Asymmetry similarity coefficient Method for Link Prediction in Homogeneous Graph Data

Rui Xie, Zhifeng Hao and Bo Liu

***Abstract*—**Predicting missing link in homogeneous networks is of both theoretical interest and practical significance in many different fields. It is found that the similar degree is different between a pair node and the similarity is asymmetry. In this paper, we deliver an efficient framework for link prediction on the basis of node similarity coefficient. A new similarity measure, motivated by the similar coefficient taking place on networks, is proposed and shown to have higher prediction accuracy than previous similarity measures. We therefore design another new measure exploiting information on the common neighbors, which can remarkably enhance the prediction accuracy. Experiments on benchmark and real-world data sets have demonstrated the effectiveness of our proposed approach**.**

***Keywords*—**asymmetry, link prediction, asymmetry similarity coefficient

I. INTRODUCTION

THE link prediction algorithm is one of the key technologies to reveal the inherent rule of network evolution [1, 2]. Link prediction algorithms aim at estimating the likelihood of the existence of links between nodes based on observed links, attributes of nodes, and structural properties of network. It can be divided into two categories, One is predicting missing links or existent yet unknown links, the other one is predicting those links that may exist or appear in the future of evolving networks [2, 3].

In recent years, due to its theoretical value and practical significance in modern science, the problem of missing link prediction has been intensively studied by researchers from different backgrounds and many methods applied to different field have been proposed [2, 3]. Therein some algorithms are based on Markov chains [4-6] and machine learning [7, 8], which first propose a proper preference assumption and then designs a corresponding loss or objective function to be optimized for two nodes. Another group of algorithms are based on node similarity, that two nodes are more likely to have a link if they are similar to each other, therefor, the essential problem for link prediction is how to calculate the similarity between nodes accurately [2, 3].

However, most of the existing work focuses on link prediction based on similarity symmetry [2, 3], and the similarity asymmetry has not been explicitly addressed. In this paper, we address the similarity asymmetry with similarity coefficient, where the miss link prediction method is built on similarity, and the similar degree between nodes is incorporated to improve it. In our simulation, we find that the similarity-coefficient-base method is very effective when applied to missing link prediction. We finally integrate the new method with the common neighbors (CN) [9], Adamic-Adar (AA) [10], resource allocation (RA) [11] methods, and find that the new method can substantially out-perform the existing methods.

The main contribution of our method can be viewed from the following aspects.

1) Our method considers the similarity asymmetry of two connected nodes in the homogeneous network, and then develops the current similarity methods by incorporating this kind of information into the link prediction.

2) Our method can be well expanded based on the existing methods. We put forward our four versions of asymmetry similarity coefficient methods based on common neighbors, Adamic-Adar and resource allocation methods, which have been found to be more accurate than themselves.

3) Substantial experiments on the benchmark and real-world data sets show that our developed methods can obtain better prediction accuracy than the existing methods.

The rest of this paper is organized as follows. Section 2 reviews the related work. The detail of our proposed approach is presented in Section 3. Experiments are shown in Section 4. Section 5 concludes the paper and discusses possible directions for future work.

II. RELATED WORK

Considering an unweight undirected simple network G (V, E) without multiple links and self-connections, V is the set of nodes and E is the set of links. For each pair of nodes $x, y \in$ V, we calculate a score sxy that measures the likelihood for node x and y whether a link between them. There are many different

This work was supported in part by the Natural Science Foundation of China under Grant 61472090, Grant 61472089 and Grant 61672169, in part by the NSFC Guangdong Joint Found U1501254, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant S2013050014133, in part by the Natural Science Foundation of Guangdong under Grant 2015A030313486, in part by the Science and Technology Project of Guangdong Province under Grant 2015B010128014 and Grant2016B010107002,in part by the Science and Technology Planning Project of Guangzhou under Grant 201707010492.Grant 201604016003,Grant 201604016067 and Grant 201604016041

Rui Xie is currently a Ph.D. Candidates in Guangdong University of Technology, Guangzhou 510006, Guangdong, China. (corresponding author; e-mail: 25457855@qq.com).

Zhifeng Hao is with the School of Computer, Guangdong University of Technology, Guangzhou 510006, Guangdong, China.

Bo Liu is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, Guangdong, China.

measures to calculate s_{xy} score, the most common and straightforward method is to calculate the similarity between node *x* and *y*.

