International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.201-210
http://dx.doi.org/10.14257/ijdta.2016.9.8.20

A Method of Network Public Opinion Analysis Based on Quantum Particle Swarm Algorithm Optimization Least Square Vector Machine

Bo Li[1,2], BaoXing Bai[1], Changsheng Zhang[3] and Yixue Jiang[2]

[1]*Changchun University of Science and Technology*
[2]*Changchun Institute of Technology*
[3]*College of Information Science & Engineering, Northeastern University*
boer0321@sohu.com

Abstract

Prediction of network public opinion is a complicated prediction featuring poor information, small samples and uncertainty. A prediction model of network public opinion based on grey support vector machine (SVM) is specified to increase prediction accuracy. First, network data are preprocessed by text clustering, hotpot extraction and data aggregation. Then a time series model GM(1,1) is established and SVM is used to modify prediction outcomes of GM(1,1). At last, simulation experiment is conducted to test performance of the model. Simulation results indicate that grey SVM improves the prediction accuracy of network public opinion compared with traditional prediction models. The predictions have certain practical values.

Keywords*: network public opinion; Grey model; Support vector machine (SVM); Prediction*

1. Introduction

Network public opinions become increasingly frequent along with rapid development of Internet and openness of network. Unlike traditional public opinion, network public opinion is gusty, direct and real time, and negative opinions on the Internet will threaten public security without correct direction and supervision. Currently, it has been a hot research topic to predict the tendency of network public opinion. Researches on this topic increase every day and they roughly fall into two types: traditional prediction and modern prediction. In traditional prediction, data of network public opinions are translated into time series and prediction methods like autoregressive moving average and exponential smoothing are used to specify models. Though these methods are simple and easy to practice, they assume that the development of network public opinion be a linear change, which does not conform to the reality and cannot get accurate predictions. Modern predictions establish models based on the nonlinear theory and achieve higher prediction accuracy than traditional methods. Major prediction models include hidden Markov model, k nearest neighbors, neural network model and SVM. Due to poor information, uncertainty and small samples in network public opinions and in order to improve prediction accuracy further, some scholars established combinational predication models by combining advantages of single model based on combinatorial optimization theory. For instance, Li Zhen specified a prediction model by combining grey model and weighted Markov model, and obtained fair good predictions. SVM is a modern machine learning algorithm specific to uncertain predictions with small samples based on modern statistical theory. It is widely applied to nonlinear time series predictions.

Considering changing characteristics of network public opinion and in order to improve prediction accuracy, this research attempts to combine grey model with SVM model to establish a new model for prediction of network public opinion. Firstly, a time

ISSN: 2005-4270 IJDTA

series model GM(1,1) is established and SVM is used to modify prediction outcomes of GM(1,1). At last, living prediction examples are used to test performance of the model.

2. Preprocessing of Network Public Opinion Data

2.1. Text Clustering

Information is collected by network crawler and get a group of messy and unordered network public opinion data. Correlation is to be built between these data through text clustering. This research employs hierarchical clustering algorithm to cluster network public opinion data and the clustering quality is evaluated by data purity. Purity is defined as:

$$P(S_r) = \frac{1}{n_r} \max\left(n_r^i\right) \tag{1}$$

where, n_r denotes the number of documents in clustering category r and n_r^i denotes the number of documents belonging to predefined category i and distributed to category r.

Then, the purity of overall results is defined as:

$$purity = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r) \tag{2}$$

2.2. Hotpot Abstraction

Hot network topics refer to information to which certain groups of people pay continuous attention. They can be acquired by following steps:

(1) Define the frequency (TF) and time (TD) of topics being talked and network click rate (CR) as features of hot topics, count up these features;

(2) Calculate values of media attention meter (MAM) and public attention meter;

(3) Set up proportional balance factors and threshold value, and calculate sub-public attention meter (SAM);

(4) The topic is defined as hot topic if its SAM is larger than the threshold value.

2.3. Data Aggregation

Network public opinion information with different variables is organized through data aggregation, and data aggregation software are used to turn the information into discrete-time series of hot topics.

