Collaboration prediction based on the Multi-node2vec algorithm by embedding tripartite citation network: A Case Study of Gene Editing

Feifei Wang
College of Economics and Management, Beijing University of Technology, Beijing,China
feifeiwang@bjut.edu.cn

Wanzhao Lu
College of Economics and Management, Beijing University of Technology, Beijing,China
luwanzhao@qq.com

ABSTRACT

The prediction of collaboration for many years was mainly based on the co-author network. However, there were few studies on the multiple path relationships among authors. Citation association is an effective mechanism to reveal the similarity between authors. The correlation between scholars based on citation relations can be further deepened. Therefore, an attempt to embed all-author tripartite citation networks can complement the author's collaborative recommendation. In this paper, we use the multi-node2vec method to calculate the similarity between authors through all-author tripartite citation networks in gene editing field. The result shows that our method is superior to the traditional indicators in co-author network and single-layer node2vec in all-author tripartite citation networks.

KEYWORDS

Link prediction, Citation analysis, Multi-node2vec, all-author tripartite citation networks

1 Introduction

In recent years, link prediction in complex social networks has attracted much attention. Link prediction in social network is to predict the possibility of future links between unbounded pairs of nodes in the network based on the existing network nodes and the topological characteristics of the network. The prediction of cooperative relationship mainly focuses on the similarity or correlation between nodes and is applied to the coauthor network. The index of correlation between the author combinations includes the common neighbor and its improvement index, the random walk index, the arrival path index, and so on. In addition, the network embedding method is used to learn the low-dimensional potential representation of nodes in the network and convert nodes into vector representations. It is also a method to study the similarity between nodes. It is an important premise of cooperative prediction that network research on multiple correlation indicators can be applied, the greater the correlation between two nodes, the greater the possibility of cooperative relationship between them[1].

Merely mining the information in the co-author network can no longer satisfy the research desire of scholars. Popescul and Ungar [2] proposed a link prediction method based on article information, author information and journal information by using citation network and logistic regression model. Naoki and Yuya[3] used machine learning method to predict missing citation links in citation network based on link prediction index in citation network.

Citation analysis is an important research content in the field of scientometrics and information metrology. It is widely used in the measurement of subject value, research frontier detection, knowledge flow detection and analysis, and scientific structure. Wang and her team[4][5] proposed all-author tripartite citation networks by combining co-citation and coupling relations as a way of weighting the strength of direct citations, proved that this new kinds of citation networks are able to help identify the most influential scholars in the field of gene editing.

This paper will predict co-authorship based on all-author tripartite citation networks by calculating the similarity of two vectors from multi-node2vec algorithm which is a new method to converts nodes to vectors by random walks on multilayer networks. Aimed to prove that all-author tripartite citation networks are also effective in collaboration prediction.

2 Data and methodology

2.1 Data

The field of gene editing in the past decade, especially since the introduction of CRISPR/Cas9 technology, it has become a worldwide hot point. CRISPR/Cas9 is a novel technology for genome directional editing. The technology has aroused extensive attention in the fields of biomedicine, genetics, and cytology. Therefore, this paper takes gene editing as the research object. This paper use a set of downloaded records retrieved from the WoS in the May 7, 2019. The search equation is (TS = (“gene edit*”) OR TS=(CRISPR) OR TS= (“clustered regularly interspaced short palindromic repeats”), including all document types, and Timespan =All years).

In this paper, the method of name disambiguation proposed by Caron was used to obtain a total of 74250 authors by name disambiguation of all the cited authors.By using the method of name disambiguation proposed by Caron and Van[6], we get 74250 authors. According to Price's law, 932 authors who published greater than or equal to 7 papers are chosen as core authors in gene editing field. This paper will focus on the papers and citations of these 932 authors.

2.2 Methodology

2.2.1Name disambiguation. Because there may be duplicate names among authors, and there are some writing irregularities, for example, sometimes the middle name are omitted, or the first name are abbreviated, etc. Therefore, in order to reduce the interference of the experimental results due to the problem of name ambiguity, we use the method proposed by Caron and Van[6] for name disambiguation.