Predicting the likelihood of the existence of links between two nodes according to their common features is a long-term problem in modern information science, and there are many methods to measure similarity based on common features. Generally speaking, two nodes are considered to be similar if they have some common important features in topology or other attributes. An review paper on these similarity indices in ref. [2,3]. One of the most straight-forward method is based on the structure of networks, which is named structural similarity [12] and it can be categorized as node-similarity [13, 14], path-similarity[15, 16] and mix methods [17]. However, these methods have serious shortcomings as it strongly favors the large degree nodes. To solve this problem, many variants, such as the Jaccard index [13] and Salton index [14], have been proposed to remove this tendency. In addition, some other methods including Katz index [15], simrank [18], hierarchical random graph [17] and stochastic block model [19, 20], are also very effective in estimating node's similarity. In addition, there are many other similarity measures for nodes in network [17]. These models provide different understanding of the similarity of nodes in network, but performances of these methods for different networks are obviously different. It is found that structural similarity-based methods can achieve satisfying algorithmic accuracies when they are applied to predicting missing links in networks [21-29].

Most of the existing methods on similarity are proposed for similarity symmetry, and the similarity asymmetry has not been explicitly addressed. In real-world applications, the similar degree is not the same for each other, it need a lot of auxiliary information to distinguish the difference of similar degree, but it is always difficult to obtain the contents and attributes of nodes to judge, Moreover, the structure of network information, especially the degree of node, which reflects the link information can sometimes be obtained, though it may not be complete and accurate, it can be utilized to describe the asymmetry of similarity between nodes and improve the missing link prediction accuracy. In this paper, we propose a asymmetry similarity coefficient that builds a missing link prediction method on complex network by incorporating the similarity asymmetry.

III. MISSING LINK PREDICTION BASED ON ASYMMETRY SIMILARITY COEFFICIENT

In this section, we will present our asymmetry similarity coefficient method to predict missing link in the homogeneous graph data. As discussed before, the previous methods such as CN, AA and RA are all based on the calculation of local structure information. They typically assign weights to all neighbor nodes uniformly and treat the number of neighbor nodes as a measure of link relevance. However, it is not adequately taken into account the ratio of the common neighbors to total neighbors relatively, which we define as asymmetry similarity coefficient to reflect the difference of relative similarity. And then it can be used to revise the similarity measure and determine whether there is a link between nodes.

We first introduce asymmetry similarity coefficient in section 3.1 and present our revised version of the existing methods in section 3.2, 3.3 and 3.4 respectively.

A. . Asymmetry similarity coefficient

As mentioned before, Given a vector V_x or V_y describing the feature of a node *x* or *y*, supposed the degree of node *x* is k_x and the neighbor set is $\Gamma(x)$, the degree of node *y* is k_y and the neighbor set is $\Gamma(y)$, so the common neighbor is $\Gamma(x) \cap \Gamma(y)$, we calculate the similarity s_{xy} of the pair nodes based on the common neighbors of V_x and V_y. For example, a simple network as shown in the following Fig.1:

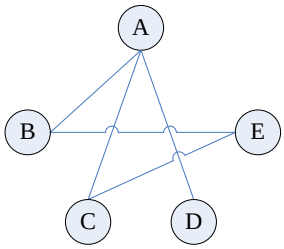

Fig. 1. A simple network

Based on the similarity define, the node A and E are similar. We note, the neighbor of node A is $\Gamma(A) = \{B, C, D\}$ and the neighbor of node E is $\Gamma(E) = \{B, C\}$, so the common neighbor is $\Gamma(A) \cap \Gamma(E) = \{B, C\}$; for the node A has 2 common neighbor nodes of its 3 neighbor nodes, but for the node E , it has all of its 2 neighbor nodes in common neighbors. We found that the degree of similarity between the two nodes is different. For the node E, it is very similar to part of node A, however, node A is partially similar to node E relatively. That the degree of similarity between A and E is relative different for each other. On these grounds, we can define a asymmetry similarity coefficient θ_z^t to reflect the similar degree of two nodes relatively. The formula as follows:

$$\theta_z^t = \frac{\Gamma(z)}{k_t} \qquad (k_t \neq 0) \tag{1}$$

Where $\Gamma(z)$ denotes the common neighbors of the pair nodes, k_t denotes the degree of each node, if any node degree k_t=0, then there is no link with other nodes, so the common neighbor$\Gamma(z)$=0 and the asymmetry similarity coefficient θ_z^t=0. For example in Fig. 1, θ_{AE}^A =2/3, θ_{AE}^E =1, the asymmetry similarity coefficient result indicates that node E is much similar to node A than node A similar to E, because it is just part of A similar to E, so the asymmetry similarity coefficient can better describe the relative similar degree between node A and E. It is easy to see that the similarity coefficient matrix is not necessarily symmetric, and it is symmetric only when the node degree is equal to each other.

