International Journal of Grid Distribution Computing
Vol. 8, No.6, (2015), pp.141-150
http://dx.doi.org/10.14257/ijgdc.2015.8.6.14

IP Network Topology Link Prediction Based on Improved Local Information Similarity Algorithm

Chen Yu*[1, 2] and Duan Zhemin[1]

[1]*Northwestern Polytechnical University,*
School of Electronics and Information, Shanxi Xi'an, 710072
[2]*Zhengzhou Institute of Aeronautical Industry Management,*
Henan Zhengzhou, 450015
chenyu3440@gmail.com

Abstract

With the expansion of the IP network scale, the topology structure of IP network has also changing. Traditional network topology research methods by adapting tools of Ping or Traceroute were used to carry out the active end-to-end detection from the origin router nodes to destination router nodes. But sometimes, because some routers set the UDP/SNMP limitation, there are some router IP address could not be detected. In addition, AS operator can adopt router syslog configuration files to obtain link connection condition, but, these files have the update life-time, it is hard to obtain the information in real-time. So, we only obtained some incomplete topology information. Due to the IP network has some local information similarity, this article adopted the complex link prediction algorithm based on local information similarity algorithm and design a kind of improved method based on the IP network characters to realize the unknown links prediction. Through the experiment, the results expressed that, the prediction index of local information common neighbor algorithm index of AA, RA obtain the most good accuracy, and the improved Bayesian CN index combining the degree value index of AA can get more higher prediction accuracy about 0.97. This article confirmed the link prediction method based on local information similarity algorithm can be effectively applied in the IP network unknown link prediction.

Keywords: *IP network, topology structure, link prediction, similarity index, local information similarity*

1. Introduction

IP network is composed of multiple autonomous area (AS), each area usually use internal gateway protocol (IGP) of ICMP, SNMP, RIP, OSPF, *etc*. Among them, using SNMP protocol to build network topology is the more common way. Through the operating of the MIB database in the managed devices, operators can extract the available information to build topology. [1] However, if the router connection relation has changed, Mib library did not upgrade, the connection relation is not accurate. With the aid of Shared Risk Link Groups (SRLGs), operator can real-time obtain current IP connection relations, but analysis is laborious. [2] In addition, by examining the router configuration file can also understand the router connection with other routers, but update exits lifetime, the information is no completely accuracy. Based on ICMP protocol of Ping detection tools to actively detect router node in network, it need to send a lot of detection data packet, which is bound to increase network traffic, and bring unnecessary obstruction to local area network (LAN).

Chen Yu is the corresponding author.

ISSN: 2005-4262 IJGDC

On large IP network connection test, because of using the AS BGP routing protocols for security reasons, many routers set a limit to ban the Ping jurisdiction, operators unable to obtain network topology information. And because of the large amount of log data in free text format, current research scholars use router syslog to understand the connection relation of each router is more difficult to parse. [3] In this paper, with the aid of link prediction algorithm, we explore the IP network structure, under understanding part of the relations of the router connection, predict the IP network topology structure.

In this paper, chapter 2 introduces the IP network model building and link prediction basis, chapter 3 introduces link prediction algorithm and related indicators, chapter 4 presents the IP network topology prediction experiments by using the AUC for accurate diagnosis, and describes the link performance between the traditional algorithm and the improved prediction algorithm in the IP network link prediction, chapter 5 concludes the paper and put forward the further work.

2. IP Network Model Build and Prediction Basis

2.1. IP Network Model

Each network node on IP network and the connection relations among them consists an un-direction connected graph. Each vertex generally refers to the network nodes such as router, switch, host, sub-net. The edge connecting each vertex expresses the link between the equipment. The ultimate goal of topology discovery algorithm is as accurately as possible to find the vertexes and edges of the un-direction connected graph. Figure 1, respectively presents an IP network containing three AS [4]. We use the numbers to identify the intra-router or inter-domain router, *etc*, and express the connection relation by the different thickness solid line. The order of ID numbers is defined by the router level from the border router to the inner router from 1 to 11. And the H1 and H2 is end-to-end detection computer, we can adopt them to realize the link connection state detection when the links are good.