The author disambiguation method for large bibliographic databases mainly uses rule-based scoring and clustering. First, we combine papers and references, then process the name into the name-block which are constructed based on the last name and first initial and the removal of all non-alphabetic characters. According to the author information of each article, we grade the author's name-block, and the scoring rules are shown in Table 1. We rate two authors with the same name-block, and give them scores according to the conditions in Table 1. Then we use single linkage clustering to get the final disambiguation result.

Table 1 Scoring criteria

Rule	Criterion	score
1	same Email	100
2	same first name	20
3	same organization	20
4	shared co-authors (one/two/three/more than three)	5/10/20/40

2.2.2 ANWDC. All-author tripartite citation networks use an indicator called ANWDC which is a normalized weighted direct citation indicator between author. The formula of *ANWDC*[5] between author A and author B is as follows:

$$ANWDC_{AB} = \sum_{r \in R} w^{A_r} * w^{B_r} * NWDC_{A_r B_r}$$

$$NWDC_{A_r B_r} = DC_{A_r B_r} + NC_{A_r B_r} + NBC_{A_r B_r}$$

$$R = \{r \mid A_i \to B_i, paperA_i \in A, paperB_i \in B\}$$

$$DC_{A_r B_r} = \begin{cases} 1, \text{ if } A_r \text{ cites } B_r \\ 0, \text{ otherwise} \end{cases}$$

$$NC_{A_r B_r} = \begin{cases} \sum_1^M \frac{1}{m_{C_r}}, \text{if } A_r \text{ and } B_r \text{ are cited by same } C_r \\ 0, \text{otherwise} \end{cases}$$

$$NBC_{A_r B_r} = \begin{cases} \sum_1^N \frac{1}{n_{D_r}}, \text{if } A_r \text{ and } B_r \text{ cite same } D_r \\ 0, \text{otherwise} \end{cases}$$

Different from Feifei Wang's method[5], here we take another method to calculate the author's contribution. Because in the field of gene editing, the person in charge of the laboratory is usually the final author, and other laboratory members or other participants are the middle collaborators. So we define the formula of contribution degree as follows：

$$w^{ith} = \begin{cases} \frac{1 - \sum_{i=2}^{n-1} \frac{1/i}{1 + 1/2 + \ldots + 1/N}}{2}, i = 1 \text{ or } i = n \\ \frac{1/i}{1 + 1/2 + \ldots + 1/n}, 1 < i < n \end{cases}$$

Where author is the *ith* author in paper, *n* is the total number of author in paper.

2.2.3 The Multi-node2vec Algorithm. In fact, all-author tripartite citation network is combination of direct citation network, co-citation network and coupling network. Multilayer networks do better in analysis relationships within and between networks. Multi-node2vec algorithm[7] is a scalable method used to generate node vectors in Multilayer networks.

For the given multilayer networks G_N^m , Using the NeighborhoodSearch procedure can identify a collection of neighborhoods of length for G_N^m through second order random walks on the network. Then, use stochastic gradient descent on the two-layer Skip-gram neural network model of context size to optimize the log-likelihood and estimates F through maximization of the log likelihood function in the follow formula[7]:

$$L(F \mid G_N^m) = \sum_{u \in N} \sum_{v \in Ne(u)} [f_v^T f_u - \log(Z_u)]$$

Where $Z_u = \sum_{w \in N} \exp\{f_w^T f_u\}$ is a normalization constant for the node node $u, u \in N$.

After get the vectors from multi-node2vec algorithm, compute the cosine similarity between the vectors. The greater the similarity, the greater the likelihood of cooperation.

3 Results and discussion

The data set contains 4525 papers and its references of 932 authors. In order to evaluate the accuracy of the method, we choose *CN* indicator, *AA* indicator, *RA* indicator, and *PA* indicator in the co-author network to contrast accuracy with multi-node2vec algorithm in all-author tripartite citation network. At the same time, we also compare the multi-node2vec algorithm with traditional node2vec algorithm in the co-author network and

all-author tripartite citation network(*ANWDC*). We select data from 2015 and before(T1) to predict new collaborations and compare it with real collaborations from 2016 to May 7, 2019(T2) to calculate AUC indicator[8].The results are shown in table 2.