From the preceding analysis, the similarity matrix is asymmetric actually and the symmetry is just only a special case when the degree of nodes are equal to each other. Taking

into account asymmetry similarity coefficient, it should revise the similarity measure, so the revise define of similarity is showed as follows:

$$s_{xy}^{t} = s_{xy} * \theta_{xy}^{t} \qquad t \in (x, y) \tag{2}$$

Where s_{xy} denotes symmetric similarity measure by aforementioned measure, such as CN, AA and RA, θ_{xy}^{t} denotes the similarity coefficient, s_{xy}^{t} denotes asymmetric similarity taking into account asymmetry similarity coefficient, which is asymmetric in general in undirected acyclic graph if and only if θ_{xy}^{t} is equal to each other, and if similarity coefficient θ_{xy}^{t}=1, it means the similarity is symmetry without similarity coefficient, otherwise it is asymmetric. That is the essentially different from symmetric similarity measure which aforementioned in general cases.

B. Asymmetry Similarity Coefficient Based on CN

The similarity measure defined in CN algorithm is the common numbers. Considering the influence of the two nodes, it has several indexes in different views and various ways, such as Salton index, Jaccard index, Sorensen index, HPI index, HDI index,LHN-1 index, etc. Under these basic similarity indexes, the greater the asymmetry similarity coefficient of node *x* and *y*, the more possibility of a link between them based on their common neighbors. Thence, the CN index can be revised based on asymmetry similarity coefficient (SC-CN), it can be defined as follows:

$$s_{xy}^{SC-CN}(x) = \propto * s_{xy}^{CN} * \theta_{xy}^{x} \tag{3}$$
$$s_{xy}^{SC-CN}(y) = \beta * s_{xy}^{CN} * \theta_{xy}^{y} \tag{4}$$

Where s_{xy}^{CN} denotes the similarity based on CN algorithm; θ_{xy}^{x} is the similarity coefficient which node *x* relative to node *y*, θ_{xy}^{y} is the asymmetry similarity coefficient which node *y* relative to node *x*; α and β are node coefficients which reflect the node's importance, weight or other factors. It is found that CN measure is the specific case of SC-CN measure. We can proof as follow:

Without considering asymmetry similarity coefficient and other factors, the parameters can be taken as:

$$\theta_{xy}^{x} = \theta_{xy}^{y} = 1 \tag{5}$$
$$\alpha = \beta = 1/2 \tag{6}$$

Then we can calculate the sum of $s_{xy}^{SC-CN}(x)$ and $s_{xy}^{SC-CN}(y)$:

$$s_{xy}^{SC-CN}(x)+s_{xy}^{SC-CN}(y) = \propto * s_{xy}^{CN} * \theta_{xy}^{x} + \beta * s_{xy}^{CN} * \theta_{xy}^{y} = s_{xy}^{CN} \tag{7}$$

It means that the similarity measure s_{xy}^{CN} which based on common neighbor (CN) is a specific case of the measure s_{xy}^{SC-CN} of asymmetry similarity coefficient based on CN (SC-CN).

In this way, we can extend CN index to SCCN index, Salton index to SCSalton index, Jaccard index to SCJaccard index, Sorensen index to SCSorensen index, HPI index to SCHPI index, HDI index to SCDHI index. In the experiments, we will investigate the extended versions and the original ones.

C. Asymmetry Similarity Coefficient Based on AA

The AA index assigns a weight for each node based on common neighbor, and the weight is the degree of the node's logarithm ($\log k_z$)-1. The basic idea of AA algorithm is the smaller degree of common neighbor node, the larger contribution for the similarity measure. Taking into account asymmetry similarity coefficient, it means the greater the asymmetry similarity coefficient of node *x* and *y*, the more possibility of a link between them. Thence, the AA index can be revised based on asymmetry similarity coefficient (SC-AA), it can be defined as follows:

$$s_{xy}^{SC-AA}(x) = \propto * s_{xy}^{AA} * \theta_{xy}^{x} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\propto * \theta_{xy}^{x}}{\log k_z} \qquad (k_z > 1) \tag{8}$$

$$s_{xy}^{SC-AA}(y) = \beta * s_{xy}^{AA} * \theta_{xy}^{y} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\beta * \theta_{xy}^{y}}{\log k_z} \qquad (k_z > 1) \tag{9}$$

Where s_{xy}^{AA} is the similarity based on AA algorithm, if k_z=1, it means nodes *x* and *y* only connect each other, then $s_{xy}^{SC-AA}(x)$=$s_{xy}^{SC-AA}(y)$=1, else if k_z=0, it means they have no neighbor, then $s_{xy}^{SC-AA}(x)$=$s_{xy}^{SC-AA}(y)$=0. It is found that AA measure is the specific case of SC-AA measure, We can proof as follow:

Without considering asymmetry similarity coefficient and other factors, the parameters can be taken as: equation (5) and equation (6), under this condition:

$$s_{xy}^{SC-AA}(x)+s_{xy}^{SC-AA}(y) = \propto * s_{xy}^{AA} * \theta_{xy}^{x} + \beta * s_{xy}^{AA} * \theta_{xy}^{y} = s_{xy}^{AA} \tag{10}$$

It means that the similarity measure s_{xy}^{AA} which based on AA algorithm is a specific case of the measure s_{xy}^{SC-AA} of asymmetry similarity coefficient based on AA algorithm (SC-AA).

