A link prediction-based method for identifying potential cooperation partners: A case study on four journals of informetrics

Lu Huang[1], Zhu Yihe[1], Yi Zhang[2], Zhou Xiao[3], Jia Xiang[1]

[1] School of Management and Economics, Beijing Institute of Technology, Beijing, PR China

[2] Decision Systems & e-Service Intelligence research Lab, Centre for Quantum Computation & Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

[4]School of economics and management, Xidian University

Abstract --Global academic exchange and cooperation have become an increasing trend in both academia and industry, but how to quickly and effectively identify potential partners is becoming an urgent problem. This paper proposes a link prediction-based model to help researchers identify partners from a large collection of academic articles in a given technological area. We initially construct a co-authorship network, and take a series of indices based on network and similarity of researchers into consideration. A fitting model of link prediction is then established, in which logistic regression analysis is involved. An empirical study on four journals of informetrics is conducted to demonstrate the reliability of the proposed method.

Ⅰ. Introduction

With the rapid development of science and technology, the difficulty of scientific research is gradually increasing, and a large number of problems need to be solved by the collaborations of experts. The collaborations are always encouraged, as they often yield a synergistic effect. The combined expertise of a research team can always produce results far surpassing the sum of the individuals' capabilities [1]. However, it is often difficult and time-consuming for a researcher to find the right expert to collaborate with.

Scientific results of cooperation are usually published in the form of scientific documents with researchers' joint signature [2]. As a result, the academic papers becomes an important way to understand the scientific collaborations.

In recent years, some scientists in the fields of biblimetrics had been studying on mining the potential links (or collaborations)between researchers through similarity of their documents .The measure of coupling or similarity are roughly in two ways：a, the methods based on citation theory, for examples co-citation analysis,bibliographic coupling analysis,which calculate the degree of co-citation or coupling of the papers to measure similarity ；b，another methods are derived from co-occurrence theory，which measure the similarity by counting the number of co-occurrences of keywords.All of these methods only reveal the possibility of cooperation from a perspective of research contents and directions. However, in realistic situation, the similarity of contents is not the sole cause to contribute to collaborations. Social relations also have an important place, like co-authorship.

Link prediction in complex networks is another approach to find out two types of links. One is future link, focusing on the dynamics of the network ;the other is existent yet unknown links or so-called missing links, which is the process of data mining. Linben Nowell generalized the methodology of link prediction initiatively. Lv and Zhou [3] [4]summed up three different topology-based methods: joint/conditional probability; models, maximum likelihood estimation, network topological structure .Link prediction was introduced into biblimetrics due to its good effectiveness and accuracy. Linben Nowell and Kleinberg [5] proposed a Near Neighbor method based on topology and evaluated several link prediction index using the co-authorship networks. Naoki Shibata [6] predict the existence of citations among papers by formulating link prediction in 5 large-scale datasets of citation networks. The results indicate that different models are required for different types of research area or networks. Milen Pavlov [1] presented a supervised learning method for building link predictors from structural attributes of the co-authorship network. Yan [7] conducted link prediction on three levels of co-authorship networks between authors, institutions and countries in library information science and found out the accuracy on country-level networks is better than that of the author- and institution-level networks.

The link prediction method receive considerable attention because that is easy to get structure information and more reliable than attribution information of researchers, such as coupling analysis. But this method ignore researchers' interests and field, since the co-authorship networks only reflected the situations of cooperation relationships between the researchers.

In this paper, we combined method of measure similarity of researchers and link prediction to find out potential links among researchers in the networks. And provide recommendations to researchers when he needs someone to work with.

II. METHOD

We use paper data to construct co-authorship networks, the nodes represent the researchers and the edges between them means they had a cooperation on papers up to now , the number of paper determined the weigh of edges. Adding the weigh to network is a better way to portray their partenrships, then improve the accuracy of predict results.

