INTELLIGENT AUTOMATION & SOFT COMPUTING, 2017
http://dx.doi.org/10.1080/10798587.2016.1267444

The Study on Evaluation Method of Urban Network Security in the Big Data Era

Qingyuan Zhou[a] and Jianjian Luo[b]

[a]Department of Economics and Management, Changzhou Administrative College, Changzhou, 213001, China; [b]Department of Information Management, Changzhou College of Information Technology, Changzhou, 213001, China

ABSTRACT

Big data is an emerging paradigm applied to datasets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. In a Smarter City, available resources are harnessed safely, sustainably and efficiently to achieve positive, measurable economic and societal outcomes. Most of the challenges of Big Data in Smart Cities are multi-dimensional and can be addressed from different multidisciplinary perspectives. Based on the above considerations, this paper combined the PSR method, the fuzzy logic model and the entropy weight method in an empirical study for feasible urban public security evaluation modeling. The PSR method was used to establish an evaluation index system regarding the essence of public security. The Entropy method was used in the weighing assignment process to verify the objectivity of this modeling. The fuzzy method was used for the quantitative analysis to determine the fuzziness of urban public security.

KEYWORDS

Urban Public Security; Big Data; Evaluation Method

1. Introduction

Big data is an emerging paradigm applied to datasets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Such datasets are often from various sources (Variety) yet unstructured such as social media, sensors, scientific applications, surveillance, video and image archives, Internet texts and documents, Internet search indexing, medical records, business transactions and web logs; and are of large size (Volume) with fast data in/out (Velocity). More importantly, big data has to be of high value (Value) and establish trust in it for business decision-making (Veracity). Various technologies are being discussed to support the handling of big data such as massively parallel processing databases (Xu et al., 2016a), scalable storage systems (Xu et al., 2016b), cloud computing platforms (Zhou & Luo, 2015a), and MapReduce. Big data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make business more agile, and to answer questions that were previously considered beyond our reach (Huang et al., 2016). Distributed systems is a classical research discipline investigating various distributed computing technologies and applications such as cloud computing and MapReduce. With new paradigms and technologies, distributed systems research keeps going with new innovative outcomes from both industry and academia. For example, wide deployment of MapReduce is a distributed programming paradigm and an associated implementation to support distributed computing over large big datasets on cloud.

Data (and then information) from people, systems and things in cities is the single most scalable resource available to City stakeholders, but difficult to publish, organize, discover, interpret, combine, analyze, reason and consume, especially in such an heterogeneous environment. Indeed data is big and exposed from heterogeneous environments such as water, energy, traffic or building. Most of the challenges of Big Data in Smart Cities are multi-dimensional and can be addressed. While research efforts in Big Data have mostly focused on the later stages of the process of making sense of the sea of data (e.g. data analytics, query answering, data visualization, etc.), in the context of Smart Cities, where heterogeneous data originates from multiple municipal and state agencies with little to no coordination, major hurdles and issues continue to impede progress toward these later stages. These key unaddressed issues are often related to information exploration, access, and linking. e.g.

(1) How to efficiently figure out and access data sources relevant to the urban public security?
(2) How to discover relations between these information sources at the data level?
(3) How to evaluate the effectiveness and efficiency of urban public security?

Today, these challenges are tackled in mostly ad hoc and labor intensive data integration efforts. It is becoming increasing clear that, without the advent of novel, scalable and semi-automated data integration techniques, this first data access and linking stage will soon represent a major bottleneck to the whole process of extracting valuable information from the increasing number data sources and volume of data available to decision makers. At present, network security situation assessment has not a unified, comprehensive definition. Network security situation assessment is to extract the network and security related factors to analyze and understand the entire network security status, and forecast the future development of network security situation. Its research focuses on the efficient organization and overall assessment of all kinds of

CONTACT Qingyuan Zhou zqy@jsut.edu.cn

complex information, which the purpose is to shorten the time from obtaining the information to make decisions. It mainly applies information fusion technology that includes intrusion detection log, firewall log, virus log, network scanning, illegal external link, and running state of equipment and real-time alarm, multi-source heterogeneous observational data. Traditional network security situation assessment provides a lot of analysis convenience for security administrator, but most of them are based on the analysis of system log. There are many problems that data source is single, real-time performance is poor, and assessment results are too dependent on experience of network management personnel (Zhou, 2016). At present, the research direction of network security is also a major change, from the initial passive security system construction to intrusion detection, defense attack. Situation assessment is one of the active security system constructions. Research on the security situation of global network has transferred to a single security problem. But the security situation assessment is still at the starting stage at home or abroad, the relevant technical theory is not mature, so it is urgent to find an effective and accurate assessment method.