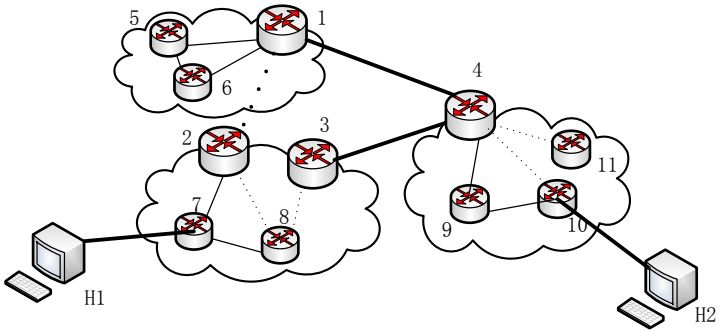

Figure 1. IP Network Model Diagram

2.2. Topology Prediction Basis

Based on the structural feature of IP network, we use link prediction method based on structure similarity, under obtained some router connection relation; we try to predict the network topology relation. To a large IP network, 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; E is a collection of links. Assumes that in a network the router node number is N, the total link number is M. In the network, it consists of N (N - 1) / 2 nodes. As shown in Figure 1, there are 11 router nodes, therefore, all of the nodes should have 55 numbers of connection links, but now only 14 numbers of actual link connection. And in fact, there may be some of link connections relationship doesn't

be detected when carrying out the active end-to-end detection. But in a local AS, the link connection has some similarity character, so, we consider a kind of link prediction algorithm to predict the unknown link connection.

In order to measure and evaluate the effect of the link prediction, namely in order to test the accuracy of algorithm, 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 condition of training link set act as known information to use, when calculating the score value between the two router, we can only use the information of this set, and the link condition to the test set don't act as the known information to predict, but in order to use them to evaluate the result of the prediction. Intuitively, there are the relations between the set as Formula (1).

$$E = E^T \cup E^P,\ E^T \cap E^P = \Phi \tag{1}$$

We can see that the link belongs to the complete set U, but not belongs to E is inexistent link in actual condition. 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). [5] When measuring the accuracy, the emphasis focus is slightly different, AUC mainly measure 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's more attention is to sort the predict edge. In this article, we use the AUC index to valid the accuracy of predicting link. The AUC can be defined as Formula (2) .

$$accurcy = (n1 + 0.5n2) / n \tag{2}$$

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. For example, in Figure 1, there are 11 numbers of nodes, 14 numbers of links can be detected, but the network total possibly has 55 numbers of link connection relationships, so, there are 41 numbers of links are inexistent connection. In order to test the algorithm precision, we need to select some numbers of links from the known links of (1,5), (5,6), (1,6), (1,4), (3,4) , (2,7), (7,8), (4,9), and (9,10) as test set, and the rest numbers of links as training test. If we select 6 training links, rest 3 numbers of link and 46 numbers of unknown links, it should carry out 46*6=276 times of comparisons. And when the score value of known link is bigger than inexistent link, the value of AUC add 1, if the score value of test link is equal to inexistent link, the value of AUC add 0.5, if the score value of inexistent link is bigger than test link, the value of AUC add zero.

3. Link Prediction based on Local Information

Considering the structure relation of IP network, link prediction method based on structure similarity is more suitable for the prediction of IP network topology. There is a premise of assumption by using the similarity of the router nodes to predict link, which is if the greater the similarity between two routers nodes, then the probability of connection relation between the two nodes is bigger. Here, the similarity mainly refers to the proximity not the tradition similarity. So, the core problem to solve is how to define the similarity between the two nodes. And from the point of the network connection, the inner router connecting to the edge router (or switch) is more; the corresponding degree value is bigger. Intuitively, when choosing the prediction index containing weight value changing with the degree value changing, the prediction effectively is better.

3.1. Similarity Index

The index of neighbor nodes is a similarity method based on the graph topology, if the intersection between the neighbor set Γ (*x*) and Γ (*y*) of the neighbor node x and y is greater, the node x and y is more similar. If for node *z*, there is the edge of < *x*, *z* > and < *z*, *y* >, so more prone to produce the link between node *x* and *y*. [6]

3.1.1. Common Neighbors Index: (CN) [7] CN index is the most simple similarity index, at the same time, it only needs to consider the local information of network, theoretical basis is: if the two nodes have the same number of neighbors, the more likely existing connection edge between the two nodes. As shown in Formula (3).

$$S_{xy} = |\Gamma(x) \cap \Gamma(y)| \tag{3}$$

In IP network, $\Gamma(x)$ are the other routers connected to the router *x*, $\Gamma(y)$ are the other routers connected to the router *y*, if a router *z* connects to the router *x* and *y*, then, router *x* may also connects to the router *y*.

3.1.2. Salton Index: [8] Salton Index formula uses common neighbor node number divided by the product of the node degrees. Degree of node refers to the number of nodes connected with other neighbors' nodes. If the node degree value is big, to some extent that the higher the importance to the nodes in the network. Calculation method is shown in the Formula (4).

