International Journal of u- and e- Service, Science and Technology
Vol.9, No. 8 (2016), pp.413-422
http://dx.doi.org/10.14257/ijunesst.2016.9.8.35

A Non-Parametric Method of Electric Power Enterprise Arrears Prediction Model

Nijiati Najimi[1], Chen Jianxin[2], Ma Bin[2], Wang Tao[2], Han Shuangli[3] and Yang Yu[2]

[1]College of Electrical Engineering, Xinjiang University, Urumqi Xinjiang Uygur Autonomous Region, 830068
[2]Xinjiang Electric Power Company Information Communication Co., Ltd., Xinjiang Uygur Autonomous Region, 830068, China
[3]Tianjin WOMOW S&T Co.LTd., Tianjin, 300170, China
yangyu@womow.com.cn

Abstract

Power supply enterprises are facing power customer promises to bring the risk of arrears, in order to avoid the risk of customer arrears, a large power customer arrear predicting model is of great significance, through analysis of large customers to quarter data to predict the coming development trend of arrears to guide the power supply enterprise to make decisions. In response to this demand put forward a nonparametric method to predict the trend of arrears. The method by comparing a fundraising events of recent trends in activity and collected plenty of historical data on electricity consumption and dynamic data trend of delinquent trend was predicted. As the power supply enterprise to develop recovery plans based on electricity, it establishes risk control principles and strategies. The experimental results show that the algorithm maintains a low error rate, which verifies the effectiveness of the algorithm.

Keywords: *Arrears prediction; Non parametric method; Prediction model; Large power customers; Power supply enterprises*

1. Introduction

The aim of this project is to get some idea of the uncertainty in PV generation profiles for better integration and management within the energy system. This report presents statistical analysis and machine learning method on correlation study between regionally close PV panels within CHINA. It analyses and describes influence factors and makes a prediction model for correlation of outputs. The influence factors as independent variables include distance of pairwise panels, weather conditions and time of year. The prediction model takes support vector regression (SVR) in machine learning to view the trend of output – correlation coefficients – as dependent variable and makes a prediction in specific conditions.

2. Preprocessing of Raw Data

Calculation of Correlation: Inner joining pairwise data sets into a frame by Date (1st level index) and Time (2nd level index). The next step is to scale each Power data to the same standard sized array, *i.e.*, normalizing data with a mean of zero and a standard deviation of one. This is called the Z-score method:

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

where μ is the mean value of data an σ is the standard deviation of data.

ISSN: 2005-4246 IJUNESST

And then, we measure correlation coefficient of the pairwise data sets (X and Y), which can be expressed as [1]:

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (2)$$

where cov is the covariance between two jointly distributed data sets (X and Y), and σ_X is the standard deviation of X.

Encoding Categorical Features: For categorical variables, such as weather condition, we convert these into dummy variables. A dummy variable takes values 0 or 1 to represent the true or false of specific feature effect [2]. In the section of multivariable analysis, we transform categorical feature Condition with 3 possible values into 3 binary features. Three possible values are ‘Fine’, ‘Cloudy’ and ‘Showers’. This means that 1 represents the existence of a value and 0 represents the non-existence of a value. In other words, if the weather condition is fine of a site, then it will be encoded as (1, 0, 0); if the weather condition is cloudy, then encoded as (0, 1, 0); if the weather condition is showers, then encoded as (0, 0, 1).

3. Univariable Analysis

Correlation of output between panels may depend on several factors. In this study, samples were analyzed for the following factors: distance, weather condition and time of year. For each factor, we first present an overview of data set, and then introduce the statistical methodology we applied and make some statistical analysis for influence factors.