Table2 AUC accuracy comparison between different indicators

Indicator	***CN***	***AA***	***RA***	***PA***
AUC	0.5830	0.6317	0.6974	0.5948
Indicator	***Node2vev (co-authors network)***	***Node2vev (ANWDC)***	***Multi-Node2vec (ANWDC)***	
AUC	0.6704	0.8014	0.8904	

It should be noted that when we calculate the accuracy of *AUC*, we use the real data of T2 period for cooperation. In fact, not every author among the 932 authors had cooperation and mutual reference data in the T1 period, so in node2vec algorithm and multi node2vec algorithm, not every author can learn feature vector. Therefore, when calculating the *AUC* accuracy, the number of input author-pairs is not the same.

The *ANWDC* is an indicator to describe the citation relevance between authors. The authors are connected with each other through the citation relationship of their published papers. The higher the citation relevance is, the higher the *ANWDC* value will be. Table 3 shows the top 10 author-pairs with the largest *ANWDC* values and its value in *DC*, *NC* and *NBC* layer in T1. It need to be noted that the *ANWDC* and *DC* are directional indicators and *NC*, *NBC* are not. We use arrows to indicate the direction of reference. In the same way, we calculate the *ANWDC* for all time without time slicing. The top 10 author-pairs with the largest *ANWDC* values are presented in Table 4.

Table 3 Top 10 author-pairs with largest *ANWDC* values at T1 period

Rank	**AuthorPairs**	***ANWDC***	***DC***	***NC***	***NBC***
1	EUGENE V KOONIN -> KIRA S MAKAROVA	37.1900	17.0740	4.9006	15.2154
2	RODOLPHE BARRANGOU -> LUCIANO MARRAFFINI	36.2089	14.4572	9.6129	12.1388
3	KIRA S MAKAROVA -> EUGENE V KOONIN	34.0023	13.8863	4.9006	15.2154
4	LUCIANO MARRAFFINI -> RODOLPHE BARRANGOU	31.9620	10.2102	9.6129	12.1388
5	EUGENE V KOONIN -> YURI I WOLF	28.1455	12.6677	2.4674	13.0104
6	YURI I WOLF -> EUGENE V KOONIN	27.5301	12.0524	2.4674	13.0104
7	RODOLPHE BARRANGOU -> PHILIPPE HORVATH	26.5439	15.6073	7.3799	3.5568
8	RODOLPHE BARRANGOU -> JENNIFER DOUDNA	24.5461	10.6900	7.5624	6.2938
9	LUCIANO MARRAFFINI -> EJ SONTHEIMER	20.8982	10.3582	4.4304	6.1096
10	EUGENE V KOONIN -> RODOLPHE BARRANGOU	20.2471	7.8535	5.8715	6.5220

Table 4 Top 10 author-pairs with largest *ANWDC* values at all time

Rank	AuthorPairs	ANWDC	DC	NC	NBC
1	EUGENE V KOONIN -> KIRA S MAKAROVA	99.3334	48.4081	14.8570	36.0683
2	KIRA S MAKAROVA -> EUGENE V KOONIN	83.0886	32.1633	14.8570	36.0683
3	EUGENE V KOONIN -> YURI I WOLF	68.4181	31.6652	7.3612	29.3916
4	YURI I WOLF -> EUGENE V KOONIN	58.0313	21.2785	7.3612	29.3916
5	RODOLPHE BARRANGOU -> LUCIANO MARRAFFINI	55.3964	24.3257	18.7510	12.3197
6	EUGENE V KOONIN -> MART KRUPOVIC	54.5128	23.4630	5.4299	25.6200
7	MART KRUPOVIC -> EUGENE V KOONIN	51.5498	20.4999	5.4299	25.6200
8	RODOLPHE BARRANGOU -> JENNIFER DOUDNA	46.8326	19.6187	17.4277	9.7862
9	LUCIANO MARRAFFINI -> RODOLPHE BARRANGOU	45.2532	14.1825	18.7510	12.3197
10	JENNIFER DOUDNA -> LUCIANO MARRAFFINI	44.3099	17.3289	19.1923	7.7887

The co-citation network and coupling network are symmetric matrices. After running the multi-node2vec algorithm in T1's networks and calculating the cosine similarity of vectors, we choose the top 10 author-pairs of vector's similarities. Then we find these author-pairs all collaborated in T2 period. Table 5 shows the detail of T1's link prediction.