Based on the above considerations, this paper combined the PSR method, the fuzzy logic model and the entropy weight method in an empirical study for feasible urban public security evaluation modeling. The PSR method was used to establish an evaluation index system regarding the essence of public security. The Entropy method was used in the weighing assignment process to verify the objectivity of this modeling. The fuzzy method was used for the quantitative analysis to determine the fuzziness of urban public security. This paper analyzed both the temporal and spatial dynamics in this area. Not only was the study case novel, but both the research issue and combined method were original. The results are interesting, and some recommendations are given at the end. In order to show this approach, we propose a case study on mobile data system to deal with the emissions in a city transportation environment. This is based on the usage of the tracking big data generated from users' mobile devices. Our main idea is to simplify the process of the ITS systems' development, and, consequently, to increase its availability to users. This approach also allows for the joining and aligning of the works of field experts on generating the mobility of big data and merging it with data from mining experts, in order to better extract knowledge from the collected data. This can be achieved by data discretization in predefined classes, performed by these field experts. This approach can also be applied for non-professional cases, such as personal data tracking, allowing its extension to the mobility habits of millions of users just by carrying cell phones in their pockets.

The organization of this paper is as follows: Related work is given in the Section 2. Section 3 introduces the basic method of the proposed model. Section 4 illustrates a platform using the proposed method. Section 5 presents a case study on mobile data. Last section makes the conclusion.

2. Related Work

The term public security was first proposed by the government of the United States (Ezeonu & Ezeonu, 2000). After that, although it has been widely studied (Kullenberg, 2012), a common accepted definition has not been given (Zhao et al., 2006). A large number of academic studies have focused on the driving forces of urban expansion (Chen, Jin, Qiu, & Chen, 2014). The environmental and public issues of the landscape changes resulting from urbanization are significant (Li et al., 2009). Sun, Wu, Lv, Yao, & Wei (2013) use network equipment service and vulnerability information using the improved evidence theory to fuse information of each node. It calculates the global overall network situation firstly, and then uses time series analysis to achieve the trend forecast. A quantitative assessment method of hierarchical network security situation threat is proposed in Zhou & Luo, (2015b). The security of service, host and network is assessed quantitatively by using intrusion detection information, network performance index and host vulnerability information. Zhang, Yang, & Li (2006) design a complete security situation assessment system prototype based on the lack of complete security situation generation framework, using D-S evidence theory to construct a network security situation assessment model. Through the knowledge discovery method to mine data-set frequent or sequence pattern, it realizes network security view's automatic generation. In Xie, Wang, & Yu (2013), a new method of network security situation assessment based on neural network is proposed, which is based on the comparison and analysis of network security situation assessment. Using RBF neural network to find nonlinear mapping relationship of network situation, it adapts genetic algorithm to optimize the network parameters and assess the network security situation (Gong et al., 2009). Existing network security situation forecast method cannot accurately reflect the changes of future security situation (Zhao et al., 2006). But it is not very good to deal with the relationship between security elements and a future network security situation. A forecast method of a network security situation based on temporal and spatial analysis is proposed. Existing network security situation forecast method can- not accurately reflect the changes of future security situation, but it is not very good to deal with the relationship between security elements and a future network security situation. A forecast method of network security situation based on temporal and spatial analysis is proposed. Assessment method is the core of network security situation assessment, anglicizing many factors such as network resources and other factors, which make reasonable explanation to the current security state of whole network (Liu et al., 2014). The main problem of traditional assessment method is that assessment range is limitation, information source is single, time and space complexity is large, and credibility is not high.