In this part, we acquired 185 different points with distance from 1.31 km to 183.82 km from 21 panels. For each pairwise panels of a distance, we have calculated correlation coefficients in the whole year of 2014. There are 45283 samples for analysis. Each sample has attributes of correlation coefficient and distance. Figure 1, shows an overview of correlation coefficients distribution related to distance in 2014. It shows the mean and standard deviation, median (50%), max value and min value, 1st quartile (25%) and 3rd quartile (75%) for each distance. Based on the figure, it is clearly that the overall trend of coefficients is decreasing at the beginning and leveling off along with the increasing of distance. The correlation coefficients are mostly distributed between -0.5 and 1.0, which indicates that coefficients lower than -0.5 can be reasonably regarded as outliers. And the fluctuation of standard deviation indicates that there are other factors that have influence on them. Other factors will be analyzed in the next parts.

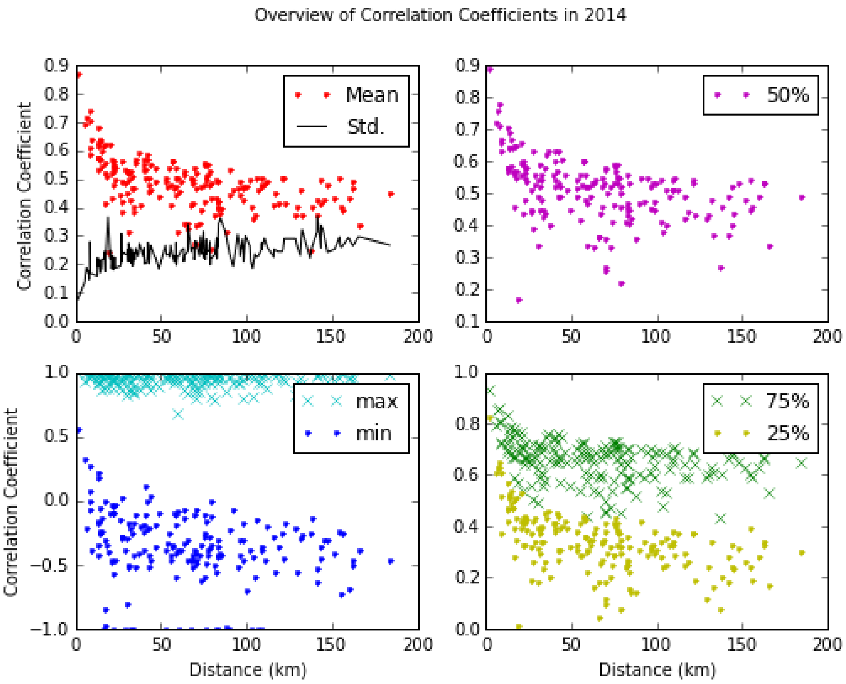

Figure 1. Distribution of Distance vs. corr_coeff in 2014

In order to analyze regional panels, we divided data of distance into 4 regions based on the distribution of raw data as showed in Figure 1, and made these 4 regions have similar volume of samples; these regions are defined as follows (in km):

Region1: [0, 30]; Region2: [30, 60); Region3: [60, 90); Region4: [90, 150].

For characterize the location and variability of correlation coefficient, we cite density estimator. Density estimation techniques are non-parametric tests for computing an estimate density function based on n independent and identically distributed samples $X = (x_1, x_2, \dots, x_n)$. The Kernel Density Estimator (KDE) [3] is:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (3)$$

where $K(\cdot)$ is the kernel function which is a non-negative function and integrated to 1, h is the bandwidth taken as a smoothing parameter, n is the number of samples. Also, the normal reference rule for optimal bandwidth h^* is:

$$h^* = C * A * n^{-1/5} \quad (4)$$

where C is calculated from the kernel, A is referred as:

$$A = min(std(X), IQR) \quad (5)$$

the IQR is the Interquartile Range, the 3rd quartile minus 1st quartile, $IQR = Q_3 - Q_1$. Note that, in this study, as the package used, we use adjusted value of A by dividing *IQR* by 1.34. *IQR/1.34* is a consistent estimate of *std.* if *X* is a normal distribution.

