, pp. 503-511, 2015.

[11] J. Zhang, H. Chen, J. Chen, et al. “Smart grid situation awareness diagram modeling and conceptual design of situation awareness visualization.” Dianli Xitong Zidonghua/automation of Electric Power Systems, vol. 38, no. 9, pp. 168-176, 2014.

[12] MAKINO, Hiroyuki, KURIHARA, Syunsuke, ARAGANE, Yosuke, et al. “Dks: denryoku kashika system: a cloud-based sensor monitoring and visualization system for saving electricity.” ieice technical report data engineering, pp. 113, 2013.