Research on the order parameter selection algorithm based on correlation analysis and principal component analysis——Taking the Logistics sector in Gansu Province as an example

Mu Dan [1a] Shi Chuwei [2]

[1.] Gansu Business Development Research Center, Lanzhou University of Finance and Economics, Lanzhou, Chjna
[2.] Gansu Key Laboratory of E-Business Technology and Application, Lanzhou University of Finance and Economics, Lanzhou, Chjna

Abstract—An order parameter selection algorithm based on correlation analysis and principal component analysis was designed according to the statistical analysis method, the selection principle of order parameters of social system, and the correlation test in correlation analysis and the variable contribution test in principal component analysis in this paper. The redundant variables were eliminated from the system by correlation analysis first, and then the variables with high contribution to the system were selected by principal component analysis, so the order parameters obtained accordingly not only have low information redundancy, but also reflect the actual information of the social system to the greatest extent. At the end of this paper, the logistics sector in Gansu Province was taken as an example to select the panel data from 2006 to 2015. Eight indices were extracted as the order parameters of the logistics sector in Gansu Province from the sixteen indices which are redundant selected by this algorithm. The order parameters selected by rational judgment reflect 99% of the original information. The results show that the order parameters in the social system can be correctly and reasonably selected by this order parameter selection algorithm based on correlation analysis and principal component analysis.

1 Introduction

Each phase of the system development experiences the transformation between quantitative change and qualitative change to different extent. From the point of view of the orderly development of the system, every phased development of the system is the result of its internal order parameter-oriented development. The so-called order parameter is the key variable of the leading system to complete the structural transformation and thereby achieve a higher level of order, which reflects the total contribution of all the internal factors to the cooperative motion of the system, and exhibits the ordered structure and type of the whole system. The order parameters have such characteristics, so the correct selection of the order parameters is very convenient to research various problems about the system.

For a system, the importance of the order parameter is self-evident. For the system in the phase transition period, its internal variables are divided into two types, fast relaxing variables and slow relaxing variables. The system has many fast relaxing variables and few slow relaxing variables. The fast relaxing variables have little influence on the development of the system due to their high decay rate, while the slow relaxing variables play a decisive role in the phase transition of the system due to their low decay rate and thereby are the key to guiding the development of the whole system. Therefore, we refer to these slow relaxing variables as the order parameters. Different forms of order parameters exist in the systems in the fields of social science and natural science. In the field of natural science, the order parameters may be some particles or some crystals with special structure [1]. In the field of social science, the order parameters may be some subsystems or some indices [2,3]. It can be seen that the order parameters exist in many forms. Correct identification of the order parameters in the system is very important to understand the self-organization behavior and cooperative motion of the system.

In the field of natural science, the order parameters are usually identified by using the experimental method for repeated phase transformation of the system based on the characteristics of the system itself. However, this method is often impractical and difficult to achieve in the field of social science. How to select and quantify the order parameters in the social system has become the key to study on the orderly development of the social system. In the social system, most researchers choose the indices that can comprehensively measure the social system as the order parameters [4-6]; some researchers also determine the order parameter of the system by such an exploration diagram method that all the factors of influencing the cooperative motion of the system and their interactive and hierarchical relationship are depicted in a diagram mainly based on the observation of the internal and external environment of the whole

[a]282608709@qq.com

system, knowledge structure and information, full imagination, and considerations of larger environment issues [7,3]; some researchers also select the order parameter by such a data mining algorithm that the reduction set is found by the information entropy reflected by the index data over the years, and the indices that do not change the information entropy of the system are eliminated[8, 9]. Different from the previous order parameter selection methods, an order parameter selection algorithm based on correlation analysis and principal component analysis was designed from a new perspective on the selection of order parameters mainly according to the statistical analysis method and self-organization theory in this paper. This algorithm takes the low redundancy and high contribution of the order parameters as the core idea, and its rationality and applicability are verified in the example of order parameter selection from the logistics sector in Gansu Province. Thus, the reasonable and objective order parameters can be selected for various social systems by this algorithm.

2 Order Parameter Selection Algorithm

A. Principle of Order Parameter Selection

The order parameters should be selected according to a certain principle. However, the order parameters in the systems in the fields of social science and natural science are selected according to different principles due to a great difference in the forms of systems between social science and natural science. In this paper, the selection principles of order parameter in social system are discussed. The specific selection principles are as follows:

1) Macro-representative

The order parameters are the parameters to describe the macro action of the whole system. In the synergetics, Hermann Haken believes that the order parameter can reflect the ordered structure and type of the whole system [10]. Therefore, the order parameters are macro-representative, and their change can reflect the overall development of the system to a certain extent.