D. Asymmetry Similarity Coefficient Based on RA

RA algorithm supposes each medium has a unit of resources and will equally distributed it to all its neighbors, and the similarity measure is defined as the number of resources that a node can accept. Considering the asymmetry similarity coefficient, it means the greater the asymmetry similarity coefficient of node *x* and *y*, the more likely a link between them. Thence, the RA index can be combined with asymmetry similarity coefficient (SC-RA), it can be defined as follows:

$$s_{xy}^{SC-RA}(x) = \propto * s_{xy}^{RA} * \theta_{xy}^{x} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\propto * \theta_{z}^{x}}{k_z} \qquad (k_z \neq 0) \tag{9}$$

$$s_{xy}^{SC-RA}(y) = \beta * s_{xy}^{RA} * \theta_{xy}^{y} = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{\beta*\theta_z^y}{k_z} \quad (k_z \neq 0) \qquad (10)$$

Where s_{xy}^{RA} is the similarity based on RA algorithm, if k_z=0, it means they have no neighbor, then $s_{xy}^{SC-RA}(x)$=$s_{xy}^{SC-RA}(y)$=0. Without considering asymmetry similarity coefficient and other factors, the parameters can be taken as equation (5) and equation (6), under this condition:

$$s_{xy}^{SC-RA}(x)+s_{xy}^{SC-RA}(y)=\propto * s_{xy}^{RA} * \theta_{xy}^{x} + \beta * s_{xy}^{RA} * \theta_{xy}^{y} = s_{xy}^{RA} \qquad (11)$$

It means that the similarity measure s_{xy}^{RA} which based on RA algorithm is a specific case of the measure s_{xy}^{SC-RA} of asymmetry similarity coefficient based on RA algorithm (SC-RA).

IV. Experiments and Results

We conduct experiments on the benchmark data sets and real-world sentiment data set to investigate the performance of our proposed method. All the experiments run on the Windows platform with a 2.8-GHz processor and 6-GB DRAM. The objectives of our experiments are as follows.

1) To evaluate the effectiveness of our method with varying numbers of training set ratio constraints.

2) To evaluate the performance variation of our method with different data sets constraints.

3) To evaluate the improvement of our method with different data sets constraints.

In this section, we will investigate the extended versions and original methods in CN, AA and RA categories. The compared methods are CN index and SCCN index, Salton index and SCSalton index, Jaccard index and SCJaccard index, Sorensen index and SCSorensen index, HPI index and SCHPI index, HDI index and SCDHI index, AA index and SCAA index, RA index and SCRA index. To analyze the performances of our measures, three groups of experimental comparisons between our measures and other well-known measures on four real networks are made in this section. All these measures are based on the information of common neighbors. Brief introduction of those well-known measures and the similar experiments were given by Zhou et al.in ref [24]. Consider an undirected simple network G(V,E) without multiple links and self-connections, where V is the set of nodes and E is the set of links. the observed links (E) are randomly divided into two parts: the training set (ET) is treated as known information, while the probe set (EP) is used for testing and no information is allowed to be used for prediction. For each pair of nodes, $x, y \in V$, every algorithm referred to in this paper assigns a score s_{xy}. This score can be viewed as a measure of similarity between nodes x and node y. All the nonexistent links are sorted in decreasing order according to their scores. And the links at the top are most likely to exist.

To test the algorithm's accuracy of similarity measure based on coefficient, We adopt a standard metric that the area under the receiver operating characteristic (ROC) curve [11,17] to quantify the accuracy of the prediction algorithms. It can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. We can define the accuracy as follows:

$$AUC = \frac{N' + 0.5n''}{n} \qquad (12)$$

If all the scores are generated from an independent and identical distribution, the accuracy should be about 0.5. Therefore, the degree to which the accuracy exceeds 0.5 indicates how much better the algorithm performs than pure chance [2,11].

A. . Description of Experiment Data

We conduct experiments on ten benchmark data sets: USAir, Yeast, NetScienc, Jazz, which are popularly used in the existing link prediction work [24]. Each data set is homogeneous Graph Data, undirected and unweight, without multiple links, self-connections.

Table. 1 Topological features of those four representative networks.