The link prediction in weight networks problem is usually described as:

Consider a network G(V,E,W) with nodes $v_i \in V$ and edges $(v_{i,}v_j) \in E$,i, ≠ j，where $w_{ij} \in W$denotes the weight of edge($v_{i,}v_j$) .Then the task to predict how likely an unobserved edges $(v_{m,}v_n) \notin \mathrm{E}$ exists between an arbitrary of nodes$(v_{m,}v_n)$in the co-authorship network.

We have combined information of weighted networks and the proximity to estimate the possibilities of link between each pair of nodes who have no cooperation before. The indices selected as follows.

(1) Link prediction indices

Common Neighbors[8] is one of the basic index in link prediction ,which means the more common neighbor have ,the more greater the likehood of link .Many other indices are proposed inspired this one .In social networks, if two people who don't know each other have a lot of common friends, they would have a great possibility to be friends，as there are lots of chances to have an encounter. So as in scientific collaboration networks. Zhou[4] analysis the performance of 10 topology-based indices in 6 realistic complex networks, three of them were proved to have an excellent performance , Common Neighbors, Adamic/Adar, Resource Allocation.This is also the reason why we choose these three indices in our work. The corresponds in weight network are listed below. To make an explanation，$\Gamma(\mathrm{x})$ denotes all of the neighborhoods x have，or the degree of x in another perspective；$\mathrm{W_{xz}}$ is the times of collaboration in papers between x and y；Sz denotes the strength of node z, namely the sum of weights of its associated links.

Common Neighbors : The index states that the probability of researchers collaborating increases with the number of other collaborators they have in common.

$$S_{xy}^{CN} = \sum\nolimits_{z \in \Gamma(x) \cap \Gamma(y)} w_{xz} + w_{zy}$$

Adamic/Adar : The index allows counting common neighbors but gives more weight to neighbors that are not shared with many others. Suppose that reseacher x and y have two common neighbor a and b, where a has 10 neighbors but b has only 2 neighbors besides x and y ,then b contributes more to the probability of collaboration between x and y in the future. This measure was initially defined in the context of social networks on the Web[9].

$$S_{xy}^{AA} = \sum\nolimits_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{zy}}{s_z}$$

Resource Allocation: the index punishes the high-degree common neiborhors to higher extent than Adamic/Adar index [10].

$$S_{xy}^{RA} = \sum\nolimits_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w_{xz} + w_{zy}}{\lg(1 + s_z)}$$

(2) Researchers coupling indices

The coupling indices measure researchers' common interest or fields. In particular, If the two researchers have the same keywords, common cited references in their papers published in academic journals, they show similarities.The indices we used are：

Number of Co-citation：There set $\mathrm{S_1(x)}$ denotes all pieces of the references x have cited, similarly, $\mathrm{S_1(y)}$ is y's references set. $\mathrm{w_{xp}}$ is the times of x refer p. For an instance，if $\mathrm{S_1(x) = \{a, b, b, c\}}$, $\mathrm{S_1(y) = \{a, b, d\}}$, the final score = 5

$$S_{xy}^{NC} = \sum\nolimits_{\mathrm{p} \in S_1(x) \cap S_1(y)} w_{x\mathrm{p}} + w_{y\mathrm{p}}$$

Number of Keywords Co-occurrence：The keywords are the representative terms to embody research contents and subject.For an example, we can get the research hotspot in a discipline field during one time through keyword frequency statistics.Just as Number of Co-citation, $\mathrm{S_2(x)}$ denotes researcher x's all keyword in his or her papers. $\mathrm{w_{xk}}$ is the number of keyword x appeared in all pieces of x published.

$$S_{xy}^{NK} = \sum\nolimits_{\mathrm{k} \in S_2(x) \cap S_2(y)} \mathrm{W_{xk}} + w_{y\mathrm{k}}$$

(3) Performance evaluate indices

AUC (Area under the receiver operating characteristic curve) value is a evaluate index of binary classification models and is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example. In link prediction on co-authorship networks, which can be understood as: select two pairs of nodes where one exist edge and the other without. If the probability score of the former is higher than that of the latter, the value add 1 point, equal add 0.5 point and lower without point. The final value is between 0.5 and 1,and the degree AUC value higher than 0.5 point measures how much accurate than the method of link two nodes randomly[11].