2) Systematic

As the internal variables of the system, the order parameters influence one another. Such influences are mapped to the whole system as the total contribution of all the order parameter to the orderly motion of the whole system. Thus, the order parameters are systematic.

3) Quantifiable

The order parameters centrally embody the ordered motion of the system. The study on the order parameters is the key to the development of the system. Thus, the order parameters are quantifiable and thereby guarantee the operability of the orderly development of the research system.

4) Simplified

The order parameters are slow relaxing variables. The system has few slow relaxing variables, but they play a leading role in the system, that is, a few dominate the majority. Thus, the order parameter should be selected based on slow relaxing variables in order to ensure its simplicity.

5) Hierarchical

A complicated system often contains several subsystems, which contains more micro subsystems, which contain the order parameters with different meanings, functions and forms. The contributions of these order parameters will be hierarchically summarized in the whole complicated system. Thus, the order parameters are hierarchical.

B. Order Parameter Selection Algorithm

Two statistical analysis methods, correlation analysis and principal component analysis, mainly were used in this paper. On one hand, the correlation analysis describes the degree of correlation between matters, so redundant variables were eliminated from the system state variables by correlation analysis in order to ensure the simplicity of variables. On the other hand, principal component analysis can reduce the dimension of data set through the idea of dimensionality reduction and ensure the maximum contributions of variances and high contribution of variables. The algorithms are introduced as follows:

1) Correlation Analysis

Assuming there is a variable set contains n n variables in the system, the correlation coefficient between variables i and j can be calculated by the following formula:

$$\rho_{i,j} = \frac{\mathrm{cov}(i,j)}{\sigma_i * \sigma_j} \quad (1)$$

Where $\mathrm{cov}(i,j)$ is the covariance between variables i and j, σ_i, σ_j are the standard deviations of variables i and j, respectively. The covariance between and standard deviations of the variables are reflected by the data distribution of the variables. $\rho_{i,j}$ is between -1 and 1. The closer to 1 $\rho_{i,j}$is, the higher the positive correlation is; the closer to -1$\rho_{i,j}$ is, the higher the negative correlation is; the closer to 0 $\rho_{i,j}$ is,, the lower the correlation is. In most studies, 0.9 is set as the critical value of high correlation of variables[11], and thereby also used as the critical value of high correlation of variables in this paper. It is noted that correlation analysis is a data analysis method, and its correlation degree depends on the distribution of the data itself. Thus, it is likely that the data is high correlative, but the actual importance of the variables is not correlative during selection of variables by correlation analysis. Thus, the actual importance of the variables should be considered, and these variables should be retained.

2) Principal Component Analysis

Assuming there are m variables, $X = (X_1, X_2, \cdots, X_m)$, in the system, the principal component model of the variable is as follows:

$$F_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{im}X_m, \quad i = 1,2,\cdots,m \quad (2)$$

Where F_i represents the i^{th} principal component, a_{ij} is the j^{th} component of the characteristic vector corresponding to the i^{th} characteristic value, X_i is the observation value of the i^{th} variable, and m is the number

of principal components. The i[th] principal component is expressed as a linear combination of the variable X to reflect the information of the original variables [12]. The steps are as follows:

The values of the variables is normalized by zscore first to calculate the correlation coefficient matrix $(R)_{m\times m}$ of m variables and obtain the characteristic value and characteristic vector of the correlation coefficient matrix.

$$|R-\lambda I|= 0 \quad (3)$$

The characteristic root λ ranks in the descending sequence, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, where λ_i reflects the original information content represented by the i[th] principal component, i.e. the total variance of the data of the explained original variables. The corresponding variance contribution is [12]:

$$e_i = \lambda_i \Big/ \sum_{i=1}^{m} \lambda_i \quad (4)$$

The variance contribution represents the ratio of the original information content represented by the i[th] principal component to the information content represented by all the principal components. The principal components with the cumulative variance contribution of greater than or equal to 85% are retained, and the principal components corresponding to the first p characteristic values are selected to obtain the factor load b_{ij} of the i[th] variable on the j[th] principal component as follows:

$$b_{ij} = a_{ij}\sqrt{\lambda_i} \quad (5)$$

The contribution of a variable to the system can be determined by the factor load. The variable with larger absolute value of factor load more significantly influences the system. The more simplified variable set can be obtained by eliminating the variables with smaller factor load.

