International Journal of Hybrid Information Technology
Vol.9, No.2 (2016), pp. 339-348
http://dx.doi.org/10.14257/ijhit.2016.9.2.30

Research and Application of Network Router Link Prediction Method

CHEN Yu*[1,2] and DUAN Zhe-Min[1]

[1]Northwestern Polytechnical University, Xi'an 710072, PR China
[2]Zhengzhou University of Aeronautics, Zhengzhou 450015, PR China
E-mail: chenyu3440@gmail.com

Abstract

Multiple Autonomous systems (ASes) of the network are usually consisted of many routers such as inter-router and intra-router connection each other. In order to detect the network routers connection relationship, many researchers at home or abroad use various methods such as active end-to-end links detection or observing SNMP MIBs to understand the links' connection relationship among the routers in network' ASes. In this paper, by using the link prediction algorithm based on the similarity, we tried to predict the links' connection relationship and got the possible network topology structure among the network' ASes when the actual connection link don't be completely detected. Through the experiments, the prediction results can be seen that the similarity algorithm index of ACT, SRW, and SimRank based on the random walk can also achieve better prediction accuracy above 0.95, which prove the similarity index especially the random walk algorithm can realize the higher link prediction accuracy to the network' ASes under the conditions of missing some known connection link, etc.

***Keywords**: Autonomous system, network topology structure, link prediction, similarity index*

1. Introduction

The link prediction to a complex network is to point to predict the unknown links that existed or the future can produce. The mainstream research point to the link prediction problem is the method based on the topological structure similarity, and the development will be more and more deepening with the corresponding research carried on. Because this method according to the existing network structure with more available and reliable, the more important this method is universal, which avoid the messy process of training to get parameters. Now, the link connection relationships are usually depended on the active end-to-end measurement, tomography, or other management information. But, because the active end-to-end measurement needs the whole routing up-down information among the ASes, though the optimization of the path and beacon position can deduce the cost as much as possible, the inference result has deviation, and the algorithm design is hard based on the heuristic method or probability theory, etc, especially, these method normally used to predict some possible failed path and identify the set of failed link [1]. Some researcher adopted the Boolean solution to realize the congested IP link location [2]. Other researcher used the share risk link group (SRLG) analysis to obtain the link connection state among the ASes of OSPF area, but because of the life-time, the information of SRLG usually inaccuracy, so, through adding the probability inference to realize the failure accuracy position [3, 4]. There are many more methods to predict the network topology or failed link, such as the data mining to the syslog or console log, *etc.*, [5, 6] through the time associating to get the router link connection relationship. But, because of the time misalign, the association rule is failure.

ISSN: 1738-9968 IJHIT

Due to the router among the ASes of the Internet usually has some character according to the each router connection style, the routing protocol has its structure feature, and the connection character has some path similarity. We can adopt some advanced sociological relations features algorithm to try to predict the ASes link connection problem among the Internet. In this paper, the current popular link prediction algorithms based on path and random walk are introduced, and through using some index of these algorithm, we estimate some common simulation Internet ASes' link connection state.

2. Network Model and Prediction Basis

2.1. Network Model

Figure 1 is a connection relationship diagram among the Internet. Among the Figure 1, Figure 1 (a) is a router connection relationship model diagram among ASes of the Internet, and Figure 1(b) is a small local network connection model diagram consisted of router, switches, and some end computer, *etc*. We can see, the different network diagram respectively has its connection character, so, in this paper, we try to predict the connection relationship. Because the paper page limited, in this paper, we only study the Figure 1 as our research object. In the Figure 1, this is a router connection diagram of Internet composed of some ASes. Through the prior research of some kind of link prediction method, the network nodes and connection relationships among them are an un-directed connected graph. Each vertex generally refers to the network nodes such as router, switch, or the host, etc. And the connecting link between each vertex is the connection relationship. The ultimate goal of link prediction algorithm is as accurately as possible to find out the vertexes and edges of the un-directed connected graph. As shown in Figure 1, we set the network routers, switches, etc as digit number from 1 to 11 respectively for identification, and show the relationship between the connections between each node through the solid lines. Among them, the numbers' rank order is defined by the level of routers from the border router to the inner router. In addition, in order to detect known link prediction effect under the condition of unknown connection state, with a dotted line said known link of unknown connection. And in this paper, we act the un-detection links as the un-connection link. In addition, we find that the number identifier setting order can be many kinds of strategy. We should define a list of setting rules. Through the experiment, we find the different number identifier setting can get different prediction result, sometimes, if the number identifier setting unsuitable, the prediction results can probably be error, this is the research key to realize Internet link prediction accurate in our further research work to improve the prediction results. In our research, we define the ID as the router level, and the inter-route as the high level, and intra-router as the low lever, as well as the AS level. The ID number of nodes listed according to the ascending order (1, 2, … , n).