Others indices are Precision, Recall, Accuracy, F1 score. The prediction models can make either a positive or a negative prediction concerning the corresponding label y , where a positive prediction means the forecast label y is true. In the positive case, if the edge is exits, the prediction is called TP (true positive); otherwise it is FP (false positive). Conversely, in the negative case, the prediction can be either TN (true negative) or FN (false negative).As in link prediction problems situation, the AUC is better than these four index[12].

$$Precision = \frac{|TP|}{|TP| + |FN|}$$

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|}$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

(4) Method process

For our purposes，

Step1 The paper data is downloaded from Web of Science and cut into two time slices A1 and A2. Then two co-authorship networks are established on basis of data A1, A2 after data cleaning.

Step2 We compute the score of the index above for each pair of nodes in A1，A2 ,then the set of scores is said to form a feature vector. Let's f_{ij} denote the feature vector for each pairs of nodes in V,not only for the subset E. And each f_{ij} has an labels y_{ij}， if edge(v_i, v_i) exits, $y_{ij} = 1$;else $y_{ij} = 0$.

Step3 We train supervised learning algorithm---Logistic Regression on training set A1(with subset f_{ij} correspond to subset y_{ij}).Since Logistic Regression can get probabilities with each sample and we can apply our own standard and custom performance metrics on this probability score to setup threshold in turn classify output in way which best fit our problems. Then the trained Logistic Regression model is tested on corresponding classification data from A2 and then the performance metrics are applied to evaluate the accuracy of the predictions. We also yields probability score of each pair of nodes without edges in A2.

Step4 At the last, we randomly select 10 researchers and list up the most likely researcher to cooperate. And we also give their common keywords. (as show in Fig. 1.)

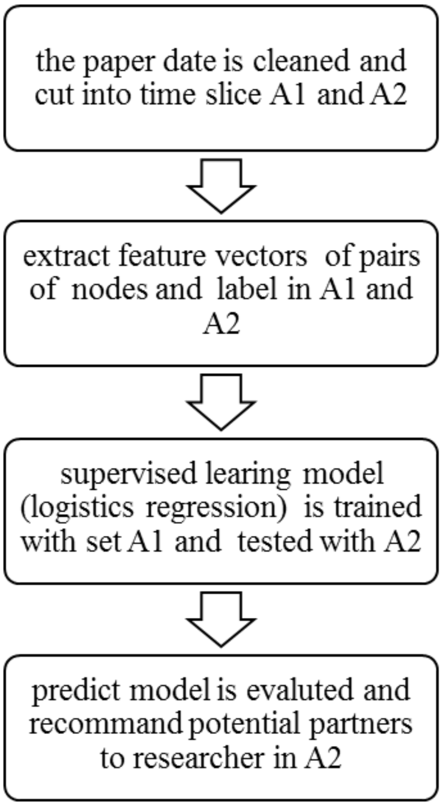
the paper date is cleaned and cut into time slice A1 and A2
extract feature vectors of pairs of nodes and label in A1 and A2
supervised learing model (logistics regression) is trained with set A1 and tested with A2
predict model is evaluted and recommand potential partners to researcher in A2

Fig.1 the process of construct link prediction model

III. Case study

(1) Data acquisition and preprocessing

We download 5980 papers of four journals from WOS data base, Scientometrics，Technology Forecasting and Social Change，Journal of the Association for Information Science，Journal of Informetrics. The type of document is article and time span is 2006-2017.

Data preprocessing is as follows, the tool is VantagePoint Software:

Divided the data in two slice, A1(2006-2012) and A2(2013-2017);

Removed the papers without a coauthor, the authors who have published less than 4,the documents which reference times less than 2 as the single document have no co-citation ,the keywords without co-occurrence and noisy ,like symbols, place name .

The details of data after cleaned are shown in TABLE1.