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|K_x \times K_y|} \tag{4}$$

3.1.3. Sørensen Index: [9**]** Sørensen index is obtained by using common neighbor number divided by the sum of the node degree; common node number does not affect the index value.

$$S_{xy} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{|K_x + K_y|} \tag{5}$$

3.1.4. Leicht-Holme-Newman Index: (LHNI) **[**10**]** Leicht-Holme-Newman index focus on that giving the node a small weights if its adjacent nodes has large moderate product. The LHNI index Formula (6) is listed below.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{K_x \times K_y} \tag{6}$$

3.1.5. Hub Promoted Index: (HPI) **[**11**]** HPI is obtained by common neighbor number divided by the nodes degrees in small degree number. In the adjacent nodes, small degree will get greater weight, nodes with big degree will have more beneficial.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{K_x, K_y\}} \tag{7}$$

3.1.6. Hub Depressed Index: (HDI) HDI is opposite to the HPI, which is to use common neighbor number divided by the large node degrees, in the node's neighbor's nodes, node with big degrees can get smaller weights. Because the denominator take big node's degrees, nodes with small degree will not be dominant, therefore, the nodes with big degree are disadvantages.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{K_x, K_y\}} \tag{8}$$

3.1.7. Adamic-Adar Index: (AA) [12] AA index give all common neighbor nodes in small degree to a larger weight.

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log K_z} \tag{9}$$

3.1.8. Resource Allocation Index: (RA) **[**13**]** RA index is similar to AA index.

$$S_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{K_z} \tag{10}$$

3.1.9. Jaccard Index: [14] The number of common neighbor divided by the merge set of two nodes, which is different from the node degree. When the node degree is same, if existing multi repeated, the higher the index is. On the other hand, the index will reduce.

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{11}$$

3.2. Improvement Algorithm Design

In chapter 3.1 we introduced the famous link prediction index based on the local information similarity, in order to carry out the IP network link prediction, among them, common neighbor similarity algorithm, considering the object is an IP network, the router clustering coefficient is low, although some nodes has no common neighbors, they are connected by some path, the real similarity is not low, but they will be assigned to zero, CN value is significantly lower than the other index's AUC. Especially, when analyzing the large-scale IP network, as the network nodes is more, using common neighbors' characters to predict link by the CN index, there may be a considerable number of nodes with the similarity score value of 0, which cause the prediction accuracy is not high. Therefore, assuming the common neighbor nodes with small degree l often larger than the large degree is more important, on the basis of CN, we consider the influence of the small common neighbor nodes. So, we adopt the node of similarity scores and combine with the index of the AA or RA into the CN index together to optimize the similarity index, combining AA index with CN index is shown as follows formula (12). As well as we can induce the RA index into the CN index. And γ is an adjust parameter of (0, 1).

$$S_{xy} = \gamma |\Gamma(x) \cap \Gamma(y)| + \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z} \tag{12}$$

In addition, the Bayesian classifier in many areas have achieved better effect in prediction, here, we consider introduce the naive Bayesian classifier algorithm into the IP network link prediction to judge the IP network link prediction effects. Due to the simple Bayesian classifier is a kind of application based on the independence assumption of simple Bayesian probability classifier, in order to more accurately describe the potential probability model as an independent feature model, a characteristics set of $E(a_1,a_2,...,a_n)$ is

given. Among them, a_n is the attribute value, the probability of a certain category C under given characteristics of E can be represented as conditional probability $P(C|E)$. [15]

$$P(C \mid E) = P(C \mid a_1, a_2, \ldots, a_n) = \frac{P(C) \times P(a_1, a_2, \ldots, a_n \mid C)}{P(a_1, a_2, \ldots, a_n)} \tag{13}$$

According to naive Bayesian theory, above formula (13) can be expressed as formula (14).

$$P(C \mid E) = \frac{P(C)}{P(a_1, a_2, \ldots, a_n)} \times \prod_{i=1}^{n} P(a_i \mid C) \tag{14}$$

We define the relationship of e for a conditional probability under given a set of neighbor node properties based on the naive Bayesian model algorithm.

$$P(e \mid \Gamma(x, y)) = \frac{P(e)}{P(\Gamma(x, y))} \times P(\Gamma(x, y) \mid e) \tag{15}$$

Then, we can separately get the conditional probability for existing e or not existing $\bar{e}$ in relationship.