3. Grey SVM Model

3.1. GM (1,1) Model

GM(1,1) is the simplest and most commonly used grey model. It is consisted of a univariate differential equation, a special case of GM(1,N). Assume original data series be GM(1,1) and modelling series be $x^{(0)}$, that is :

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \ldots, x^{(0)}(n)) \tag{3}$$

One-accumulation of $x^{(0)}$ generating 1-AGO series, introducing $x^{(1)}$, then

$$x^{(1)} = (x^{(1)}(1), x^{(1)}(2), \ldots, x^{(1)}(n)) \tag{4}$$

where, $x^{(1)}(1) = x^{(0)}(1)$; $x^{(1)}(k) = \sum_{m=1}^{k} x^{(0)}(m)$。

Define $z^{(1)}$ as average series of $x^{(1)}$ to get $z^{(1)}$ as below:

$$z^{(1)}(k) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1) \tag{5}$$

Then, the grey differential equation model of GM(1,1) is:

$$x^{(0)}(k) + az^{(1)}(k) = b \tag{6}$$

The equation above is substituted by $k = 2,3,...,n$ and get:

$$\begin{cases} x^{(0)}(2) + az^{(1)}(2) = b \\ x^{(0)}(3) + az^{(1)}(2) = b \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ x^{(0)}(4) + az^{(1)}(2) = b \end{cases} \tag{7}$$

The equation above can be transformed into following matrix equation:

$$y_N = BP = [x^{(0)}(2), x^{(0)}(3), ..., x^{(0)}(n)]^T \tag{8}$$

Where, B, y_N P and is data matrix, data vector and parameter vector respectively, and

$$\begin{cases} B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdots & \cdots \\ -z^{(1)}(n) & 1 \end{bmatrix} \circ \\ P = [a,b] \end{cases}$$

Least square method is used to get:

$$P = (a,b)^T = (B^T B)^{-1} B^T y_N \tag{9}$$

Put $P = [a,b]$ into the equation (4) and solve the differential equation to get the content type expression of GM(1,1):

$$\hat{x}^{(0)}(k) = u^{k-2} \cdot v \tag{10}$$

where $u = \dfrac{1-0.5a}{1+0.5a}$, $v = \dfrac{b - a \cdot x^{(0)}(1)}{1+0.5a}$.

Then test the results. Define $\varepsilon(k)$ as residual, that is:

$$\varepsilon(k) = \frac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)} \times 100\% \tag{11}$$

The common condition is $\varepsilon(k) \le 20\%$ and the best condition is $\varepsilon(k) \le 10\%$.

3.2. SVM Model

Calculation speed decreases with increasing complexity of quadratic programming problems related with SVM solution. Least square SVM (LSSVM) is an improved version of SVM and reduces the complexity of solution. Thus this research employs LSSVM as prediction model. For the time series of network public opinion, regression function of LSSVM is:

$$f(x) = w^T \varphi(x) + b \tag{12}$$

where ω is weight vector and b is offset quantity.

The equation (10) can be transformed to a quadratic optimization problem by introducing structure risk function:

$$\min \|w\|^2 + \frac{1}{2}\gamma \sum_{i=1}^{n} \xi_i^2 \tag{13}$$

The constraint condition is:

$$y_i = w^T \varphi(x) + b + \xi_i \tag{14}$$

where γ is regularization parameter and ξ_i is slack variable.

Introduce lagrangian multiplier to get:

$$L(w,b,\zeta,\alpha)=\frac{1}{2}w^T w+\frac{1}{2}\gamma\sum_{i=1}^{n}\zeta_i^2+\sum_{i=1}^{n}\alpha_i(w^T\varphi(x_i)-b+\xi_i-y_i) \tag{15}$$

where α_i is lagrangian multiplier.

according to KKT condition:

$$\frac{\partial L}{\partial w}=0,\frac{\partial L}{\partial b}=0,\frac{\partial L}{\partial \xi_i}=0,\frac{\partial L}{\partial \alpha_i}=0 \tag{16}$$

Therefore, the final solution is:

$$\begin{bmatrix}0 & 1^T\\ 1 & x^t x+\gamma^{-1}I\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{17}$$

Transform the equation (15) by introducing kernel function k(xi,xj) to get the LSSVM prediction model as follows:

$$f(x)=\sum_{i=1}^{n}\alpha_i k(x_i,x_j)+b \tag{18}$$

In this research, RBF function is used as the kernel function of LSSVM. The final LSSVM prediction model is:

$$f(x)=\sum_{i=1}^{N}\alpha_i\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right)+b \tag{19}$$

where σ2 is RBF kernel bandwidth.