3. Basic Methodology

3.1. Pressure–state–response (PSR) for Urban Security Index

Various methods have been used for public security evaluation: Exposure-response method, integrated index number method, public capacity analysis and some public models. After comparing the other methods, the authors recommended the pressure–state–response (PSR) concept model to establish the urban public security index due to its clear structure and logic. In addition, the analysis resulted in a PSR framework that can be easily understood and applied by decision makers in practice. The concept of the environmental PSR index was proposed by Organization of Canadian Economic Cooperation and Development (OECD) and the United Nations Environment Program (UNEP). The index system was established from three aspects that affect or are related to urban public security, i.e., the

urban public pressure (the pressure from population growth and environmental resource assumption) the eco-environment state and the response (measures and policies that are adopted to solve eco-environmental issues). With the help of expert consultation, frequency statistics and literatures investigation, the evaluation index of the BTH region was established. In addition to the essence of the urban public security assessment, the choice of this index system was considered via other two aspects. The first aspect was the acceptability of the data. As there were changes in yearbooks during the study case period, some indicators did not have continuous data. The second aspect was the comparability of the data. Compared to scale indicators, percentage indicators are easily used in comparison between cities with different developed scales.

The key issue in evaluating public security (ES) from the 18 observed indicators was how to weigh each indicator with minimal subjectivity. The entropy-based weight method was used. In natural sciences, thermo dynamic entropy is used to measure the disorder of a system. In social sciences, entropy information means the degree of uncertainty of a system. It is generally believed that the more entropy information one system has, the more balanced is its structure and the smaller is its difference. From this perspective, the entropy information for each index was calculated, and smaller entropy means greater weight. The urban public security evaluation method evolved from a qualitative description in the early stage to today's quantitative model analysis, which mainly includes fuzzy synthesis, gray relation method, matter-element evaluation method, mutation progression method, landscape analysis, public foot print method, spatial statistical methods, etc. As the concept of "security" is fuzzy, i.e., not exact, there is no common idea about the exact threshold to divide safety and danger. Therefore, a fuzzy division for different security levels is required, and this paper chose the fuzzy synthesis evaluation method. Fuzzy logic usually contains fuzzification, the application of the rule base to fuzzy data, the inference of fuzzy results and the defuzzification of fuzzy results. Fuzzification is a process that transforms the observed (real) data into a fuzzy form using a membership function that is defined by the feature of the tar-get question. The rule base defines the relationship among the membership functions and the form of the resulting membership function. Defuzzification provides the real value from the resulting membership function. The establishment of the membership function is the key step in fuzzy evaluation. The study used up half of a trapezoid membership function and the security was divided into 1–5 scales in which 1stands for insecure and 5 stands for secure. Because the selected evaluation index of public security has positive and negative effects, the membership function can be divided into positive and negative membership functions. There is currently no standard in urban public security evaluation. Previous studies have mainly been based on the international, national, and local and industry rules of urban development and evaluation standards, such as environmental quality, public health standards issued by the national or international organization, environmental safety evaluation standard and regulation issued by various industries, governmental planning goals, regional background values and local standards. The analogical standard uses the ecosystem with no serious human interference as the standard of public security. According to relevant standards of urban security evaluation, we developed an urban public security evaluation index system and classification standard for our research region.

3.2. Situational Factor Selection

In practical application, the information of network security situation assessment is generally derived from three major categories: The running information, the configuration information and the related IDS system information. This information can be acquired through the distribution of Netflow, IDS, Firewall, VDS and other components in the network.

Indexes related to running information such as CPU utilization, memory, network traffic flow rate, the total number of sub network data flow, the distribution of different sizes of data packets in the sub network, etc., constitute the component basis running ability(Run); Indexes related to configuration information such as vulnerability, system configuration, security software installed, the number of key equipment vulnerabilities and level, the number of sub network security devices, and the total number of key equipment in the sub network, constitute component vulnerability; Indexes related to IDS system log library such as DDOS, worm attack, Trojan horse and common virus, subnet bandwidth usage, network data flow rate, and the rate of net inflow into the IDS system log library, constitute the component threat ability. In order to facilitate the study, according to the probability theory knowledge, each observation index corresponding to a random variable X, take two observation value of discrete or continuous type.