TABLE1 the detail of data A1&A2

	A1(training set)	A2(test set)
Papers	1688	3021
Researchers	145	343
Keywords	844	2139
References	4754	17399

(2) Extract feature vectors and training model

We generate two co-authorship networks according to data A1 and A2, the A1 network has 145 nodes with 343 edges and A2 has 343 nodes with659 nodes. 5 indices scores of every pair of node were computed after and the code based on Python3.6 is in Appendix. At last, there are 10400(C_{145}^{2}) samples in A1 and 58653(C_{342}^{2})samples in A2.

We use the Logistic Regression classifier in scikit-learn package (a Python module for machine learning) to train 10400 samples in A1. Threshold is set to 0.5,which means if the final score is greater than 0.5,the predict model class these pair of node as group that would have edge in the future. The prediction model was evaluated at last. Results is in TABLE2.and Fig2

TABLE2. The performance of predict model

	0	1
Support sample	57994	659
Precision	1	0.87
Recall	1	0.63
F1-score	1	0.73
AUC	0.82	

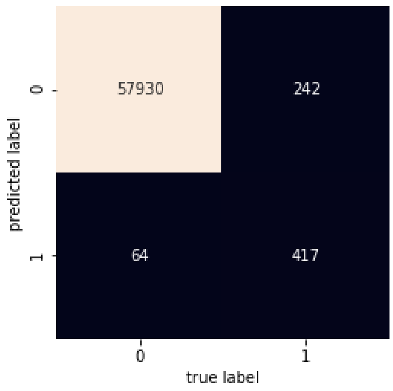

Fig 2 the performance of predict model

(3) Analysis

Through analyzing the network, we find out that the density value (0.011) of the A2 network is low and the network diameter (14) is large that means these 343 researchers are poor connectivity, so it is not conducive to the exchange and communication of knowledge. We list up top10 researchers according to betweenness centrality, these people with high score is important as they collaborators),which means they are active in scientific cooperation. And they also have published a great number of papers with diversified subjects. Keywords illustrate this point. For an instance, In terms of the keywords, Bornmann link everyone in the co-authorship network. For any researchers who what to find someone to cooperate with, they play a role as bridge. In TABLE3, most of them have many Collaborators (the Degree is one’s number of partners) Lutz have 79 keywords, like altmetrics，citation analysis，climate change,societal impact; Leydesdorff Loet have these keywords, like triple helix, classification, social network analysis, citation impact, algorithmic historiography. There are also some researchers have not many collaborators and their subjects is not diversified，like Carley Stephen，Rafols，Ismael，one of the cause is that they have worked with high-scoring Leydesdorff Loet and Porter Alan L. we can get this point in Fig 3.

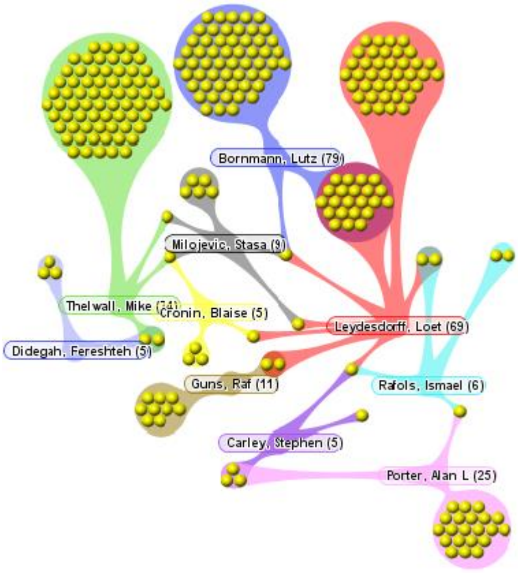

Fig 3.the co-authorship network of top10 score in betweennees centrality

Table 3. top 10 researchers in betweenness centrality score

Researcher	Betweenness centrality	Papers	Degree	Main keywords
Leydesdorff ,Loet	10959.76	69	31	triple helix,classfication
Thelwall, Mike	2994.39	74	17	citation analysis,Altmetrics,webometrics
Porter, Alan L	2704.89	25	16	text mining,tech mining,patent analysis
Carley, Stephen	2449.20	5	6	Interdisciplinary,bibliometrics
Milojevic, Stasa	2349.11	9	10	Aging,scholarly communication
Bornmann, Lutz	2295.69	79	20	Bibliometrics, citation analysis, Altmetrics
Cronin, Blaise	2255.01	5	9	scholarly communication,citation analysis
Didegah, Fereshteh	2159.01	5	6	inter-organization,collabration strategion
Guns, Raf	2081.08	11	8	Networks,collabration,link prediction
Rafols, Ismael	1972.92	5	6	science map ， topic model