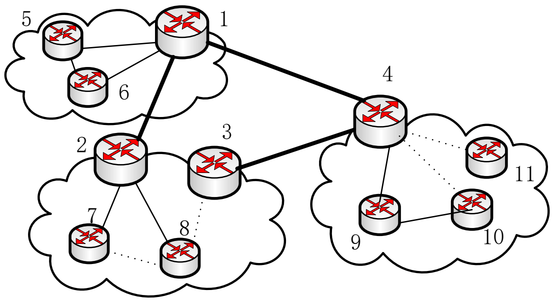

Figure 1. Connection Relationship Diagrams Among The Internet Nodes

2.2. Basic Method of Link Prediction

To a small scale of AS network connection model, we build a model of a powerless and un-direction connected network $G(V, E)$.Among them, V represents a collection of nodes, such as the routers, switches, etc. E is a set of links. Assumes that in a network the router node number is N, and from the Figure 1, we can get N=11, the total link number is M, in the network model of Figure 1, we can get $M=N(N-1)/2=55$ numbers of nodes. But now, only 10 numbers of link connection can be detected. And in fact there are 14 numbers of link connections in the AS with 4 un-detected links. Consider a kind of link prediction algorithm, we give a score value S_{xy} to each pair of nodes (x, y), because the AS network G is un-directed, therefore, the scores are equivalent, namely $S_{xy}=S_{yx}$. Then, according to the size of the score value, we make descending order to pair of node (x, y) from big to small, and assumes that probability to the pair of node existing link connection is greater when the score value at the top.

In order to measure and evaluate the effect of algorithm in link predicting, we need to divide the end-to-end detects link information (namely known link E) random into two parts of the training set E^T and the testing set E^P. The link connection situation in the training set act as the known information to use, when calculating the score value between the two routers, we can only use the information of this set. But we did not act the link condition of testing set as unknown information; the testing set is used to evaluate the result of the prediction. Intuitively, there are relations between the set: $E = E^T \cup E^P$, $E^T \cap E^P = \Phi$. We can see that the link belongs to the complete set U but not belongs to E is inexistence link.

Measuring the accuracy index of the link prediction method mainly has three kinds: respectively includes the area under the receiver operating characteristic curve (AUC) index, precision index (PI) and ranking score index (RSI). [7] When measuring the accuracy, the emphasis focus is slightly different, AUC mainly measures the accuracy of the algorithm from the overall, PI is a part index, only consider the top numbers L of predict edge whether accurate or not, RSI pays more attention is to sort the predict edge. In this paper, we use the AUC score value to valid the accuracy of predicting link. AUC refers to the possibility of the similarity under randomly selecting the link higher than randomly selecting not exist link edge. According to this thought, we randomly selects a edge from a test set very time, and let it compared with randomly selected non-existent edge, if the similarity of the test set edge is greater than the edge of the non-existent edge, add score of 1, if equal, added score of 0.5, otherwise, add no score. So, hypothesis comparing n times, if there is $n1$ times of the similarity of the edge in the test set more than the similarity of the non-existent edge, and n2 time is equal to two similarity, the AUC can be defined as:

$$accurcy = (n1 + 0.5n2) / n \tag{1}$$

We can see, the value of AUC gets more than 0.5, which expresses how much degree the prediction value is superior to the random selection. In this paper, we adopt AUC index to evaluate the prediction performance of the router's connection relationship of the Internet.

3. Link Prediction Method

Considering the link structure relationship among the routers in the Internet, the link prediction method based on structure similarity is more suitable for the forecast of IP network topology. When predicting link connection relationship based on the similarity of the router nodes, there is a premise hypothesis, if the greater the similarity between two routers nodes, the probability of connection relationship between the two nodes is bigger. From the point of the network connection, connected to the switch of the host more, therefore, switch the corresponding degree of value is bigger. In this paper, we only monitor whether the structure similarity algorithm can realize link prediction or not, and try to find the most suitable to the IP network link prediction method, and provide a research basic to realize IP network link topology finding. There are mainly two kinds of similarity index such as based on the method of path similarity and random walk character.