From the modeling process of LSSVM above, we can see that parameter (γ, σ2) needs to be defined for LSSVM model based on radial kernel function. Generally, N-fold cross validation is used to choose parameters for LSSVM. But there are several groups of parameters to be validated and large calculation quantity slows down rate of convergence, and parameters obtained by cross validation may not be the optimal solution. For this reason, it is hard to use cross validation to get the optimal LSSVM model. Genetic algorithm is employed in this research to determine parameters for LSSVM model due to its powerful global searching ability.

3.3. Workflow of Prediction Model of Network Public Opinion

Grey model is not suitable for prediction of time varying and nonlinear data series though it is able to indicate development tendency of data. SVM model is fit for describing nonlinear data series with small samples. Thus the two models can be combined to establish a grey SVM model for prediction of network public opinion. Steps are as below:

Step 1: Collect network data by network crawler and save data after eliminating impure information;

Step 2: Divide subject domains, cluster texts by hierarchical subject trees and hierarchical clustering, calculate clustering purity;

Step 3: Abstract hot topics, aggregate data by using aggregating software to get time series of network public opinion;

Step 4: Use GM（1,1）model to predict network public opinion and get corresponding predictions;

Step 5: Calculate residual of GM（1,1）predicted values from actual values;

Step 6: Determine the number of time delay of GM（1,1）predicted residual (m), take m residuals as input vector of LSSVM an actual predicted residual as desired output. Construct LSSVM learning sample;

Step 7: Optimize parameter of LSSVM by using genetic algorithm, and establish corresponding prediction model;

Step 8: Get predicted value of residual by prediction value, modify predictions of GM（1,1）to get the final prediction results of network public opinion.

To sum up, the workflow of prediction model of network public opinion based on grey SVM is shown in Figure 1.

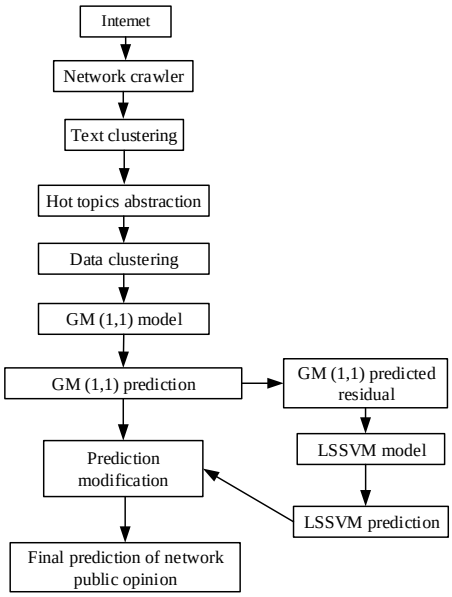


Figure 1. Workflow of Grey SVM Prediction Model of Network Public Opinion

4. Simulation Experiment

4.1. Data Source

Algorithm is implemented through VC++ programming in the Windows 2000 hardware environment of P4 3.0G CPU, 2G RAM in order to validate the performance of grey SVM model in predicting network public opinion. When predicting a hot network topic, 40 data are collected. Details are illustrated in Figure2.