Name	Type	Nodes	Edges	weighted	Description
USAir	Information Network	1532	2126	weighted	Network of the USA airline
NetScience	Social Networks	1536	2742	weighted	network of coauthorships
Jazz	Collaboration Networks	198	5484	unweighted	network of Jazz musicians
Yeast	Biological Networks	2375	11693	unweighted	network or protein interaction

(1) USAir. The network of the US air transportation system. Which contains 1532 airports and 2126 airlines[1].

(2) Yeast. A network representing the interactions among proteins .there are 2375 nodes and 11693 links in this biological networks.

(3) NetScience . A network of co-authorships between scientists. Containing 1536 scientists and 2742 cooperation-ships[30].

(4) Jazz: a music collaboration network obtains from the Red Hot Jazz Archive digital database. Here it includes 198 bands that performed between 1912 and 1940, with most of the bands in the 1920 and 1940[31].

Table. 1 summarizes the basic topological features of those four representative networks.

[1] http://vlado.fmf.uni-lj.si/pub/networks/data/.

B. Comparison Experiment

The four data sets are randomly divided into training set(ET) and test set(EP) respectively, E= ET∪EP and ET∩EP = ∅， The ratio of the training set is c= ET /E. For simplicity, we do not take into account other factors and adopt $\alpha = \beta = 1/2$ in experiments. Respectively, we test the algorithm based on asymmetry similarity coefficient in the following different situations, each experiment is conducted 100 times and AUC is the average of 100 times.

(1) Experiment 1: Different ratios of the training in the same data.

To analyze our proposed approach, we test the AUC of missing link prediction based on asymmetry similarity coefficient in USAir data set with different ratios of the training by SCCN, SCAA an SCRA methods respectively, the results as shown in Fig.2.

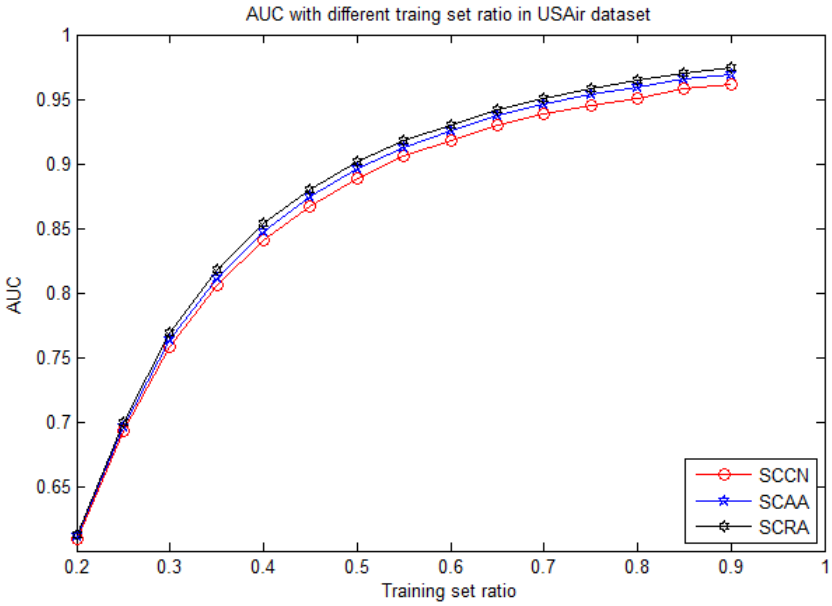

Fig. 2. AUC with different training set ratio in USAir data set

As shown in Fig. 2, the AUC value increases when the ratio of training set goes up, it is found that the accuracy of missing link prediction AUC is rising. With the increase of the ratio of the training set in USAir data set. The accuracy of link prediction of SCRA and SCAA algorithms has more than 90% when the ratio of the training set is 0.6, and it has reached more than 95% when the ratio is 0.9. The similar findings can also be observed in other three datasets, then experimental results show that the link prediction method based on asymmetry similarity coefficient is effective in missing link prediction..

(2)Experiment 2:Ratio constant with different data sets

Based on the previous experiment, we chose the ratio is 0.9 in the experiments, and we use our proposed approaches, SCCN, SCSalton, SCJaccard, SCSorens, SCHPI, SCHDI, SCAA, SCRA to test the accuracy of missing link prediction based on similarity coefficient in four data sets, the experiment results as shown in table 2:

Table. 2 c=0.9,the AUC of different predictions bases on coefficient in different data sets

data set	SCCN	SCSalton	SCJaccard	SCSorens	SCHPI	SCHDI	SCAA	SCRA
USAir	0.9604	0.9301	0.9262	0.9256	0.898	0.9229	0.969	0.974
NetScience	0.9925	0.9924	0.9924	0.9924	0.992	0.9923	0.993	0.993
Yeast	0.9155	0.9145	0.9147	0.9146	0.914	0.9146	0.916	0.916
Jazz	0.967	0.9663	0.9659	0.9661	0.957	0.9637	0.97	0.975

Table 2 presents the AUC on the benchmark data sets when the number of training set ratio is 90%n, where n is the size of the data set. The AUC of our proposed approaches have exceeded 90% in these data sets, especially in the NS network, more than 99%. From the above example, it can be seen that our proposed approaches have a good predictive effect in these data sets.