3.1. Similarity Index based on Path

In the previous study, we found that the similarity index based on common neighbor has good performance in higher aggregation coefficient network, but for low cluster coefficient network field, the prediction accuracy is poor. The similarity index based on the path defines the similarity from the aspect of path information, and the definition of accuracy is mainly based on whether the structure characteristics of the target network can be grasped. The method based on path has four:

3.1.1. Full Path Index Katz Algorithm [8] Katz algorithm consider the entire path set, this is different with the common neighbor algorithm that only consider node neighbors in first level. Its definition is: the more paths between two nodes, the shorter path express the stronger relationship, namely that the shorter path has heavy weight. The mathematical expression formula of this class is listed as formula (2):

$$S_{xy} = \sum_{l=1}^{\infty} \beta^1 \mid path_{xy}^{<l>} \mid = \beta A_{xy} + \beta^2 (A^2)_{xy} + \beta^3 (A^3)_{xy} + ... \qquad (2)$$

Among the formula (2), $path_{xy}^{<l>}$ expresses the entire path set with the length of 1 between the node x and y. β expressed the attenuation factor of path weight, which to control the weight of path. With the β smaller, S_{xy} is more close to the neighbors; because the path is longer the contribution is smaller.

3.1.2. Local Path (LP) Algorithm [9**]** LP algorithm has too much calculation amount, which limits its application in the complex networks. And LP algorithm considers the influence of neighbor nodes' similarities, and the amount of calculation is small. LP algorithm on the one hand takes into account first level neighbor and on the other hand considers the secondary level [4].Its computation formula is listed as formula (3):

$$S_{xy} = (A)_{xy}^2 + \varepsilon (A)_{xy}^3 \qquad (3)$$

Among the formula (2), ε is adjustment parameter, and can be given different value according to network style. When the ε is 0, the index can be change into common neighbor index. A is the neighbor matrix of the network, $(A)_{xy}^2$ expresses the path length of 2 between the node *x* and *y*, and its value is the number of common neighbor, $(A)_{xy}^3$ represents the path length of 3. If *x* connects with *y* directly, its path number is 1.

3.1.3. Leicht-Holme-Newman Index (LHNI) [10**]** Index of LHNI is a kind of deformation of index of Katz algorithm. And its similarity is defined as formula (4).

$$S_{xy} = \varphi AS + \Psi I = \Psi (I - \varphi A)^{-1} = \Psi (I + \varphi A + \varphi^2 A^2 + ...) \qquad (4)$$

Among the formula (4), *Φ* and Ψ is the adjustment parameter with value less than 1, the smaller Φ gives the shorter path bigger weight, which can control the balance between the two similar nodes, and *D* is diagonal matrix. To nodes of *x* and *y*, if the neighbor of node x and y has same similarity, x and y has high similarity. If we set Ψ =1, LHN-II similar to the index of Katz algorithm.

3.2. Similarity Index based on Random Walk

The similarity algorithm based on walk definition include the indexes of average commuting time (ACT), SimRank, Cosine (Cos^+), local random walk (LRW), random walk with restart (RWR) and superposed local random walk (SLRW), *etc*. The basic hypothesis of random walk is: carry out the random walk from the node *x*, the more less the average step is, the similarity between node *x* and *y* is, if reach the required the less the average number of nodes in the *y*, the similarity of the node *x* and *y* is higher.

3.2.1. ACT [11] If suggesting $m(x,y)$ is average step a random particle walk from node *x* to node *y*, then, the average commuting time between the node *x* and node *y* can be expressed as formula (5).

$$act(x,y) = m(x,y) + m(y,x) \tag{5}$$

Its data calculation can be obtained through calculating the pseudo inverse l^+ of the Laplace matrix of network. The formula is shown as bellow (6).

$$act(x,y) = M(l^+_{xx} + l^+_{yy} - 2l^+_{xy}) \tag{6}$$

The smaller of the average commuting time is, the similarity is higher between the two nodes. The Internet networks usually have the common aggregation effect, therefore, the closer to the nodes, the links are easier to produce, and the similarity definition ignoring the constant *M* based on ACT is listed as follows.