The Cumulative Distribution Function (CDF) of the KDE can be expressed as the integral of $\hat{f}(x)$:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} \widehat{f_X}(x)\, dx \quad (6)$$

Skewness and kurtosis are good measurements in statistics to describe the distribution. Skewness is an indicator of the asymmetry of probability distribution [4]. For skewness > 0, the mass of data are concentrated on the left of the mean with maximum value skewed to the right, that is called Positive skew or Right skewed distribution. For skewness < 0, the mass of data are concentrated on the right with maximum value skewed to the left, that is called Negative skew or Left skewed distribution. For skewness = 0, it means that the mean equals to the median, and the distribution of data is symmetrical. Thus, the larger of the absolute value of skewness, the more the skewing level. For univariate data $Y = (y_1, y_2, \dots, y_n)$, the Fisher-Pearson Coefficient of skewness (g_1) is expressed as:

$$g_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3/n}{\sigma^3} \quad (7)$$

where $\bar{y}$ is the mean value, σ is the standard deviation of Y and n is the size of data. Notice that, in this study, as the package used, the formula (7) has adjusted as G_1 [5]:

$$G_1 = \frac{\sqrt{n(n-1)}}{n} \frac{\sum_{i=1}^{n}(y_i - \bar{y})^3/n}{\sigma^3} \quad (8)$$

Kurtosis is an indicator of peakedness or flattening relatively to a normal distribution [4]. For kurtosis > 3, there is a 'sharpen' distribution with high probability for extreme value that near the mean and thick tails. For kurtosis < 3, there is a 'flatten' distribution with relatively lower probability for extreme values compared with normal distribution and a flat top. For kurtosis = 3, it indicates the standard distribution. Thus, the larger of the kurtosis, the denser of data points around the mean. For univariate data $Y = (y_1, y_2, \dots, y_n)$, the kurtosis ($k$) is expressed as:

$$k = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4/n}{\sigma^4} \quad (9)$$

where $\bar{y}$ is the mean value, σ is the standard deviation of Y and n is the volume of data. If the distribution is a standard normal distribution, then k is equal to 3. Note that, in this study, as the package used, we use adjusted formula called 'excess kurtosis' (K) [5], which is expressed as:

$$K = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^4/n}{\sigma^4} - 3 \quad (10)$$

In this section, we use the Gaussian Kernel, $K(u) = (1/\sqrt{2\pi})e^{-u^2/2}$, with univariable KDE to approximate the distribution of correlation coefficient in different regions. We take the value of C in formula (4) is equivalent to 1.059 referring Scott's rule [6]. Then, we get arrays of correlation coefficient as endogenous variable to calculate the coefficient of skewness, and to fit the KDE model for each region.

Figure 2 illustrates the observed data as a function of KDE and CDF. Table 1 shows result for regions, including the maximum density with corresponding coefficient, skewness and kurtosis. All regions show negative skewness and positive kurtosis, which means the mass of distribution is located on the right of the mean and the probability for maximum values is higher than a normal distribution. Region1, followed by region 2, shows highest skewing level and sharpest distribution than others, and extreme values are located around 0.64. Region3 and region4 illustrates similar distribution of the correlation coefficient. The difference between these two regions of skewness and kurtosis are small compared with other regions; in other words, the distribution tends to be stable for distance from 60km to 150km. Also, short distances have higher probability of positive correlation than long distances as showed in CDF. In conclusion, the correlation coefficients are mostly located round 0.5 and correlation between two panels with a short distance is higher than that with a long distance, and the distribution of correlation coefficient is stable in long-distance range (indicated by region3 and region4).