3) Order Parameter Selection Algorithm Based on Correlation Analysis and Principal Component

To select the order parameters from the social system, the common indices in the evaluation system have been taken as the order parameters of the system in some studies so that it is very likely that the indices are high correlative. In addition, some indices may not well reflect the actual development of the system and ultimately affect the follow-up research.

Based on the above problems, an order parameter selection algorithm based on correlation analysis and principal component analysis is designed in such an idea that on the basis of the redundant variable set selected from the system, the correlation analysis of the variables is carried out and the redundant variables in the selected redundant variable set are eliminated first to ensure the simplicity; then the principal component analysis of the remaining variables is carried out accordingly and the variables with high contribution to the system are selected by the factor load in order to obtain the order parameter of the system. This method ensures that the selected order parameters have low redundancy and high contribution.

At the end of the selection, it is necessary to determine the rationality in order to ensure that the selected order parameters can fully reflect the development of the system. The rationality is determined mainly based on whether the information content of the variance ratio test-based order parameter accounts for more than 85% of the information of the originally redundant variable sets or not [11]:

$$ln = trS_s/trS_h \quad (6)$$

Where trS_s represents the sum of variances of the selected order parameters, trS_h represents the sum of variances of the originally redundant variable sets, and the ratio of trS_s to trS_h represents the information contribution of the order parameter. For the process flow diagram of the order parameter selection algorithm based on correlation analysis and principal component, see Fig. 1.