$$S_{xy} = \frac{1}{l^+_{xx} + l^+_{yy} - 2l^+_{xy}} \tag{7}$$

3.2.2. SimRank [12] The definition of SimRank algorithm is shown as bellow: in some degree, if two node has the similar neighbor node(s), then, the two nodes are similar. In addition to the special case, *similarity* (*x*, *x*) = 1, otherwise, the computation formula is shown as formula (8).

$$S_{xy} = \gamma \frac{\sum_{a\in\Gamma(x)} \sum_{b\in\Gamma(y)} similarity(a,b)}{|\Gamma(x)| \bullet |\Gamma(y)|} \tag{8}$$

Among the formula (7), $\gamma \in [0,1]$ is attenuation function in transferring the similarity.

3.2.3. Cos^+ [13] The index of Cosine is based on the vector inner product. In European space, $v_x = \Lambda^{\frac{1}{2}} \cup^T e_x$,element l^+_{xy} in the elements' set *l*+ express the inner product of vector v^T_x and v_x , *U* is a standard orthogonal matrix, and the matrices are ordered according to the size of the corresponding feature vector. *Λ* is diagonal matrix composed of characteristic vector, and $\vec{e}_x$ is a vector of *N* line and 1 volume, its first x^{th} element is 1, the rest of the elements are 0, *T* expresses the matrix transpose.

$$S_{xy} = \cos(x,y)^+ = \frac{v^T_x v_y}{|v_x||v_y|} = \frac{l^+_{xy}}{\sqrt{l^+_{xx} \bullet l^+_{yy}}} \tag{9}$$

3.2.4. RWR [14] The index of RWR is an application for Page Rank. The main method is to assume when the particles carry out every step in the random walk will return to the starting position at a certain probability. Assumes that at starting node *x*, we randomly select a neighbor node as the next step with the probability of *c*, and the probability of returning to the node *x* for 1-*c*. q_x expresses the probability vector of *x* arriving to each node y in network. So, $q_x = cP^T q_x + (1-c)e_x$. P is the transition probability matrix, if x connects y, P_{xy}=1/K_x, and vice of 0.This formula can also be written as $q_x = (1-c)(I - cP^T)^{-1} q_x$ directly. e_x is the initial state of an one-dimensional vector with the first x^{th} element value of 1, the other elements value of 0. The expression formula is as follows.

$$S_{xy} = q_{xy} + q_{yx} \tag{10}$$

3.2.5. LRW [15] The index formula is shown as formula (11).

$$S_{xy} = q_x \pi_{xy} + q_y \pi_{yx} \tag{11}$$

q is an initial distribution, and defined as $q_x = K_x / M$. The similarity measurement process of node x and y is of the initial state began random walk from the node. Therefore, the initial vector is for πx (0)=ex, when $t > 0$, vector can be express as: $\pi_x(t+1)=P\text{T}\pi_x(t+1)$.

3.2.6. SRW As the name implies, the index of SRW is a superposition to the front few steps of the index of LRW. This index's focus is to give a greater similarity value to the target node and its neighbors. The formula of SRW is shown as formula (12).

$$S_{xy} = \sum_{l=1}^{t} S_{xy}(l) = \sum_{l=1}^{t} q_x \pi(l_{xy}) + \sum_{l=1}^{t} q_y \pi(l_{yx}) \tag{12}$$