As we can tell in the result of prediction in T1, the *ANWDC* indicator is effective in link prediction. We input the all-time citation information of the 932 authors to predict the collaboration and present the top 10 author-pairs with largest similarity in Table 6.

In this paper, the cooperation between authors is judged according to the vector similarity between authors. The greater the similarity means that the more similar the authors are in the all-author tripartite citation network, the greater the possibility of cooperation.

Table 5 Top 10 author-pairs with greater similarity in prediction at T1 period

Rank	AuthorPairs	Cos (vectors)	Counts (papers)
1	FRIEDRICH FAUSER HOLGER PUCHTA	0.9760	2
2	BO HUANG BAOHUI CHEN	0.9744	4
3	OMAR ABUDAYYEH JULIA JOUNG	0.9719	8
4	TIM WANG DAVID SABATINI	0.9717	5
5	WEILI YANG ZHUCHI TU	0.9716	5
6	CHRISTINE POURCEL GILLES VERGNAUD	0.9712	1
7	FLORIAN SCHMIDT DIRK GRIMM	0.9711	2
8	REBECCA M TERNS MICHAEL P TERNS	0.9698	4
9	JINXING LIU JINLONG QIU	0.9696	2
10	WEI CHEN XU ZHANG	0.9694	3

Table 6 Top 10 author-pairs with greater similarity in prediction at all time

Rank	AuthorPairs	Cos (vectors)
1	JOFFREY MIANNE LYDIA TEBOUL	0.9749
2	ZHAOMING LIU NANA FAN	0.9700
3	HAO LI HONGBIN SONG	0.9686
4	WATARU FUJII KUNIHIKO NAITO	0.9683
5	GEMMA CODNER LYDIA TEBOUL	0.9683
6	ZHAOMING LIU ZHEN OUYANG	0.9678
7	SEVERINE MENORET IGNACIO ANEGON	0.9674
8	YANDI GAO LI XU	0.9673
9	JORGE MANSILLASOTO MICHEL SADELAIN	0.9670
10	HAO LI JING XIE	0.9666

4 Conclusions

In this paper, the all-author tripartite citation network is included in the discovery of potential co-authorship in the field of gene editing. Divide the all-author tripartite citation network into a three layers networks, realizes the predicting collaboration between authors by using multi-node2vec algorithm.

Research has found that all-author tripartite citation network can be used in link prediction. By calculating the *AUC* accuracy of the predicted results with real data, we found that the node2vec method in citation network was generally superior to the *CN, AA, PA, RA* indicators in traditional co-author network. And the multi-node2vec method do better than single-layer node2vec method. But this study is based only on the field of gene editing. The future can be looked forward to in a variety of fields.

There is one thing to note. In co-author network, the prediction is limited to new links. But for the prediction in citation network, whether or not cooperation has occurred in the past, only predict links in the future.

REFERENCES

[1] Wang, P. , Xu, B. W. , Wu, Y. R. , & Zhou, X. Y. . (2014). Link prediction in social networks: the state-of-the-art. ece China. Information ences, 58(1), 11101-011101.

[2] Popescul, A., Ungar, L. H., Lawrence, S., & Pennock, D. M. (2003). Statistical relational learning for document mining. international conference on data mining.

[3] Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. Journal of the Association for Information Science and Technology, 63(1), 78-85.

[4] Wang, F., Wang, X., & Yang, S. (2017). Mining author relationship in scholarly networks based on tripartite citation analysis.. PLOS ONE, 12(11).

[5] Wang, F., Jia, C., Wang, X., Liu, J., Xu, S., Liu, Y., & Yang, C. (2019). Exploring all-author tripartite citation networks: A case study of gene editing. Journal of Informetrics, 13(3), 856-873.

[6] Caron, E., & van Eck, N-J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), Proceedings of the Science and Technology Indicators Conference 2014 (pp. 79-86). Universiteit Leiden.

[7] Wilson, J. D., Baybay, M., Sankar, R., & Stillman, P. E. (2018). Fast embedding of multilayer networks: An algorithm and application to group fMRI.. arXiv: Social and Information Networks.

[8] Lv, L.(2010).Link prediction on complex networks, Journal of University of Electronic Science and Technology of China, 039(005), 651-661.