Fig4.Fig5 visualize real and predictive giant component of networkA2(the threshold value=0.5, number of future link=64) and select several researchers and give the potential partner with highest possibility score. And we list two common keywords with high frequency, which represent their common fields or subjects to some extent.

TABLE 4. SOME RESEARCHERS IN network A2 and his or her potential partner

Researcher	Potential partners	Possibility Score	Common keywords
Bornmann, Lutz	Waltman, Ludo	0.99	Citation analysis，Citation impact
Leydesdorff, Loet	Porter, Alan L	0.99	Bibliometrics，Citation analysis，
Castillo, Javier	Bornmann, Lutz	0.89	Scientometrics, Catation impact
Wouters, Paul	Thelwall, Mike	0.78	Research evalution,Altmetrics
Gonzalez-Albo, Borja	Costas, Rodrigo	0.95	Networking centres,Gastroenterology,
Huang, Mu-Hsuan	Guan, Jiancheng	0.49	Patent, co-citation

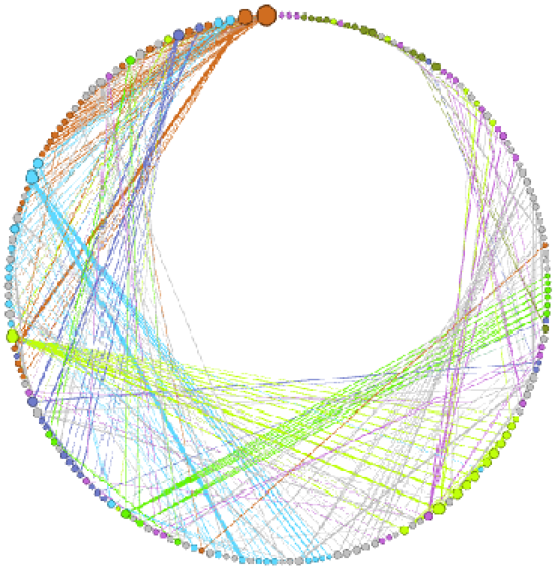

Fig4 the real giant component of network A2

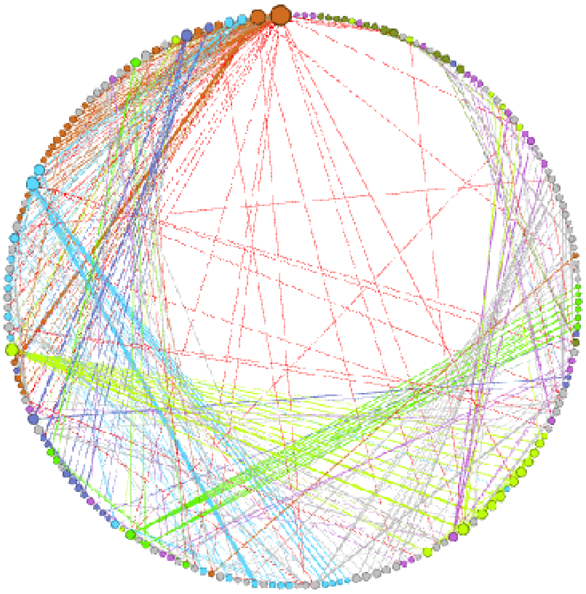

Fig5 the predictive giant component of network A2 in the future

.