4. Experiment and Analysis

This chapter carries out the prediction to the AS network link connection diagram of Figure 1. We can see from the Figure 1, there are 11 nodes that can produce 55 numbers of links, but in fact, the numbers of links made up of node pair only have numbers of 14, respectively is (1,5), (1,6), (5,6), (1,2), (1,4), (2,7), (2,8), (3,4), (4,9), (4,10), (3,8), (7,8), (4,11) and (9,10). But through the link tomography, the link pair of (3,8), (7,8), (4,10), and (4,11) are un-detection, so, there are 41 numbers of links are un-connection link and 4 numbers of links are un-detection, total 45 numbers of links are un-known. That is to say, the known connection links' pair respectively is (1,5), (1,6), (5,6), (1,2), (1,4), (2,7), (2,8), (3,4), (4,9), and (9,10). In the process of algorithm running, for example, we divide the known links into two sets, respectively is training set and test set, and the training set including (1,5), (1,6), (5,6), (1,2), (1,4), (2,7), (2,8), (4,9) and (9,10), test set including link pair of (3,4). In order to verify the link prediction algorithm, we use the training set of 9 numbers of links to predict other 46 numbers of links (including other inexistent 41 numbers of link, and 4 numbers of un-detection links, as well as one test links of (3,4)). We use the above two kinds of different link prediction algorithms to predict these links' connection conditions. In order to obtain the performance evaluation and assessment of the link prediction algorithm, we use the AUC to verify the accuracy of the algorithm and compare the score value between the test link set and inexistent or un-detection link set, because this experiment we select one test link and 45 numbers of inexistent or un-detection link, we take 1*45 times' comparison. And based on the AUC computation method in formula (1), we get the AUC value of each algorithm. In the process of algorithm design, we can monitor multiple times and take the average to the data.

4.1. Result Analysis to Similarity Index based on Path

The AUC value of LP, Katz, and LHNI algorithm based on the path similarity index respectively is shown as Table 1. And the results are listed in increasing order.

Table 1. AUC Value based on Path Similarity Algorithm

Index	Katz_0.01	LP	LHNI_0.9
AUC	0.7813	0.8483	0.8839

From the Table 1 we can see, the prediction accuracy of AUC is poorer, in the aggregation coefficient higher network, the performance of similarity algorithm based on path is very good, but in lower concentration coefficient IP network, the prediction accuracy is bad. When choosing the parameter of 0.01 in index of Katz, we can achieve the maximum prediction accuracy. Selecting the parameter of 0.9 we can achieve he maximum AUC in index of LHNII. However, from the result of LPI, Katz, and LNHII, the AUC value is obvious low, and greatly influenced by parameters' selection. Because the path similarity index is defined through the angle of information of the path, the accuracy is mainly based on whether the target network structure can be grasped or not, so, to the ASes, the path similarity index is not effectively in IP network link prediction.

4.2. Result Analysis to Similarity Index based on Random Walk

The AUC value of ACT, SimRank, Cos^+, RWR, LRW, and SRW algorithm based on the random walk similarity index is shown as Table 2. And the results are listed as increasing order.

Table 2. AUC Value based on Random Walk Similarity Algorithm

Index	RWR0.85	LRW_4	Cos^+	SimRank	SRW_3	ACT
AUC	0.8712	0.8795	0.8951	0.9374	0.9576	0.9673

From Table 2 we can see, the index of ACT, SRW, SimRank can realize better prediction accuracy, and other index all can beyond 0.85, which can obtain better AUC value. The parameter selection of SRW has great influence to the value of AUC, when electing three steps, the AUC of SRW can achieve maximum value of 0.9576, less or more steps will cause the value of AUC decreased. SWR has better prediction accuracy; the change of the parameters always gains the high prediction value of AUC with slight impact, which can always get higher accuracy because of the SWR index giving greater similarity value to the target node and its neighbor node. Because ACT index carry out similarity judgment according to the average commuting time, namely, the close nodes are more likely to produce link, which has the same character with the IP network link connection structure, so, the AUC value is less than 0.96.

5. Conclusion and Future Work

Through the analysis and research to the index of link prediction algorithm among the ASes in the internet, we can get the results from the experiment. The similarity indexes based on random walk algorithm can achieve good results in the link prediction. Due to the similarity algorithm based on path is suitable for the higher network aggregation coefficient, therefore, for the low concentration coefficient of IP network, the prediction accuracy is poor. This paper only carry out the simulation and experiment to a certain AS IP network model, and from the prediction accuracy and stability, the similarity index of SWR based on random walk has highest prediction accuracy in the data missing conditions, and with reasonable selection of the parameters, the index of SRW can gain prediction accuracy more than 0.95. Therefore, the index based on random walk algorithm is suitable to the IP network link prediction than path similarity index. But, when detecting the link connection relationship, we found the number ID definition sometimes has certain affection to the prediction result. How to mark the ID number is also the key of research. In addition, grasp the topology normal routing rules is also important to realize the link failure inference. And when the fact connection link is un-detection, in this paper we have not analyzed, and act it as un-connection link, so, how to accurate predict the missing link is also an important key to improve the link prediction accuracy.