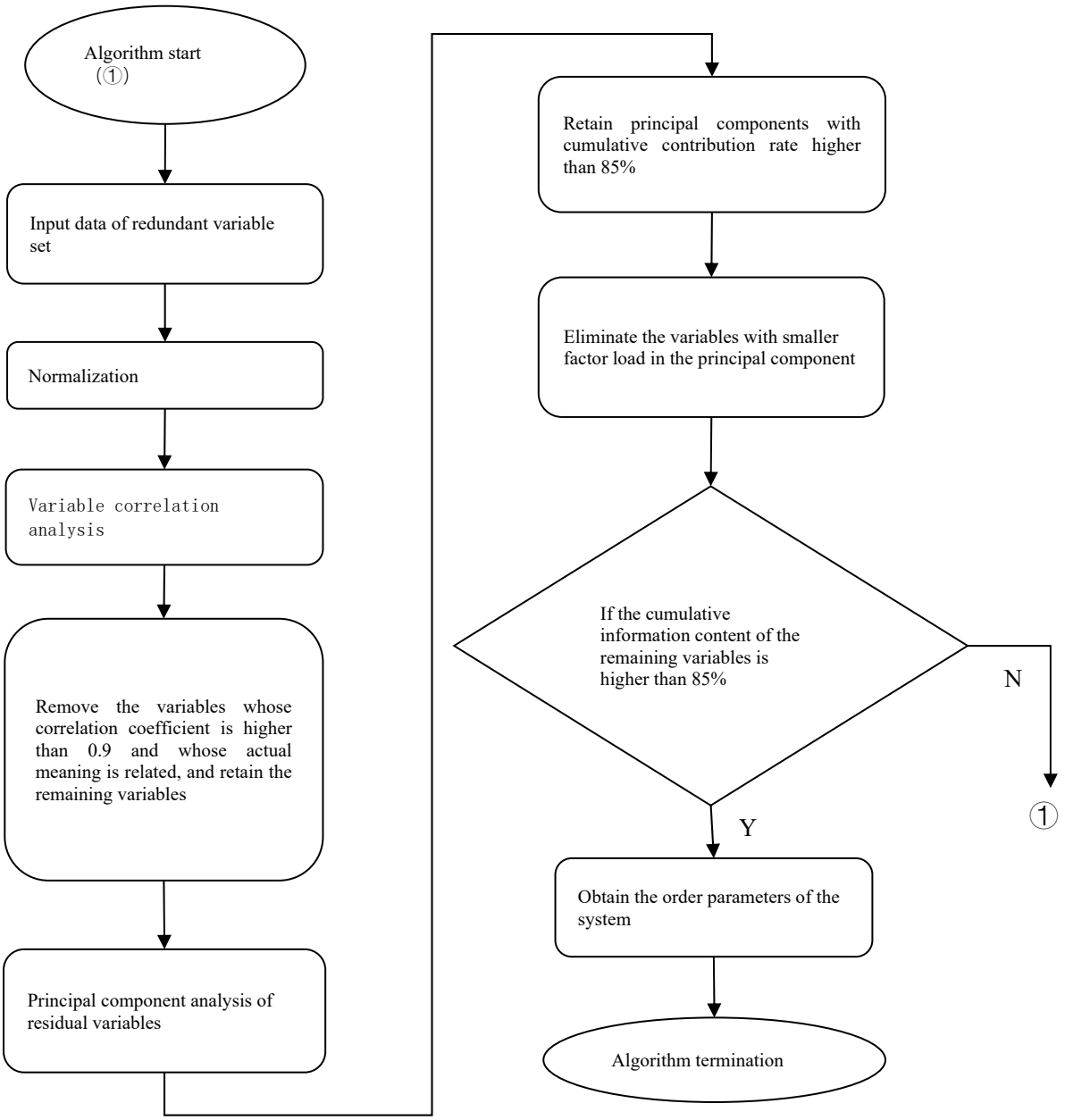

Fig. 1 Process Flow Diagram of The Order Parameter Selection Algorithm Based On Correlation Analysis And Principal Component

3 Selection Of The Order Parameters from The logistics sector in Gansu Province

According to the relevant statistical indices of the logistics sector in China *Statistics Yearbook* and the researchers' research conclusions of the index system of the logistics sector, the redundant variable set was mainly determined on the four aspects, capital scale, demand scale, infrastructure and supply capacity, as shown in Table 1.

TABLE 1 REDUNDANT VARIABLE SET SELECTED FROM THE LOGISTICS SECTOR

Primary Indexes	Secondary Indexes(Unit of Measurement)	Variable Name
Fund Size	The Value-added of the Traffic, Transport, Storage and Post Sectors(Billion yuan)	x1
	The Total Post Business Volume(Billion yuan)	x2
Demand Scale	The Fixed Assets Investment in the Traffic, Transport, Storage and Post Sectors(Billion yuan)	x3
	The Express Delivery Volume(Ten Thousand Pieces)	x4
	The Number of Packages(Ten Thousand Pieces)	x5
	The Freight Volume(Ten Thousand Tons)	x6
Infrastructure	The Turnover Volume of Freight Transport(Hundred Million Ton-kilometers)	x7
	The Number of Postal Business Outlets(Places)	x8
	The Length of Postal Routes(Kilometres)	x9
	The Railway Operating Kilometrage(Ten Thousand Kilometers)	x10
	The Highway Kilometrage(Ten Thousand Kilometers)	x11
Supply Capability	The Number of Employees in the Logistics Sector(Peoples)	x12
	The Possession of Vehicles for Highway Business Transportation(Ten Thousand Vehicles)	x13
	The Possession of Freight Cars for Highway Business Transportation(Ten Thousand Vehicles)	x14
	The Possession of Civil Freight Cars(Ten Thousand Vehicles)	x15
	The Possession of Private Freight Cars (Ten Thousand Vehicles)	x16

The panel data of Gansu Province from 2006 to 2015 were obtained from the website of National Bureau of Statistics of the People's Republic of China. The order parameters of the logistics sector in Gansu Province were selected by the order parameter selection algorithm based on correlation analysis and principal component analysis. The data were analyzed and processed by SPSS software. Table 2 shows the correlation analysis results of the redundant variable set selected from Gansu Province.

TABLE 2 CORRELATION COEFFICIENT MATRIX OF THE REDUNDANT VARIABLE SET SELECTED FROM THE LOGISTICS SECTOR IN GANSU PROVINCE