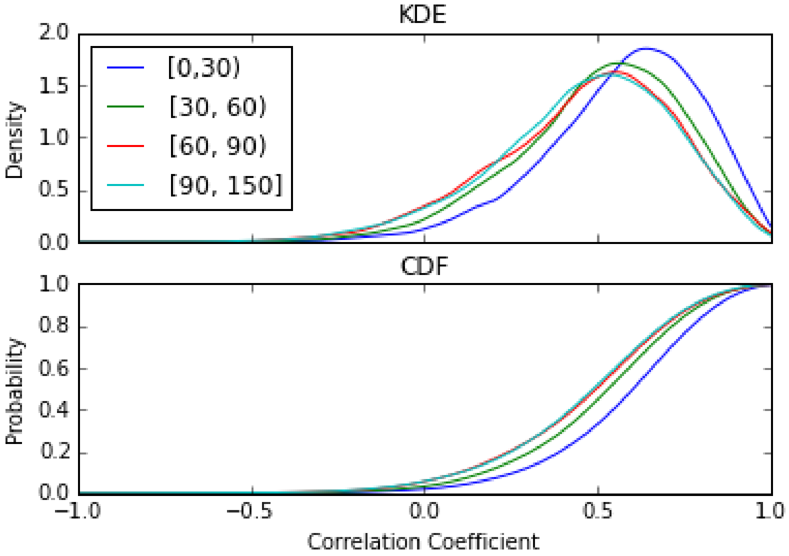

Figure 2. Density Estimation of Regions

Table 1. Description of Regions

	Region1	Region2	Region3	Region4
Mean	0.574	0.506	0.466	0.461
Skewness	-0.905	-0.715	-0.680	-0.680
Kurtosis	1.562	0.866	0.683	0.629
(coef, density)	(0.50, 1.44)	(0.50, 1.64)	(0.50, 1.58)	(0.50, 1.58)
(coef, max_density)	(0.64, 1.85)	(0.56, 1.71)	(0.55, 1.62)	(0.53, 1.60)

4. Time of Year and Correlation Coefficient

In this section, we acquired samples with features of correlation coefficient, distance and date. Samples are grouped by ‘distance’, and we select data points which ‘distance’ has 150 or more data points. There are 131 distances. For each distance, we analyze the trend of correlation coefficient throughout the year 2014; in other words, analyze the time series of coefficients. Figure 3 gives examples of time series in different distances. It shows the trend of correlation coefficient as time goes on. By observing each plot, we primary get an idea that in general the value is changing from 0.0 to 1.0 with some outliers.

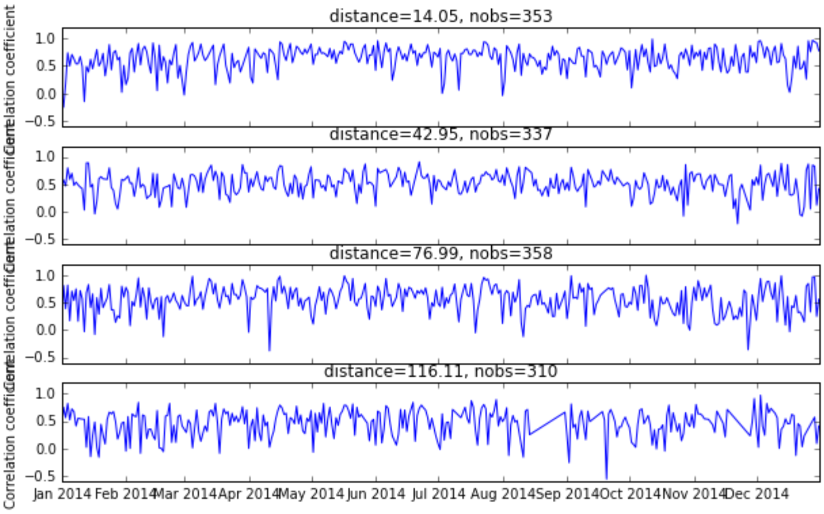

Figure 3. Examples of Time Series

Note: 'nobs' means the number of observations.

5. Regression Model for Correlation Coefficient

In this part, we take the support vector machine for regression as a prediction model to forecast the correlation coefficient of pairwise panels. And we take two criterions to evaluate the model: Mean Square Error (MSE) (11) and Mean Absolute Error (MAE) (12).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \quad (11)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \quad (12)$$

where n is the number of data points, Y_i is the true data, $\hat{Y}_i$ is the prediction data.