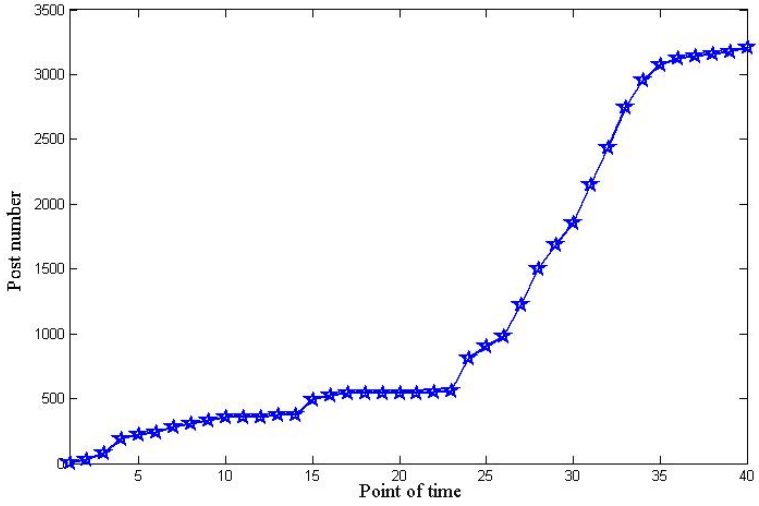


Figure 2. Time Series of Network Public Opinion

4.2. Data Preprocessing

The time series of network public opinion are preprocessed and normalized in [0,1] in order to accelerate training of the model and more effectively reflect changing tendency of network public opinion. That is:

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{20}$$

where $x_i^{'}$ denotes normalized data. $x_{\max}$ and $x_{\min}$ denote the maximum value and minimum value of time series of network public opinion respectively.

4.3. GM(1,1) Prediction

Data are divided into two parts. The first 30 data constitute a training sample set and the rest 10 data constitute a test sample set. Input training sample set into GM(1,1) to establish model and the model is used to obtain fitted values for training samples. Fitted values and actural values are show in Figure3.

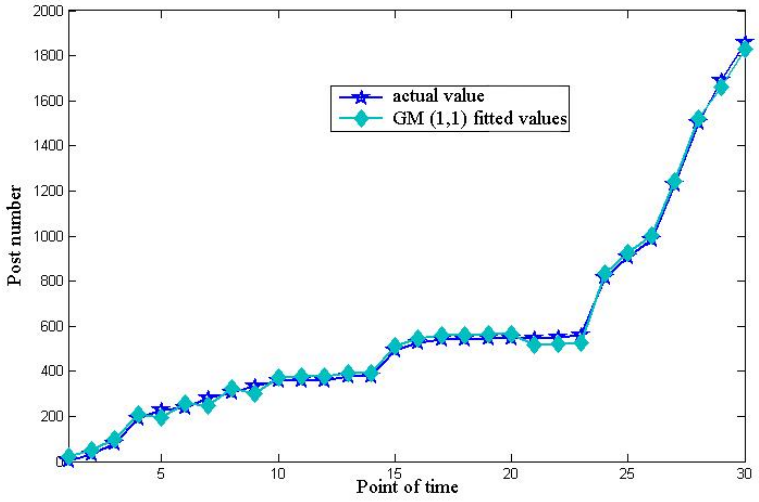


Figure 3. GM(1,1) Fitted Curve for Training Samples

As shown in Figure2, GM(1,1) model achieve an ordinary fitted result for the tendency of network public opinion, and the fitting precision reaches 90.10% and exceeds 85%. Thus GM(1,1) model is able to describe the basic changing tendency of network public opinion, but it cannot accurately reflect the time–variability and non-stationary of network public opinion. Fitting precision needs to be improved.

4.4. Modification of Residual by SVM

Calculate the residual series of predicted values from actual values and input into LSSVM for learning. Fit the residual series and modify GM(1,1) predictions. The fitted results are shown in Figure4.