(3)Experiment 3: Compare with existing methods

We compare our proposed approaches with existing methods in different data sets on ratio 0.9 to verify the validity of our measure algorithms. We test our proposed approaches with existing methods in eight group experiments, the experiment results as shown in Fig. 3, Fig. 4 , Fig. 5 and Fig. 6.

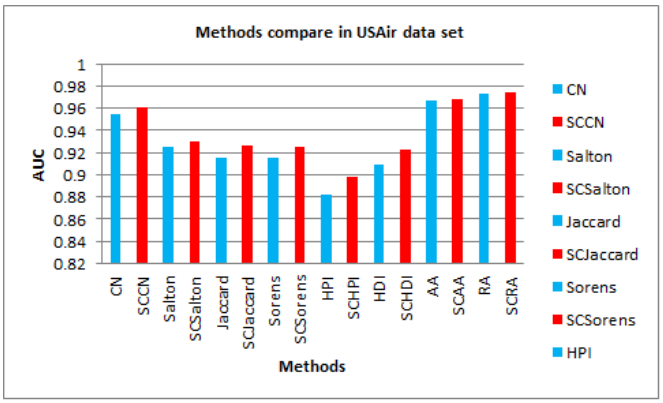

Fig. 3. Compare in USAir data set

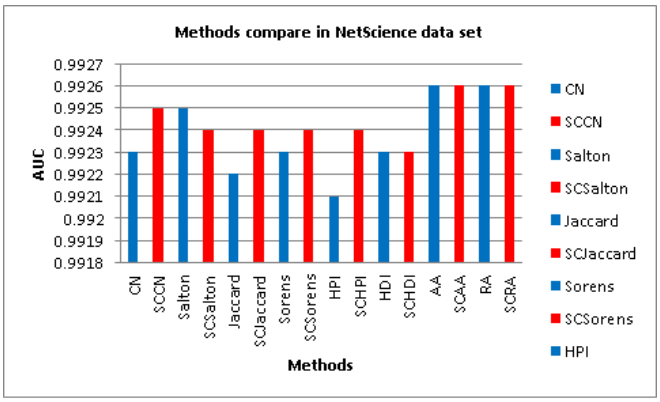

Fig. 4. Compare in NetScience data set

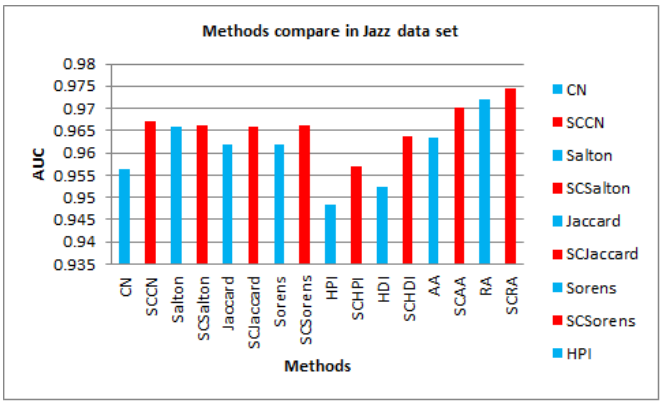

Fig. 5. Compare in Jazz data set

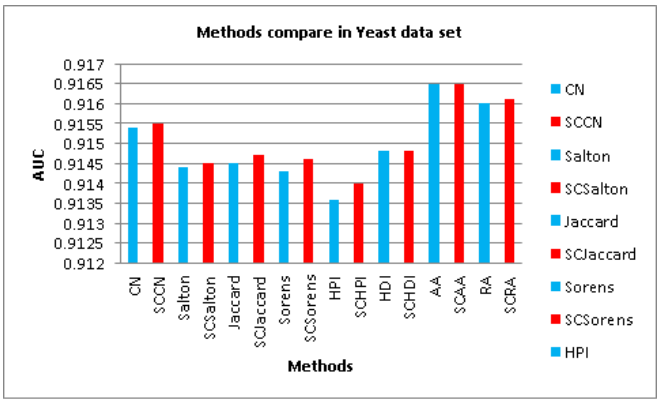

Fig. 6. Compare in Yeast data set

As shown in Figure 3, the different methods have different AUC in USAir data set. The best one is SC-RA while HPI index is poor. It is found that the asymmetry similarity coefficient method is better than the existing method in the eight groups of experiments. The results show that our proposed method has good prediction effect in information network.