We mainly take two features, distance and weather condition in this part. As analyzed above, we integrate weather conditions Cloudy, Mostly Cloudy and Partly Cloudy into one factor as Cloudy, because of in-significance impaction. So, the input factors are: Distance (from 0 to 150 km), Condition and Co-condition (Fine, Cloudy, Showers). The output of the model is the value of correlation coefficients of pairwise panels.

Firstly, we take a daily example to analysis. We only select data on 2014-07-15 in this part. There are 136 data points in total, 75% of them are taken randomly as a training set. Figure 10 shows the real value and predict value for the test set on 2014-07-15. The MSE equals to 0.0255, the MAE equals to 0.1227. For a monthly example to analysis, we select July with 4022 samples in this part. 75% of them are taken randomly as the training set. 1006 samples are taken as testing set. The MSE equals to 0.0397, the MAE equals to 0.1512. The errors are larger than the former test. The correlation coefficient of pairwise panels has high possibility to around 0.6 in July according to the figure. Then, we take data in 2014 with 45283 samples to analysis. There are 11321 samples in the testing set. The predicted value is mainly in the range from 0.4 to 0.6. There are gaps between real value and predicted value. The MSE equals to 0.0569, the MAE equals to 0.1846. For larger dataset, the outputs of the model tend to stable at the range. But the real value may have some outliers. There may be other factors influence the correlation coefficients.

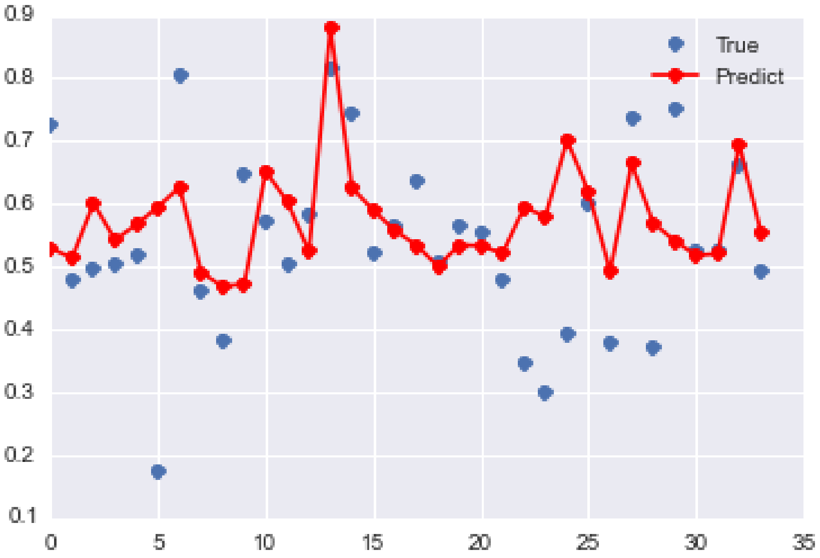

Figure 4. Test Set on 2014-07-15

Then, we take the polynomial function (11) as a kernel function to develop the model. We defined coefficients as follows: $degree = 2$; $C = 1.0$; $\varepsilon = 0.1$. Also, we set 75% of target data set as training set and the remaining as testing set. Figure5 shows the trend of MSE and MAE of Polynomial. The x-axis represent of the number of data that used for training and testing, where 75% of data is taken as training set and the remaining as testing set. The y-axis represents the value of corresponding error. The data set was selected randomly from the raw data. The figure shows that both MSE and MAE tend to stable with increasing the number of samples and they will stable around 0.06 and 0.17 respectively, especially for the large data set. That is to say, the model is suitable for a large data set, it avoids problem of over fitting.