In the process of network security situation assessment, there are many complex observation indexes, which are conflicting, non-public and non-determination. Some index of network security situation assessment plays a very important role, and some index of network security situation assessment is minimal. There may exist redundancy between indexes. In practical application, the index is too many to produce huge space and time complexity, so it is necessary to select some index with typical representative. The indexes form required factors for network security situation assessment, through excluding some redundancy of security situation assessment.

To determine whether there is redundancy for observation indexes of x_i and x_j, it can be judged by calculating the correlation coefficient of them. If their absolute value of correlation coefficient ρ is very close to 1, indicating there is redundancy, need to eliminate an index.

The correlation coefficients ρ of x_i and x_j are calculated:

$$\rho_{xixj} = Cov(x_i, x_j) / \sqrt{D(x_i) * D(x_j)}$$

$Cov(x_i,x_j)$ is the covariance of x_i and x_j. There is:

$$\begin{aligned} Cov(x_i, x_j) &= E\{[x_i - E(x_i)][y_j - E(x_j)]\} \\ &= E(x_i x_j) - E(x_i)E(x_j) \end{aligned}$$

It has expression:

$$E(x_i) \approx \overline{x_i} = \frac{1}{n}(x_{i1} + x_{i2} + \cdots + x_{in})$$

$$D(x_i) \approx S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \overline{x_i})^2$$

$$E(x_i x_j) \approx \overline{x_i x_j} = (x_{i1}x_{j1} + x_{i2}x_{j2} + \cdots + x_{in}x_{jn})/n^2$$

$E(x_i)$, $E(x_j)$ and $D(x_i)$,$D(x_j)$ represent the mathematical expectation and variance of x_i and x_j respectively. Supposed x_i and

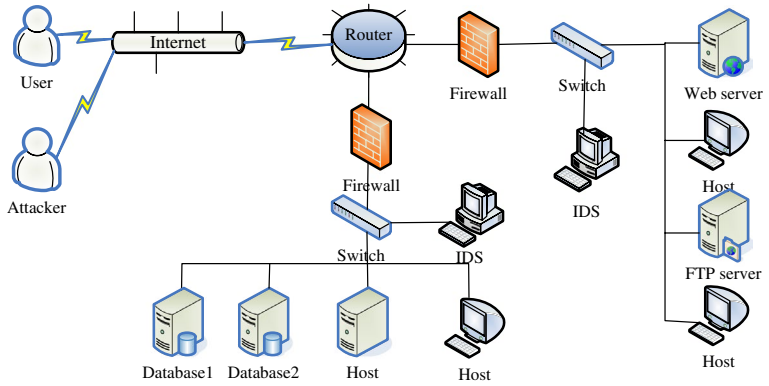

Figure 1. Experimental Network Topology.

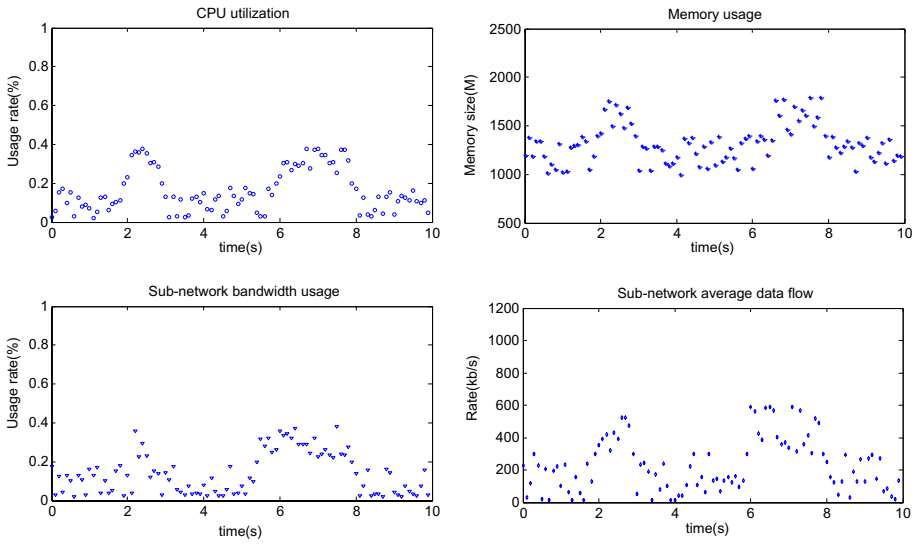

Figure 2. Original Data Sampling Diagram.