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16
x1	1.000	.694	.627	.713	-.518	.745	.952	.486	.723	.477	.858	.620	.820	.840	.873	.867
x2	.694	1.000	.912	.952	-.823	.844	.772	.873	.877	.879	.756	.756	.888	.892	.905	.910
x3	.627	.912	1.000	.966	-.743	.935	.752	.926	.957	.953	.630	.890	.931	.917	.897	.902
x4	.713	.952	.966	1.000	-.824	.929	.801	.942	.951	.939	.727	.859	.937	.928	.927	.933
x5	-.518	-.823	-.743	-.824	1.000	-.636	-.577	-.806	-.693	-.766	-.415	-.620	-.679	-.689	-.712	-.717
x6	.745	.844	.935	.929	-.636	1.000	.829	.844	.984	.831	.693	.939	.966	.962	.944	.948
x7	.952	.772	.752	.801	-.577	.829	1.000	.576	.840	.584	.858	.773	.922	.930	.944	.939
x8	.486	.873	.926	.942	-.806	.844	.576	1.000	.873	.975	.529	.798	.803	.780	.772	.783
x9	.723	.877	.957	.951	-.693	.984	.840	.873	1.000	.859	.694	.961	.980	.971	.952	.956
x10	.477	.879	.953	.939	-.766	.831	.584	.975	.859	1.000	.552	.764	.803	.778	.763	.773
x11	.858	.756	.630	.727	-.415	.693	.858	.529	.694	.552	1.000	.541	.774	.776	.802	.804
x12	.620	.756	.890	.859	-.620	.939	.773	.798	.961	.764	.541	1.000	.930	.914	.887	.889
x13	.820	.888	.931	.937	-.679	.966	.922	.803	.980	.803	.774	.930	1.000	.996	.989	.989
x14	.840	.892	.917	.928	-.689	.962	.930	.780	.971	.778	.776	.914	.996	1.000	.995	.995
x15	.873	.905	.897	.927	-.712	.944	.944	.772	.952	.763	.802	.887	.989	.995	1.000	1.000
x16	.867	.910	.902	.933	-.717	.948	.939	.783	.956	.773	.804	.889	.989	.995	1.000	1.000

As seen from Table 2, we can make the following analysis.

a) The coefficient of correlation between the value-added of the traffic, transport, storage and post sectors (x1) and the turnover volume of freight transport (x7) is 0.925, indicating a strong correlation. The increase of the turnover volume of freight transport is correlative with the increase of the output value of traffic, transport, storage and post sectors, so the two indices are high correlative. One of them can be retained. The value-added of the traffic, transport, storage and post sectors is eliminated herein.

b) The coefficient of correlation between total post business volume (x2) and the express delivery volume (x4) is 0.952, indicating a strong correlation and overlapping and inclusive relationship between x2 and x4.

c) The coefficient of correlation between total post business volume (x2) and fixed assets investment in the traffic, transport, storage and post sectors (x3) is 0.912, indicating a strong correlation. The fixed assets investment in the post section promotes the increase of total post business volume. Thus, the fixed assets investment in the traffic, transport, storage and post sectors can be eliminated.

d) The possession of vehicles for highway business transportation (x13), possession of freight cars for highway business transportation (x14), possession of

e) civil freight cars (x15), possession of private freight cars (x16) are inclusive and high correlative. Only the possession of civil freight cars is retained herein, and the remaining three should be eliminated.

f) The total post business volume (x2), possession of civil freight cars (x15) and possession of private freight cars (x16) are high correlative too, but are the evaluation indices in different fields and thereby are not correlative in actual importance and should be retained. In the same way, the freight volume (x6), length of postal routes (x9) and the number of employees in the logistics sector (x12) are not correlative in actual importance and should be retained. The railway operating kilometers (X10) and number of postal business outlets (x8) are not correlative in actual importance and should be retained.

The following indices were selected from the logistics sector in Gansu Province by correlation analysis: total post business volume, number of packages, freight volume, turnover volume of freight transport, number of postal business outlets, highway kilometer, the number of employees in the logistics sector, possession of civil freight cars, and length of postal routes. The next step of the algorithm was principal component analysis. For principal component analysis and component matrix, see Tables 3 and 4.

TABLE 3 PRINCIPAL COMPONENT ANALYSIS

Component	Initial Eigenvalue			Extract Square Sum Load		
	Total	Percent of Variance(%)	Percent of Cumulative(%)	Total	Percent of Variance(%)	Percent of Cumulative(%)
1	7.113	79.032	79.032	7.113	79.032	79.032
2	.952	10.575	89.607	.952	10.575	89.607
3	.471	5.231	94.839			
4	.320	3.552	98.391			
5	.063	.697	99.088			
6	.039	.433	99.521			
7	.035	.390	99.912			
8	.008	.086	99.998			
9	.000	.002	100.000			