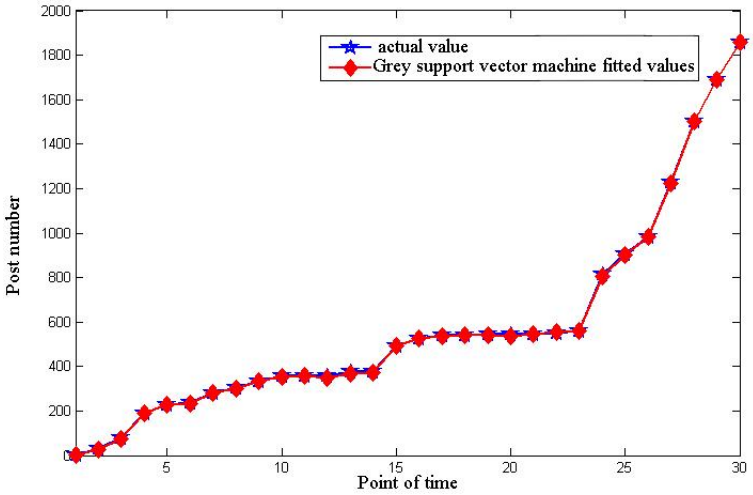


Figure 4. Grey SVM Fitted Curve for Training Samples

As shown in Figure4, the fitting precision of grey SVM model reaches as high as 97.81%, 7.08% higher than the 90.10% of GM(1,1) model. This indicates that combination of grey model and SVM model is an effective way to complement each other's advantages and overcome each other's shortcomings. It helps to increase the fitting precision of network public opinion. The fitted results show that grey SVM is an effective and feasible model for prediction of network public opinion and a preliminary verification for the rationality and feasibility of combing grey model and SVM model to predict network public opinion.

4.5. Performance Comparison with Other Models

It is generalization ability rather than fitting ability that matters in evaluating the performance of prediction models. The single GM(1,1) model and SVM model are specified to carry on experimental contrast in order to reflect the advantages of grey SVM model. The predictions of same test samples by several models are shown in Figure5.

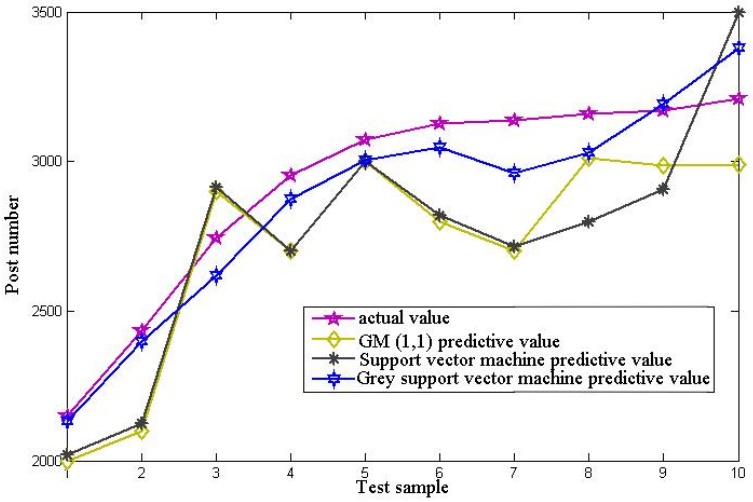


Figure 5. Predictions of Test Samples by Several Models

According to analysis on predictions of test samples by several models as shown in Figure5, predicted values got by grey model and SVM model differ greatly with actual values and both the two models bring about big errors. That is to say, single model can only describe the fragmental and partial changing rules of complex network public

opinions, but cannot accurately reflect their time-varying and nonlinear features. Grey SVM model has the advantages of both grey model and SVM model, so it overcomes shortcomings of a single model. The model is able to reflect the poor information and small samples in network public opinions, and accurately predicts the nonlinear and uncertain changing rules. It is more effectively to reflect the tendency of network public opinions and improves the prediction accuracy.

5. Conclusions

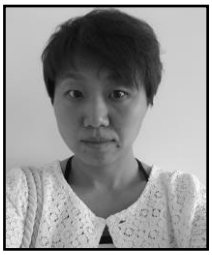
Network public opinion is a complex, time-varying, random and volatile system. Accurate prediction of its development tendency helps network supervisory authorities to discover hidden risks and direct network public opinion towards healthy development. This paper combines the advantages of grey model and SVM model to establish a grey SVM model for prediction of network public opinion. Simulation results indicate that grey SVM model is able to effectively and accurately predict the tendency of network public opinion and it has a promising application prospect in management of network public opinion.