x_j have n observation data, it can use $\overline{x_i}$ to estimate $E(x_i)$ of x_i mathematical expectation and use S_i^2 to estimate variable $D(x_i)$. Since the product of two random variables x_i and x_j have the number of n^*n data, it can use $\overline{x_i x_j}$ estimate mathematical expectation $E(x_i x_j)$ of $x_i x_j$. Given a real number $0 < \varepsilon < 1$, if correlation coefficient $|\rho_{xixj}| > \varepsilon$, variable x_i can be considered very relevant to x_j, that retain only a observation index playing major role, called a situation factor.

4. Simulation Experiment

In order to verify the rationality and correctness of proposed abnormal detection technology and algorithm for network security situation, we use MATLAB 7 to simulate an experiment with setting up a network experimental environment (as shown in Figure 1). In this environment, the security situation is assessed experiment quantitatively. Ordinary user (User) and attacker (Attacker) can access the network host through Internet. Specific attack steps are as follows:

(1) Ordinary users can access the server Web server, the main database and file transfer FTP server;
(2) SQL injecting loopholes to attack the main database server;
(3) SQL injecting attacks from the host (Host) to the backup database server Database2;
(4) UDP flooding attack Web server;
(5) The worm attack on the FTP server;
(6) Using the (2) ~(5) to attack the main database, server Web server and FTP server et al.

Collecting the IDS attack information, Nessus scanning information, Snort log alarm information and router Netflow network traffic information, it can be as this simulation source of multi-source heterogeneous data sources.

(1) Original data

Attacker can launch various attacks against the network environment constantly. In order to draw clearly, it collects 300 samples dynamically every 10 s. Taking four observable indexes (CPU utilization, sub network bandwidth usage, average data flow of the sub network, scanning alert information) sample, the original data is shown in Figure 2. When the host is attacked, the sample real time data can produce some fluctuation, and the data is related with each other:

(2) Discrete data

For the continuous original sample data in Figure 2, it can be processed to a real number between 0~1 with the 2.4 section formula (1), that get five corresponding discrete values. For ease of expression, the data in Figure 2 are all translated to the corresponding position in the middle value. Instead of taking the discrete value directly or else it may become a broken line, that the difference doesn't be expression between the data, as shown in Figure 3. After discretization, the data in the corresponding discrete value are near upper and lower amplitude

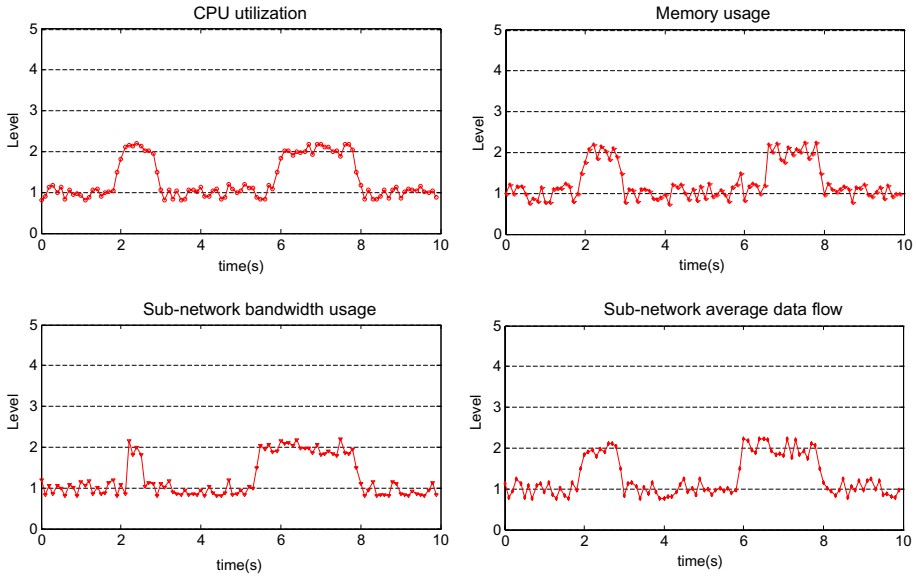

Figure 3. Discrete Data Sampling Diagram.

fluctuations. In this application, the data on the fluctuation of level of *i* near the upper and lower levels is taken as the discrete value *i*, which is convenient and easy to operate.