$$P(e \mid \Gamma(x, y)) = P(e \mid \omega_1, \omega_2, \ldots, \omega_n) = \frac{P(e)}{P(\Gamma(x, y))} \times \prod_{i=1}^{n} P(\omega_i \mid e) \tag{16}$$

$$P(\bar{e} \mid \Gamma(x, y)) = P(\bar{e} \mid \mid \omega_1, \omega_2, \ldots, \omega_n) = \frac{P(\bar{e} \mid)}{P(\Gamma(x, y))} \times \prod_{i=1}^{n} P(\omega_i \mid \bar{e} \mid) \tag{17}$$

Among them, ω_i is a given neighbor nodes of node pair (x, y), so, the similarity degree S_{xy} of (x, y) may be defined as the ratio of (13) and (14) .

$$S_{xy} = \frac{P(e)}{P(\bar{e})} \times \prod_{i=1}^{n} \frac{P(\bar{e})}{P(e)} \times \frac{P(e \mid \omega_i)}{P(\bar{e} \mid \omega_i)} \tag{18}$$

$$s = \frac{P(e)}{P(\bar{e} \mid)} \times \prod_{i=1}^{n} \frac{P(\bar{e} \mid)}{P(e)}$$

The left part of formula (18) is prior condition probability for a given IP networks, because $P(e)$ is a ratio of network links and possible link. The right part of $R_{\omega i} = \frac{P(e \mid \omega_i)}{P(\bar{e} \mid \omega_i)}$ is the contribution for every neighbor user ω_i, so, the CN index based on the naive Bayesian relation formula (19) is listed as bellow.

$$S_{xy} = \mid \Gamma(x, y) \mid \times \log s + \sum_{\omega \in \Gamma(x, y)} \log R_{\omega} \tag{19}$$

Similarly, we can use naive Bayesian relationship easy to improve the index of AA or RA, *etc*.

4. Analysis of Experiment Result

We carry out the AS connection link prediction in chapter 2 AS connection diagram of Figure 1. AS shown in Figure 1 there are 11 nodes and get the link pair of 11 * 10/2 = 55, but in the actual IP network end-to-end detection, the actual connection link node pair consists of (1, 5), (1, 6), (5, 6), (1, 2), (2, 7), (2, 8), (3,8), (1, 4), (3, 4), (4, 9), (9, 10), (4, 10), (7, 8) and (4, 11). And some links could not obtain their connection relationship. We divide these actual known links into two set, test link set and training link set through a certain ratio value. By using the above local information algorithm and corresponding improvement method, we try to predict the link connection character. The training set is expressed as E^T and testing set as E^P, the test link set E^P and those not been observed link (U - E^T) has a similarity score values Score(x, y), said the possibility size of the existing link between x and y. According to the similarity algorithm, we get the similarity score values to all the nodes of (x, y), but we do only pay attention to the size of the disconnection node pairs' similarity scores. That is to say, when we carry out the end-to-end detection, if some link could not receive the response information in ICMP packet, and missing this link connection information, through the local similarity link prediction we can obtain the most likely missing link. During the experiment, we set a certain ratio among the training set and testing set, and take the average to the data of 100 times.

4.1. Result of the Similarity Algorithm based on Traditional Common Neighbor

The AUC values to different algorithm index based on local information common neighbor algorithm respectively is shown as Table 2.

Table 1. AUC Value based on Local Information Common Neighbor Algorithm

index	CN	Salton	Jaccard	Sørensen	HPI	HDI	LHN	AA	RA
AUC	0.8967	0.9302	0.9306	0.9308	0.9157	0.9226	0.9309	0.9568	0.957

From the Table 1 common neighbor similarity index we can see, the prediction accuracy all can reach above 0.9, which can achieve good results. And Salton, Jaccard, Sørensen, LHN, AA, and RA all can reach about 0.91, which shows better prediction accuracy. We can find the prediction accuracy of index of AA and RA are the best two indexes. The prediction accuracy of the PA index is the worst less than 0.5, because the definition of PA index is to give greater weight to having big value of the two nodes degrees' product.

4.2. Result of Improved Algorithm Model

In chapter 3, we put forward an improvement method to the index of CN based on the Bayesian model. Through the experiment, the result of link prediction (IM_CN) combining with the AA index model is 0.9569. The result is listed at Table 2, form the test result we can see an obvious improve, the origin prediction result is 0.8967. And under the Bayesian model, the result of CN index (Bay_CN) is 0.957, which has not obvious improved. Through using the same method, we try to improve the Bay_CN algorithm, and through experiment, the result of improved Bay_CN (IM_Bay_CN) is 0.9699, which beyond the entire traditional local information prediction index. Through analysis because the Bayesian model formula carry out the stronger Bayesian hypothesis to the neighbor role R_w, which represents the other user's contribution to the candidate

user are independent, which is suitable to the IP network node link structure, the independence assumption conform to the connection properties of IP network. Therefore, the result after Bayesian hypothesis is better than the traditional CN index, and through improved again, the monitor accuracy is higher than others. So, the local information similarity algorithm based on Bayesian model and its improved is most suitable for the IP network link prediction.