In this paper, we mainly used the two kinds of link prediction method to some routes' network connection relationship, although the ASes network has some normal structure character, but, in the Internet, there are many kinds of ASes, so, these method can be applied into the all kinds of link prediction style or not is the further work. And based on the different Internet connection style, the improved algorithm should be presented to complete the accurate link prediction.

References

[1] Y. Huang, N. Feamster and R. Teixeira, "Practical Issues with Using Network Tomography for Fault Diagnosis", ACM SIGCOMM Computer Communication Review, vol. 38, no. 5, **(2008)**, October, pp. 53-57.
[2] H. X. Nguyen and P. Thiran, "The Boolean Solution to the Congested IP Link Location Problem: Theory and Practice", INFOCOM, 26th IEEE International Conference on Computer Communications, **(2007)** 6-12 May, pp. 2117-2125.
[3] D. Katabi, J.-P. Vasseur and S. Kandula, "Shrink: A Tool for Failure Diagnosis in IP Networks", SIGCOMM'05 Workshops, Philadelphia, PA, USA, **(2005)**, 22-26 August; pp. 173-178.
[4] R. R. Kompella, J. Yates, A. Greenberg, A. C. Snoeren, "IP Fault Localization Via Risk Modeling", NSDI'05 2nd Symposium on Networked Systems Design and Implementation, Boston, Massachusetts, USA, **(2005)** May 2-4. pp. 57-70.
[5] K. Yamanishi and Y. Maruyama, "Dynamic Syslog Mining for Network Failure Monitoring", KDD'05 , Chicago, USA, **(2005)**, 21-24 August, pp. 499-508.
[6] W. Xu, L. Huang, A. Fox, D. Patterson and M. Jordan, "Detecting Large-Scale System Problems by Mining Console Logs", the 22nd ACM Symposium on Operating Systems Principles (SOSP' 09), Big Sky, MT, USA, **(2009)** October, pp. 37-46.
[7] L. Linyuan, "Link Prediction on Complex Networks", Journal of University of Electronic Science and Technology of China, vol. 39, no. 5, **(2010)**, pp. 651-660.
[8] L. Katz, "A New Status Index Derived from Sociometric Analysis", Psychometrika, vol. 18, no. 1, **(1953)**, pp. 39-43.
[9] L. Linyuan, C. Jin and T. Zhou, "Similarity Index Based on Local Paths for Link Prediction of Complex Networks", Physical Review E, vol. 80, no. 4 (Pt 2), **(2009)**, pp. 593-598.
[10] E. A. Leicht, P. Holme and M. E. J. Newman, "Vertex Similarity in Networks", Physical Review E, vol.73, no. 2, **(2006)**, pp. 026120-1-10.
[11] D. J. Klein and M. Randic, "Resistance distance", Journal of Mathematical Chemistry, vol. 12, no. 1, **(1993)**, pp. 81-95.
[12] G. Jen and J. Widom, "SimRank: A measure of structural-context similarity", Proceedings of the ACM SIGKDD, Edmonton, Alberta, Canada, **(2002)**, pp. 538-543.
[13] F. Fouss, A. Pirotte, J.-M. Render and M. Saerens, "Random-walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation", Knowledge & Data Engineering, IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3, **(2007)**, pp. 355-369.
[14] S. Brin and L. Page, "The Anatomy of a Large-scale Hyper Textual Web Search Engine", Computer Networks and ISDN Systems, vol. 30, **(1998)**, pp. 107-117.
[15] W. Liu and L. Linyuan, "Link Prediction based on Local Random Walk", EPL: A Letters Journal Exploring, the Frontiers of Physics, vol. 89, no. 5, **(2010)**, pp. 58007-58012(6).

Authors

CHEN Yu, He is currently pursuing Ph.D. degree from Northwestern Polytechnical University, School of Electronics and Information. Since 2001, he has been working as a teacher in Zhengzhou University of Aeronautics, assistant professor. His research interests are circuit and system, data collection and signal process, network information and network security, *etc*.

DUAN Zhe-Min, He is a professor in Northwestern Polytechnical University, School of Electronics and Information. In 2011, he was awarded a prize of national teaching masters. His electronic series basic course teaching team was named the national teaching team in 2010. His research interests are circuit and system, data collection and signal process, integrated circuit analysis and design, electrical theory and new technology, *etc*.