Abnormal data continue to flow in; there are a certain number of abnormal situations in the network, through the situation factor collection and after discretization. All the situation factors are at a certain time.

5. The Case Study on Mobile Data for its Security

From the mobile device sensor data, it is possible to identify the transportation mode that the user takes to go from A to B. All of these approaches use past data to build a classification model that identifies the transportation mode and most of these approaches use a combination of GPS and accelerometer data from the three axes. From this data, it is possible to calculate the speed and the position. We apply our methodology based on the discrete approach using the predefined data classes.

We describe the current approach based on the case studied where we have a training set of 250 cases: 60 represent car travel, 50 buses, 30 trains, 40 underground, 25 walking, 20 boats, and 25 motorcycles. Since we want a generic approach, the main effort to perform is the transformation of raw data into these pre-defined classes. These classes can be increased to cover new situations, and when not used should be treated as empty fields. Speed information from GPS is used to differentiate among walking, bicycling, boat, and other transportation modes. Periodic stops are used to differentiate among car/motorcycle and underground, bus or train. Motorcycle is better discerned from car transportation in high traffic periods, because the average speed is higher and the position pattern is different. In order to distinguish metro from train, we use the following heuristics: (1) Underground usually runs below ground without a GPS signal; (2) Distances between stops in underground transportation are usually smaller; (3) Altimetry information. We have all GPS data and time stored in a user mobility profile, in a cloud database, with the information about time, routes (XML graph with time and GPS coordinates). To this data we can add transportation mode and carrier, where we store the information about distance (in km) per transportation mode, performed in a month period. For example, for a certain user we have the following data for a specific month (May 2014); car (246 km), bus (23 km), underground (41 km), walking (2 km) and biking (15 km). It is possible to present the route representation for that month with associated information of the transportation mode, the number of times the route was performed, and also the temporal periods. Thus, it is possible to represent the time that a user spent in a month in these locations. This information represents the user mobility activity captured by the mobile device sensors. We are aware that, if the user turns off the data acquisition, or if the mobile device runs out of power, mobility data will be lost.

Based on the speed data and the location information (GPS data), it is possible to identify traffic situations as conditions that occur on road networks as the number of vehicles increase, and are characterized by slower speeds and longer trip times. From the information of a current road (matching of position against available routes), it is possible to check traffic situations using the relation of current speed divided by the maximum road speed (we call this ratio *Vt*). If the average *Vt* of current users in a specific road is below 0.25, the system tries to classify this traffic into four classes: (1) Green if the user's average *Vt* is above 0.25; (2) Yellow for average *Vt* between 0.1 and 0.25; (3) Red when the average *Vt* is below 0.1; (4) Black when average *Vt* is zero. We disregard $Vt = 0$ at public transportation stops if transportation mode is a public transportation, as well as at road intersections, because we assume (for simplification purposes) the existence of a traffic light on each road intersection. To avoid this we would need the GPS coordinates of the traffic lights, which are taken from road graph information. There may be times where the number of online users is reduced, which results in a reduction of information in the user database. Thus, it is necessary to use external information from traffic web services. The current average speed information of the users in a given route is used as input in the route advisor.

6. Conclusions

In a Smarter City, available resources are harnessed safely, sustainably and efficiently to achieve positive, measurable economic and societal outcomes. Most of the challenges of Big Data in Smart Cities are multi-dimensional and can be addressed from different multidisciplinary perspectives e.g., from Artificial Intelligence (Machine Learning, Semantic

Web), Database, Data Mining to Distributed Systems communities. Enabling City information as a utility, through a robust (expressive, dynamic, scalable) and (critically) a sustainable technology and socially synergistic ecosystem could drive significant benefits and opportunities. Based on the above considerations, this paper combined the PSR method, the fuzzy logic model and the entropy weight method in an empirical study for feasible urban public security evaluation modeling. The PSR method was used to establish an evaluation index system regarding the essence of public security. The Entropy method was used in the weighing assignment process to verify the objectivity of this modeling. The fuzzy method was used for the quantitative analysis to determine the fuzziness of urban public security.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Qingyuan Zhou received Ph.D. degrees from Sichuan University (SCU), Chengdu in 2014, respectively. Currently, he is an associate professor in Changzhou Administrative College, China. He has joined and accomplished 5 national and 6 provincial research programs. He has also authored/co-authored over 50 papers in international/national journals and conferences. His current research interests include cluster computing, risk management and optimization theory.