As shown in Figure 4, the different methods have different AUC in Yeast data set. The best one is SCAA while HPI index is poor. It is found that the asymmetry similarity coefficient method is better than the existing method in the eight groups of experiments. The results show that our proposed method has good prediction effect in information network.

As shown in Figure 5, the different methods have different AUC in Yeast data set. The best one is SCAA while HPI index is poor. It is found that the asymmetry similarity coefficient method is better than the existing method in the eight groups of experiments. The results show that our proposed method has good prediction effect in information network.

As shown in Figure 6, the different methods have different AUC in Yeast data set. The best one is SCRA while HPI index is poor. It is found that the asymmetry similarity coefficient method is better than the existing method in the eight groups of experiments. The results show that our proposed method has good prediction effect in information network.

The comparison results show that the missing link prediction based on coefficient has a higher accuracy than existing methods in most data sets; it shows that the method proposed in this paper has a general scope of application and can improve the accuracy of prediction.

In addition, it is also found in this paper that the accuracy is not high enough with low ratio of training set, because the asymmetry similarity coefficient reflects the relative similar degree between nodes and it is difficult to obtain better performance while insufficient training. Therefore, this method will be better performance after sufficient training.

We have noticed that out proposed approaches all most have better accuracy than existing methods in these data sets, and only a few our methods have the same effect as the original methods in NetScience and Yeast data sets.

V. CONCLUSION AND DISCUSSION

In this paper, we analyze the relative similar degree between nodes and found that the similarity is asymmetric between them in general. So, the asymmetry similarity coefficient was presented to reflect the degree of relative similarity between nodes. Based on it, we design a new method to measure the similarity and propose a new approach use to missing link prediction, it was experimented in several real data sets and the results demonstrate that new method can significantly improve the accuracy of missing link prediction compared with other well-known measures. On the other hand, the proposed algorithm in this paper has high complexity$O(n^2)$; it is not efficient in dealing with large-scale network. We will focus on asymmetry similarity coefficients and find ways to reduce the complexity of the algorithm.

REFERENCES

[1] Yang, Xu Hua, et al. "Link prediction based on local community properties", *International Journal of Modern Physics B*, vol. 30, no. 31, pp.437-448, 2016.

[2] Lü, Linyuan, and T. Zhou. "Link prediction in complex networks: A survey", *Physica A Statistical Mechanics & Its Applications*, vol. 390, no. 6, pp. 1150-1170, 2010.

[3] Martínez, Víctor, F. Berzal, and J. C. Cubero. "A Survey of Link Prediction in Complex Networks", *Acm Computing Surveys*, vol. 49, no. 4, pp. 69-81, 2016.

[4] Bilgic, M, G. M. Namata, and L. Getoor. "Combining Collective Classification and Link Prediction", *IEEE International Conference on Data Mining Workshops IEEE Computer Society*, 2007, pp.381-386.

[5] Wei, Zhuoyu, et al. *Link Prediction via Mining Markov Logic Formulas to Improve Social Recommendation. Knowledge Graph and Semantic Computing: Semantic, Knowledge, and Linked Big Data*. Springer Singapore, 2016.

[6] Zhu, Jianhan, J. Hong, and J. G. Hughes. "Using Markov Chains for Link Prediction in Adaptive Web Sites", *Lecture Notes in Computer Science*,no. 2311, pp. 60-73,2001.

[7] Wang, Chao, V. Satuluri, and S. Parthasarathy. "Local Probabilistic Models for Link Prediction", *IEEE International Conference on Data Mining IEEE*, 2007, pp.322-331.

[8] Popescul, Rin, and L. H. Ungar. "Statistical Relational Learning for Link Prediction", *In Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*, 2003, pp.221-235.

[9] Newman, M. E. "Clustering and preferential attachment in growing networks", *Working Papers*, vol. 64, no. 2, pp.025102-, 2001.

[10] Adamic, Lada A, and E. Adar. "Friends and neighbors on the Web", *Social Networks*, vol. 25, no. 3, pp. 211-230, 2003.

[11] Zhou, Tao, L. Lü, and Y. C. Zhang. "Predicting missing links via local information", *European Physical Journal B*, vol. 71, no. 4, pp. 623-630, 2009.

[12] Guo-Dong Lyu, et al. "Predicting missing links via structural similarity", *International Journal of Modern Physics B*, vol. 29, no.15, pp. 1550095-, 2015.

[13] Jaccard, Paul. "Etude de la distribution florale dans une portion des Alpes et du Jura", *Bulletin De La Societe Vaudoise Des Sciences Naturelles*, vol. 37, no. 142, pp. 547-579, 1901.