TABLE 4 COMPONENT MATRIX

		x2	x5	x6	x7	x8	x10	x11	x12	x15
Component	1	.952	-.494	.950	.864	.900	.893	.796	.891	.968
	2	-.066	.390	.059	.451	-.384	-.352	.525	-.011	.202

After extraction of the principal components 1 and 2, the cumulative contribution exceeds 85%. It is found from the factor load of each index in Table 4 that the number of packages (x5) has a smaller factor load in the principal components 1 and 2 and thereby can be eliminated first, while the other indices have bigger factor loads and higher contributions to the system, reflect an aspect of development of the logistics sector and thereby should be retained. It is found from identification of the rationality according to Formula (6) after primary selection of the order parameters that the information content of the selected order parameters accounts for 99% of the information content of the originally selected index set, which meets the requirements for identification of the rationality. Finally, the following order parameters of the logistics sector in Gansu Province are determined: total post business volume, freight volume, turnover volume of freight transport, number of postal business outlets, highway kilometers, the number of employees in the logistics sector, possession of civil freight cars, and length of postal routes.

The development of the logistics sector in Gansu Province mainly relies on highways and railways. The logistics sector in Gansu Province, which is a mountainous province, mainly relies on motor vehicles. Moreover, the post sector is a pillar sector in the logistics sector in Gansu Province. The order parameters selected from the post sector comply with the actual development of Gansu Province. Thus, the order parameters selected by this algorithm not only have high contribution and low redundancy, but also reflect the actual development of the system and have rationality and availability.

4 Conclusion

Correct selection of the order parameters from the system is very important to the orderly development of the research system. If the index system of the evaluation system is taken as the order parameters of the system, the follow-up research will deviate due to redundancy,

contribution, etc. According to the principle for selection of the order parameters from the social system, the order parameter selection algorithm based on correlation analysis and principal component analysis not only achieves the low redundancy of the order parameters, but also the high contributions of the order parameters to the system.

Acknowledgement

This research was funded by The Application open fund subject and Silk Road Economic Research Institute Project of Lanzhou University of Finance and Economics(JYYY201807) and The Open fund project of Gansu Key Laboratory of E-Business Technology(2018GSDZSW63A13).

References

1. Cai Yuping. Symmetry and Order Parameters of DisPlcaive Phsea Trasnit. Diss. Hebei University of Technology, 2000.
2. Diao Xiaochun, and Su Jingqin. " Evolutionary form of eco-industrial networks based on order parameter identification." Studies in Science of Science 03(2008):62-66.
3. Sun Bing, and Zhen Chuiyong. " Study on the order parameters of innovation system of industrial cluster." Statistics and Decision, 06(2009):142-144.
4. Song Mingzhen. Evaluation of Coordinated Development Between Regional Logistics and Regional Economic Taking Chengdu for Example. Diss. Southwest Jiaotong University, 2015.
5. Fu Weizhong. and Li mengyu. " Research on Collaborative Development of Regional Logistics and Regional Economy of Beijing-Tianjin-Hebei" Journal of Hefei University of Technology (Social Sciences) 030.006(2016):1-8.
6. Zhong Ming , Wu Yanyun , and Luan Weixin. " Model of synergy degree between port logistics and urban economy." Journal of Dalian Maritime University 37.1(2011):80-82.
7. Xu Xusong, and Zheng Xiaojing. "Qualitative Analysis Tools: Probe Graph, Cycle Graph and Structure Graph." Technology Economics 05(2010):3-8.
8. Lan Hongjie, Liu Zhigao, and Wang Ruijiang. " Research on Food Cold Chain Logistics System Collaboration Based on Attribute Reduction in Rough Set." Journal of Beijing Jiao Tong University (Social Sciences Edition) 02(2010):36-40.
9. Li Diansheng , and Zhang Lihong. " A Study on Port Coupling Level Based on Rough Set Theory." Journal of Ocean University of China(Social Sciences) 3(2012).
10. Haken, Hermann . Synergetics. Springer, 1983.
11. Chi Guotai, Cao Tingting, and Zhang Kun. " The establishment of human all-around development evaluation indicators system based on correlation-principle component analysis." Systems Engineering-Theory & Practice 32.1(2012):111-119.
12. Tong Qihui. " The Application of Primary Element Analyzing Methods in the Index Synthetic Evaluation." Journal of Beijing Institute of Technology (Social Sciences Edition) 4.1(2002).