Table 2. AUC Value of Improved Algorithm

Index	IM_CN	Bay_CN	IM_Bay_CN
AUC	0.9569	0.957	0.9699

From the Table 2, IM_CN, Bay_CN and IM_Bay_CN respectively get improvement compare with the index of traditional local information prediction index. The same, the Bayesian model adopting the improved method combining with the index of AA can also get some degree improvement. In the future work, we will focus on the complex large-scale IP network link prediction, and verify the prediction accuracy. In addition, to find if the method can be used into the IP network missing link recovers.

5. Conclusion

Through the analysis and research to the link prediction algorithm and corresponding index, and the simulation experiment on the IP network model, we can see from the experimental results that the local information similarity indexes and their optimized index can accurately predict the link connection. And based on the Bayesian independence assumption is also suitable to the structure of the IP network properties, and after introducing improved index based on index of AA or RA into CN, Bay_CN, and the prediction accuracy can gains a certain degree improvement. This paper only discuss a kind of fixed AS model experiment of IP network, from the prediction accuracy and stability, the algorithm based on the common neighbor index and their improved index has higher prediction accuracy, the prediction accuracy can reach more than 0.96. As a result, the neighbor algorithm based on the local information link prediction index is suitable to the IP network link prediction. But considering the IP network structure with complex connection, and especially in large-scale IP network, it is hard to judge, whether the link predict can also obtain good results is the future research direction.

References

[1] R. Rao Kompella, J. Yates, A. Greenberg, A. C. Snoeren, “IP Fault Localization Via Risk Modeling”, 2nd Symposium on Networked Systems Design and Implementation, Boston, Massachusetts, USA, **(2005)** May 2-4.
[2] S. Kandula, D. Katabi, J-P Vasseur, “Shrink: A Tool for Failure Diagnosis in IP Network”, SIGCOMM'05 Workshops, Philadelphia, PA, USA, **(2005)** August 22-26.
[3] T. Qiu, Z. Ge, D. Pei, “What Happened in my Network? Mining Network Events from Router Syslogs”, IMC’10, Melbourne, Australia, **(2010)** Nov.1-3.
[4] X. Xiren, “Computer Network, Dalian University of Technology Press, Dalian **(2000)**. p. 174.
[5] 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.
[6] R. Albert, A L, Barabási, “Statistical mechanics of complex networks”, Reviews of modern physics, vol. 74, no. 1, **(2002)**. p. 47.
[7] L.Lu, C.H. Jin, T. Zhou, “Similarity index based on Local paths for link prediction of complex network”, Phys. Rev. E, 80, **(2009).**
[8] G. Salton, M J McGill, “Introduction to Modern Information Retrieval”, New York: McDraw-Hill Co., **(1983)**, pp.30-42.
[9] T. Sørenson, “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons”, Biol Skr, vol. 5, no.4, **(1948)**. pp. 1-34.

[10] E A Leicht , P. Holme , Newman M E J. "Vertex similarity in networks", Physical Review E, vol.73, no.2. **(2006)**.
[11] E. Ravasz, A L Somera , D A Mongru, *et al*. "Hierarchical organization of modularity in metabolic networks", science, **(2002)**, vol. 297, no. 5586, pp. 1553-1555.
[12] Adamic L A, Adar E. Friends and neighbors on the web. Social Networks, vol. 25, no. 3, **(2003)**, pp.211-230.
[13] T. Zhou, L. Lü, Y C. Zhang, "Predicting missing links via local information", The European Physical Journal B, vol. 71, no. 4, **(2009)**. pp. 623-630.
[14] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Societe Vaudoise des Science Naturelles, no. 37, **(1901)**. pp. 547-579.
[15] W. Jie-hua, Z. An-qing, C. Xue-lian, Z. Xiao-lan, "Hidden naive Bayesian model for social relation recommendation", Application Research of Computers, vol. 31, no. 5, **(2014)**. pp. 1382-1383.

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 Institute of Aeronautical Industry Management, Department of Electronic Communication and Engineering, assistant professor. His research interests are circuit and system, data collection and signal process, network information and network security, *etc*.

Duan Zhemin, 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*.