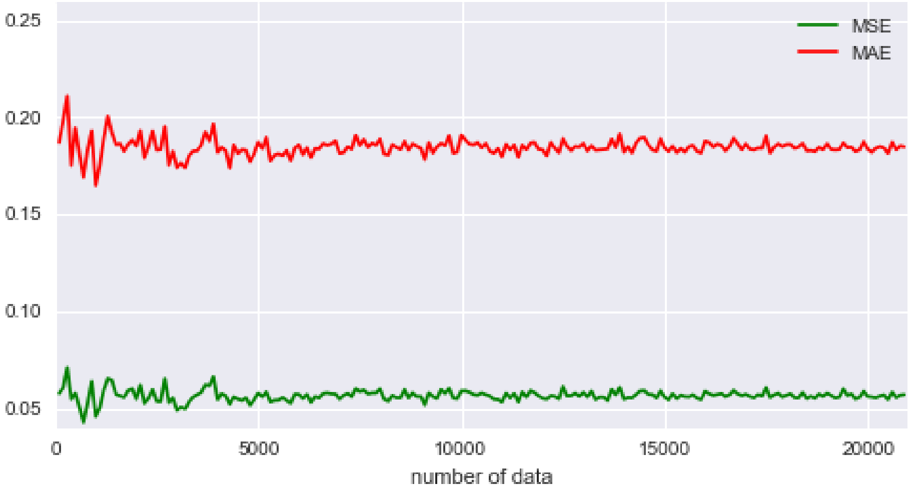

Figure 5. MSE and MAE of Polynomial

Then, we take the Gaussian function (12) as a kernel function to develop the model. We have defined coefficients as follows: $\gamma = 1/n$, n is the number of features of training set; $C = 1.0$; $\varepsilon = 0.1$; *degree* = 3. Also, we set 75% of target data set as training set and the remaining as testing set. Figure 12 shows the trend of values of MSE and MAE of RBF. Same axes are set as above. Compared with Figure 5, the

Gaussian kernel indicates appearance fluctuation with larger errors when the number of data is small (less than 25000 samples).

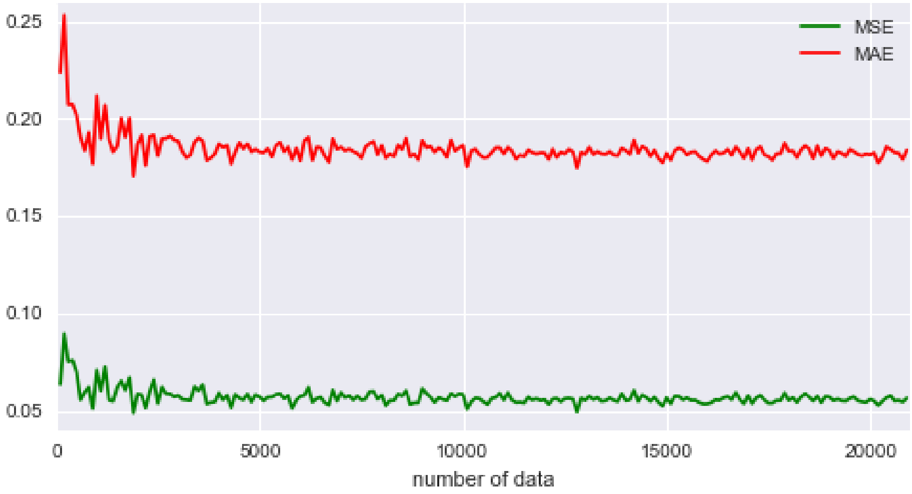

Figure 6. MSE and MAE of RBF

In conclusion, by comparing two kinds of kernels used in SVR, we say that polynomial kernel is more suitable for evaluating the correlation coefficients of pairwise panels, especially for the number of samples less than 25000. SVR model can be applied for a large data set, it avoids problem of over fitting.