IV. Conclusion

we established a predict model based on link prediction method and take the similarity of researchers in consideration. The model can give some advice when a researcher are finding right partners. Compared to link prediction method, we also can give recommend to the researcher who have not cooperated with others in the perspective of similarity. However, there are also some other factors which may influence the cooperation, such as geographical position, gender, institution. In the future study, we will try to improve the predict model in this direction.

References

[1] Pavlov, M., & Ichise, R. (2007). Finding experts by link prediction in co-authorship networks. International Conference on Finding Experts on the Web with Semantics (pp.42-55). CEUR-WS.org.

[2] Qiu, J. P., & Liu, G. H. (2014). Research on author collaboration based on social network analysis and keyword analysis: taking the field of knowledge management in domestic as an example. Information Science.

[3] Lv,L.Y. "link prediction in complex networks." Journal of University of Electronic Science and Technology of China 39.5(2010):651-661.

[4] Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. European Physical Journal B, 71(4), 623-630.

[5] Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. John Wiley & Sons, Inc

[6] Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. John Wiley & Sons, Inc

[7] Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: an author-, institution-, and country-level analysis. Journal of Informetrics, 8(2), 295-309

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

[9] Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the web. Social Networks, 25(3), 211-230.

[10] Zhao, J., Miao, L., Yang, J., Fang, H., Zhang, Q. M., & Nie, M., et al. (2015). Prediction of links and weights in networks by reliable routes. Scientific Reports, 5, 12261

[11] Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. Scientometrics, 101(2), 1461-1473..

[12] Rousseau, R., Delsaerdt, P., & Guns, R. (2012). Missing links: predicting interactions based on a multi-relational network structure with applications in informetrics. Dissertations & Theses - Gradworks. I. V. Wartburg, T. Teichert and K. Rost, "Inventive progress measured by multi-stage patent citation analysis," Research Policy, vol. 34, pp. 1591-1607, 2005.

Appendix

(1) the calculate code of 3 indices:CN AA RA

```
''''''''''''''''''''''''''''''

common_neibors
Adamic_Adar
resource_allocation
''''''''''''''''''''''''''''''

import networkx as nx
import numpy as np
import math
graph=nx.read_pajek('coauthorshipnetwork.net')
adjmartrix = nx.to_numpy_matrix(b)
a=np.array(adjmartrix)
for i in range(0,len(a)-1):
    for j in range(i+1,len(a)):
        if a[i][j]!=0 and a[j][i]!=0:
            label=1
        else:
            label=0
        score1=0 #common_neibors
        score2=0 #Adamic_Adar
        score3=0 #resource_allocation
        for x in range(len(a)):
            if a[i][x]!=0 and a[j][x]!=0:
                s=0
                for z in range(len(a)):
                    s = a[z][x] + s
                else：
                    s=0
                    score1 = score1 + a[i][x] + a[j][x]
                     score2 = score2 + (a[i][x] + a[j][x])/math.log10(1+s)
                     score3 = score3+ (a[i][x] + a[j][x])/s
             print(score1,score2,score3,label)
```

(2) the code of train and evaluate LogisticRegression model

```
'''''''''''''''''''''

LogisticRegression
'''''''''''''''''''''

import numpy as np
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report
from sklearn.metrics import roc_curve
from sklearn import metrics
import matplotlib.pyplot as plt
import seaborn as sns
dataA1 = np.loadtxt('traindata.txt', dtype=float, delimiter=',')
dataA2 = np.loadtxt('testdata.txt', dtype=float, delimiter=',')
a=np.array(dataA1)
b=np.array(dataA2)
X_train = a[:,[1,2,3,4,5]]
y_train = a[:,6]
X_test = b[:,[1,2,3,4,5]]
y_test = b[:,6]
sc = StandardScaler()
sc.fit(X_train)
X_train = sc.transform(X_train)
X_test = sc.transform(X_test)
LRModel = LogisticRegression(C=1000.0, random_state=0)
LRModel.fit(X_train, y_train)
predictions = LRModel.predict(X_test)
print(classification_report(y_test,predictions))
test_auc = metrics.roc_auc_score(y_test,predictions)
print('AUC value',test_auc)
```