References

[1] J. Hu and Z. Gao, "Distinction immune genes of hepatitis-induced heptatocellular carcinoma", Bioinformatics, vol. 28, no. 24, **(2012)**, pp. 3191-3194.
[2] J. Yan, B. Chen and J. Zhou, "A Low-Power and Portable Biomedical Device for Respiratory Monitoring with a Stable Power Source", Sensors, vol. 15, no. 8, **(2015)**, pp. 19618-19632.
[3] X. Li, Z. Lv and J. Hu, "Traffic management and forecasting system based on 3d gis", Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on, **(2015)**, pp. 991-998.
[4] S. Zhang and H. Jing, "Fast log-Gabor-based nonlocal means image denoising methods", Image Processing (ICIP), 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 2724-2728.
[5] J. Hu and Z. Gao, "Distinction immune genes of hepatitis-induced heptatocellular carcinoma", Bioinformatics, vol. 28, no. 24, **(2012)**, pp. 3191-3194.
[6] X. Song and Y. Geng, "Distributed community detection optimization algorithm for complex networks", Journal of Networks, vol. 9, no. 10, **(2014)**, pp. 2758-2765.
[7] J. Hu and Z. Gao, "Modules identification in gene positive networks of hepatocellular carcinoma using Pearson agglomerative method and Pearson cohesion coupling modularity", Journal of Applied Mathematics, vol. 2012, **(2012)**.
[8] D. Jiang, Z. Xu and Z. Chen, "Joint time-frequency sparse estimation of large-scale network traffic", Computer Networks, vol. 55, no. 15, pp. 3533-3547.
[9] J. Hu, Z. Gao and W. Pan, "Multiangle Social Network Recommendation Algorithms and Similarity Network Evaluation", Journal of Applied Mathematics, vol. 2013, **(2013)**.
[10] M. Zhou, G. Bao, Y. Geng, B. Alkandari and X. Li, "Polyp detection and radius measurement in small intestine using video capsule endoscopy", 2014 7th International Conference on Biomedical Engineering and Informatics (BMEI), **(2014)**.
[11] G. Yan, Y. Lv, Q. Wang and Y. Geng, "Routing algorithm based on delay rate in wireless cognitive radio network", Journal of Networks, vol. 9, no. 4, **(2014)**, pp. 948-955.
[12] Y. Lin, J. Yang and Z. Lv, "A Self-Assessment Stereo Capture Model Applicable to the Internet of Things", Sensors, vol. 15, no. 8, **(2015)**, pp. 20925-20944.
[13] K. Wang, X. Zhou and T. Li, "Optimizing load balancing and data-locality with data-aware scheduling", Big Data (Big Data), 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 119-128.
[14] L. Zhang, B. He and J. Sun, "Double Image Multi-Encryption Algorithm Based on Fractional Chaotic Time Series", Journal of Computational and Theoretical Nanoscience, vol. 12, **(2015)**, pp. 1-7.
[15] T. Su, Z. Lv and S. Gao, "3d seabed: 3d modeling and visualization platform for the seabed", Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on. IEEE, **(2014)**, pp. 1-6.
[16] Y. Geng, J. Chen, R. Fu, G. Bao and K. Pahlavan, "Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine", IEEE transactions on mobile computing, vol. 1, no. 1, **(2015)**, pp. 1-15.
[17] J. He, Y. Geng, F. Liu and C. Xu, "CC-KF: Enhanced TOA Performance in Multipath and NLOS Indoor Extreme Environment", IEEE Sensor Journal, vol. 14, no. 11, **(2014)**, pp. 3766-3774.
[18] N. Lu, C. Lu, Z. Yang and Y. Geng, "Modeling Framework for Mining Lifecycle Management", Journal of Networks, vol. 9, no. 3, **(2014)**, pp. 719-725.

[19] Y. Geng and K. Pahlavan, "On the accuracy of rf and image processing based hybrid localization for wireless capsule endoscopy", IEEE Wireless Communications and Networking Conference (WCNC), **(2015)**.
[20] Z. Lv, A. Halawan and S. Feng, "Multimodal hand and foot gesture interaction for handheld device", ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 1, **(2014)**, pp. 10.
[21] G. Liu, Y. Geng and K. Pahlavan, "Effects of calibration RFID tags on performance of inertial navigation in indoor environment, 2015 International Conference on Computing", Networking and Communications (ICNC), **(2015)**.

Authors



Bo Li, received her M.S. degree in software engineering from Jilin University in Changchun, China. She is currently a lecturer in the Changchun Institute of Technology. She is also a doctorate in computer college of Changchun University of Science and Technology. Her research interest is mainly in the area of Computer Software, swarm intelligence algorithms. She has published several research papers in the above research areas.

Baoxing Bai, received his doctor degree in computer from Changchun University of Science and Technology, China. He is a professor in Changchun University of Science and Technology. His research interest is mainly in the area of Computer Software, the image processing. He has published several research papers in the above research areas.