6. Conclusion

The aim of this project is to get some idea of the uncertainty in PV generation profiles for better integration and management within the energy system. The report studies the correlation coefficients of regionally close pairwise PV panels within CHINA. It uses statistical methodology and machine learning method. It analyses influence factors and makes a regression model for correlation of outputs. We take influence factors as independent variables, which include distance of pairwise panels, weather conditions and time of year. And we take correlation coefficients as dependent variable.The primary analysis of this report consists of two sections: univariable analysis and multivariable analysis of factors. There are three parts in univariable analysis. The first part is on distance. We divide range of distance into four regions as mentioned above. By analyzing KDE and CDF and viewing the density distribution, the coefficients are around 0.5, and the coefficients from short distances are relatively higher than that in long distances, and the distribution of correlation coefficient is stable in long-distance range. The second part is on weather condition. We use MWW test to test the distribution of coefficients on different kind of weather conditions. There are 15 groups of weather conditions for coefficients. By visual analysis and hypothesis test, we find that the conditions of Fine and Showers have significant effect on the total level of the coefficients, but as we studied, in some pairs as showed in Table 4, the changing of weather condition in pairwise group does not have significant impaction of correlation coefficients. The third part is on time of year. We apply ARMA model to analyze the time series in 2014. The results are evaluated by AIC, the smaller the better. The process of the coefficients in time series is a stationary stochastic process. Also the residual series is a white noise series. The results also indicate that correlation coefficients of two panels are fluctuated around 0.5 over time, but there still exists some gaps between predicted results and true values.

Acknowledgment

The research is supported by Project of National Natural Science Foundation (61163026, 60865001).

References

[1] T. Su, W. Wang, Z. Lv, W. Wu and X. Li, “Rapid Delaunay Triangulation for Random Distributed Point Cloud Data Using Adaptive Hilbert Curve”, Computers & Graphics, **(2015)**.

[2] J. Hu, Z. Gao and W. Pan, “Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation”, Journal of Applied Mathematics, vol. 2013, **(2013)**.

[3] Z. Lv, T. Yin, Y. Han, Y. Chen and G. Chen, “WebVR-web virtual reality engine based on P2P network”, Journal of Networks. 6, no. 7 **(2011)**, pp. 990-998.

[4] J. Yang, B. Chen, J. Zhou and Z. Lv, “A portable biomedical device for respiratory monitoring with a stable power source”, Sensors, **(2015)**.

[5] D. Zhao, “FusionFS: Toward supporting data-intensive scientific applications on extreme-scale high-performance computing systems”, Big Data (Big Data), 2014 IEEE International Conference on IEEE, **(2014)**.

[6] S. Dang, J. Ju, D. Matthews, X. Feng and C. Zuo, “Efficient solar power heating system based on lenticular condensation”, Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference, **(2014)** April 26-28.

[7] X. Zhang, Y. Han, D. Hao and Z. Lv, “ARPPS：Augmented Reality Pipeline Prospect System”, 22th International Conference on Neural Information Processing (ICONIP **2015**), Istanbul, Turkey. In press.

[8] J. Hu and Z. Gao, “Distinction immune genes of hepatitis-induced heptatocellular carcinoma”, Bioinformatics, vol. 28, no. 24, **(2012)**, pp. 3191-3194.

[9] K. Wang, “Overcoming Hadoop Scaling Limitations through Distributed Task Execution”.

[10] S. Zhang, X. Zhang and X. Ou, “After we knew it: empirical study and modeling of cost-effectiveness of exploiting prevalent known vulnerabilities across iaas cloud”, Proceedings of the 9th ACM symposium on Information, computer and communications security, ACM, **(2014)**.

[11] G. Bao, L. Mi, Y. Geng and K. Pahlavan, “A computer vision based speed estimation technique for localiz ing the wireless capsule endoscope inside small intestine”, 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), **(2014)** August.

[12] W. Gu, Z. Lv and M. Hao, “Change detection method for remote sensing images based on an improved Markov random field”, Multimedia Tools and Applications, **(2016)**.

Authors

Nijiati Najimi, is currently studying at Department of Economic Management from North China Electric Power University in Beijing, China, as a Ph.D. candidate. He is currently the deputy general manager of the State Grid Xinjiang Electric Power Company Information and Communications Company. His research interest is mainly in the area of Information and communication, Electric power automation. He has published several research papers in scholarly journals in the above research areas and has participated in several books.