Jianjian Luo received Ph.D. degrees from Shanghai University, Shanghai in 2007, respectively. Currently, he is an associate professor in Changzhou College of Information Technology, China. He has joined and accomplished 2 national and 9 provincial research programs. He has also authored/co-authored over 30 papers in international/national journals and conferences. His current research interests include optimization theory and cluster computing.

References

Chen, T., Jin, Y.Y., Qiu, X.P., & Chen, X. (2014). A hybrid fuzzy evaluation method for safety assessment of food-waste feed based on entropy and the analytic hierarchy process methods. *Expert Systems with Applications, 41*, 7328–7337.

Ezeonu, I.C., & Ezeonu, F.C. (2000). The environment and global security. *The Environmentalist, 20*, 41–48.

Gong, J.Z., Liu, Y.S., Xia, B.C., & Zhao, G.W. (2009). Urban ecological security assessment and forecasting, based on a cellular automata model: A case study of Guangzhou. *China Ecology Model, 220*, 3612–3620.

Huang, D., Yang, Y., Tang, L., Zhang, J., & Wang, X. (2016). An approximation method of optimal scheduling for multicommodity flows in cloud-service scenarios. *Intelligent Automation & Soft Computing, 22*, 143–152.

Kullenberg, G. (2002). Regional co-development and security: A comprehensive approach. *Ocean Coastal Manage, 45*, 761–776.

Li, H., Apostolakis, G.E., Gifun, J., Van Schalkwyk, W., Leite, S., & Barber, D. (2009). Ranking the risks from multiple hazards in a small community. *Risk Analysis, 29*, 438–456.

Liu, Y., Feng, D., Lian, Y., Chen, K., & Wu, D. (2014). Network situation prediction method based on spatial-time dimension analysis. *Journal of Computer Research and Development, 51*, 1681–1694.

Sun, C., Wu, Z.-F., Lv, Z.-Q., Yao, N., & Wei, J.-B. (2013). Quantifying different types of urban growth and the change dynamic in Guangzhou using multi-temporal remote sensing data. *International Journal of Applied Earth Observation and Geoinformation, 21*, 409–417.

Xie, L., Wang, Y., & Yu, J. (2013). Network security situation awareness based on neural networks. *Journals of Tsinghua University Science and Technology, 53*, 1750–1760.

Xu, Z., Mei, L., Hu, C., Liu, Y. (2016a). The big data analytics and applications of the surveillance system using video structured description technology. *Cluster Computing, 19*, 1283–1292. doi: 10.1007/s10586-016-0581-x

Xu, Z., Hu, C. & Mei, L. (2016b). Video structured description technology based intelligence analysis of surveillance videos for public security applications. *Multimedia Tools and Applications, 75*, 12155–12172. doi: 10.1007/s11042-015-3112-5

Zhang, Y., Yang, Z.F., & Li, W. (2006). Analyses of urban ecosystem based on information entropy. *Ecological Model, 197*, 1–12.

Zhao, Y.Z., Zou, X.Y., Cheng, H., Jia, H.-K., Wu, Y.Q., Wang, G.-Y., … Gao, S.-Y. (2006). Assessing the ecological security of the Tibetan plateau: Methodology and a case study for Lhaze County. *Journal of Environmental Management, 80*, 120–131.

Zhou, Q. (2016). Research on heterogeneous data integration model of group enterprise based on cluster computing. *Cluster Computing, 19*, 1275–1282. doi:10.1007/s10586-016-0580-y

Zhou, Q., & Luo, J. (2015a). Artificial neural network based grid computing of E-government scheduling for emergency management. *Computer Systems Science & Engineering, 30*(5), 327–335.

Zhou, Q., & Luo, J. (2015b). The risk management using limit theory of statistics on extremes on the big data era. *Journal of Computational and Theoretical Nanoscience, 12*, 6237–6243.