[14] T. A. Sørensen. "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons", *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab*, vol. 5, pp. 1-34, 1948

[15] Katz, Leo. "A new status index derived from sociometric analysis", *Psychometrika*, vol. 18, no. 1, pp.39-43, 1953.

[16] Lü, Linyuan, C. H. Jin, and T. Zhou. "Effective and Efficient Similarity Index for Link Prediction of Complex Networks", *Physics*, vol. 40, pp. 1-9, 2009.

[17] Lü L and Zhou T, " Link Prediction", *Higher Education Press*, Beijing, 2012.

[18] Jeh, Glen, and J. Widom. "SimRank: a measure of structural-context similarity", *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining ACM*, 2002, pp.538-543.

[19] Lü, L, C. H. Jin, and T. Zhou. "Similarity index based on local paths for link prediction of complex networks", Phys Rev E Stat Nonlin Soft Matter Phys, vol. 80, no. 2, pp.046122-, 2009.

[20] Lin, L. I. "A concordance correlation coefficient to evaluate reproducibility", *Biometrics*, vol. 45, no. 1, pp.255-268, 1989.

[21] Lü L, Zhou T. Link prediction in weighted networks: The role of weak ties[J]. Epl, 2010, 89(1):18001.

[22] Liu, Zhen, et al. "Link prediction in complex networks: a local naïve Bayes model", *Epl*, vol. 96, no. 4, pp. 48007, 2011.

[23] Liao, Hao, A. Zeng, and Y. C. Zhang. "Predicting missing links via correlation between nodes", *Physica A Statistical Mechanics & Its Applications*, vol. 436, no.1, pp. 216-223, 2015.

[24] Upasana - Sharma, S. K. Khatri, and L. M. Patnaik. "Link Prediction in Social Networks: A Neuro-Fuzzy Approach", *International Journal of Advances in Soft Computing & Its Applications*, vol. 45, no. 2, pp. 122-136, 2016.

[25] Liu, Jie, et al. "A link prediction algorithm based on label propagation", *Journal of Computational Science*, vol. 16, no. 1, pp. 43-50, 2016.

[26] Lu, Y., Y. Guo, and A. Korhonen. "Link prediction in drug-target interactions network using similarity indices", *Bmc Bioinformatics*, vol. 18, no.1, pp. 39-51, 2017.

[27] Dai, Caiyan, et al. "Link prediction in multi-relational networks based on relational similarity", *Information Sciences*, vol. 394, no. 395, pp. 198-216, 2017.

[28] Yao, Lin, et al. "Link Prediction Based on Common-Neighbors for Dynamic Social Network", *Procedia Computer Science*, vol. 83, pp. 82-89, 2016.

[29] Farkas, I., et al. "Networks in life: Scaling properties and eigenvalue spectra", *Physica A Statistical Mechanics & Its Applications*, vol. 314, no.1, pp. 25-34, 2012.

[30] Von, Mering C, et al. "Comparative assessment of large-scale data sets of protein-protein interactions", *Nature*, vol. 417, no. 6887, pp. 399-403, 2002.

[31] Newman, M E J. "Finding community strcuture in networks using the eigenvectors of matrics", *Physical Review E*, vol. 74, no. 3, pp. 036104-, 2006.

Rui Xie was born on Feb. 28, 1977. He received his B.S. in electrical engineering and automation from Dalian Maritime University in 2000, and the M.Sc. in Computer Science from Guangdong University of Technology in 2003. He is currently a Ph.D. Candidates in Guangdong University of Technology. His research interests cover a variety of different topics including machine learning, cloud computing, data mining and their applications.

Zhifeng Hao was born on May 10, 1968. He received the BScdegrees in Mathematics from the Sun Yat-Sen University in 1990, and the PhD degree in Mathematics from Nanjing University in 1995. He is currently a Professor in the School of Computer, Guangdong University of Technology, and School of Mathematics and Big Date, Foshan University. His research interests involve various aspects of Algebra, Machine Learning, Data Mining,EvolutionaryAlgorithms.

Bo Liu was born on Apr 19, 1978. He is currently with the School of Automation, Guangdong University of Technology, Guangzhou, China. He has authored papers in the IEEE TRANSACTIONS ON NEURAL NETWORKS and the IEEE TRANSACTIONS ON KNOWLEDGE ANDDATA ENGINEERING. His current research interests include machine learning and data mining.

Xu Shengbing was born on Jun 21, 1975. He received his B.S. in computational mathematics & applied software from Xiangtan University in 1997, and the M.Sc. in Applied mathematics from Xiangtan University in 2001. He is currently a Ph.D Candidates in Guangdong University of Technology. His research interests cover a variety of different topics including mathematical modeling, machine learning, data mining and their applications.