Research on Heterogeneous Enhanced Network Embedding for Collaboration Prediction

Xin Zhang[1][0000-0001-8784-3788] ,Yi Wen[1][0000-0002-6520-2733] and Haiyun Xu[2][0000-0002-7453-3331]

[1] Chengdu Library and Information Center,CAS, Chengdu 610041, Sichuan, China
[2] Business School, Shandong University of Technology ,Zibo 255049,Shandong, China
[1]zhangxin@clas.ac.cn

Abstract. Scientific research collaboration has always been an important research content of information science, and collaboration prediction is an important issue in personalized information services. This article constructs an author-centered heterogeneous information fusion schema, and uses causal analysis to quantitatively study the influence of same institutions, co-word and citation on collaboration, compares the effects of different network embedding algorithms in collaboration prediction, and built heterogeneous information fused network embedding model for collaboration prediction. Take the field of stem cells as an empirical case, experiments show that the matrix decomposition based network embedding algorithms(like NetMF) are balance of performance and accuracy. Institutions, keywords and citations can improve the prediction effect of collaboration, and (same institutions > citations > co -words) among them. Multiple features fusion models are generally better than single information fusion, and the model of (collaboration + same institution + citations) performs outstanding in collaboration prediction.

Keywords: collaboration Prediction , Network embedding, Heterogeneous information.

1 Introduction

Since the birth of information science, scientific research collaboration has become an important research topic. The prediction of collaboration has important theoretical and practical significance for the analysis of S&T trends and the recommendation of personalized service information. Collaboration prediction is also a very challenging task. Scholars in different fields have invested in this topic. Most computer scientists design more and more advanced and complex network representation learning algorithms, and find ways to incorporate different types of information such as text into representation learning. Intelligence experts pay more attention to the application of these algorithms for collaborative recommendation

Newman[1] was the first to introduce the network analysis into research collaboration. Liben-Nowell and Kleinberg [2] put forward the problem of link prediction in social networks, they gave some similarity measurements based on

network structure, and applied two series of index methods based on nodes and paths, empirical analysis is carried out in the collaboration network of authors in five fields of physics. Since then, quit a lot of work focused on improve the indicators in this article. In the book "Link Prediction", Lv Linyuan introduced a series of similarity-based link prediction indicators, such as common neighbor index (CN), cosine similarity, Jaccard Indicators, Adamic-Adar (AA) indicators, resource allocation (RA) indicators, LHN-I indicators and others based on the local information of nodes, as well as local path indicators (LP), Katz indicators, restart random walks, etc. based on the information of edges or paths[3]. Intelligence experts has conducted a series of empirical studies based on these indicators, such as R Guns and R Rousseau [4] ,they constructed a weighted collaboration network in the field of malaria and tuberculosis to carry out cooperative prediction and recommendation work. Yan E et al. [5] used the papers of 59 journals in the field of library and information as an empirical case to compare the prediction results under various indicators. Shan Songyan et al. [6] summarized and reviewed the author similarity algorithm for cooperative prediction. In addition, cooperative prediction also has methods based on maximum likelihood estimation, methods based on probability graph models, etc. The latter two approaches are based on statistical ideas and have achieved certain effects on specific networks. The computational complexity of those method is too high, and it is difficult to implement on large-scale networks.

In response to the large-scale sparse network prediction problems encountered in actual research, researchers continue to propose new methods to learn a low-dimensional dense representation of the network. Then use low-dimensional representation for structural prediction work. For example, Zhang Jinzhu et al[7] introduced the method of network representation learning into the collaboration prediction. In this paper, the LINE model is used to construct the vector representation of author, and the cosine similarity is used to measure the possibility of collaboration. Yu Chuanming et al[8] studied the application of network representation learning methods in collaborative recommendation, proposed an integrated recommendation model, and conducted empirical analysis in the financial field.

Moreover, it must be noted that collaboration is affected by many complicated factors. In addition to the network structure, collaboration prediction should also use rich heterogeneous structure information. Wang Zhibing et al. [9] merged the attribute characteristics such as the node mechanism into the similarity index of the network structure, and carried out collaboration prediction. Liu Ping et al. [10] constructed an LDA-based author interest model based on the community division, and then analyzed the author’s relevant literature to achieve the purpose of scientific research collaboration recommendation; Yu Chuanming et al. [11] constructed the collaboration network in the financial field and adopted a link prediction method based on feature fusion of individuals, institutions and regions. Lin Yuan et al. [12] combined the heterogeneous information of scholars, institutions, and keywords to construct a scientific research collaboration network, and then used node2vec to

express learning methods for cooperative prediction. Zhang Xin et al. [13] proposed a scientific research collaboration prediction method that combines network representation learning and author's topic characteristics .On the basis of these studies, this paper constructs a cooperative prediction method of feature fusion.

2 Ideas and Methods

2.1 Research Framework

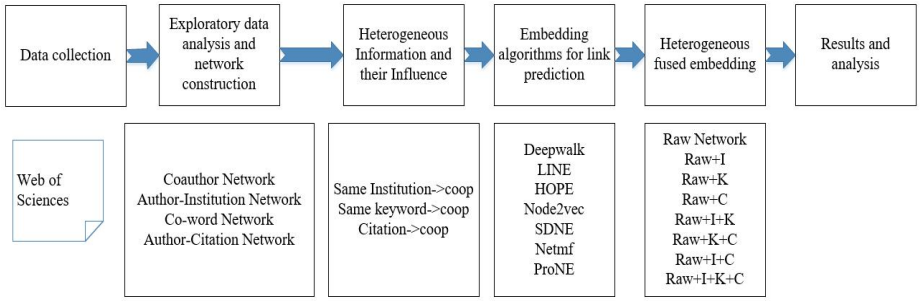


Fig. 1. Research Framework of this paper

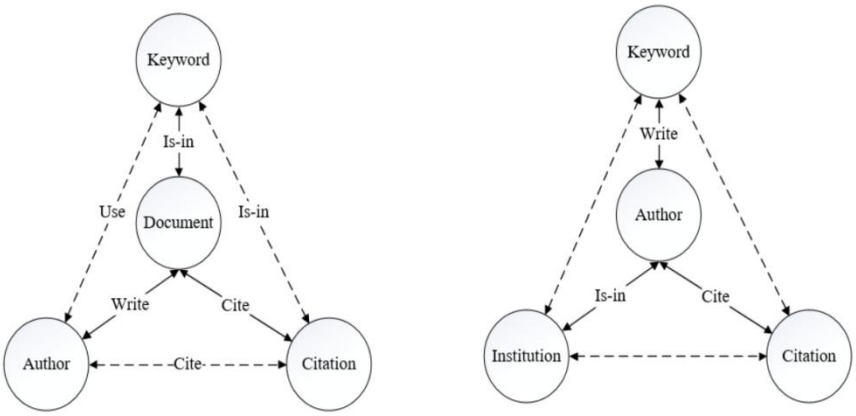
The research framework of this paper is shown in Figure 1. It can be roughly divided into 6 stages, namely (1) data collection -> (2) exploratory data analysis and network construction -> (3) the impact of heterogeneous information on scientific research collaboration -> (4) embedding algorithms evaluation for link prediction -> (5)
The construction of heterogeneous features fused network embedding approaches and (6) Results and analysis. Retrieve the documents to be analyzed from the Web of Sciences database, and then conduct exploratory analysis on the collaboration relationship among them, and construct a collaboration network, author-institution network, author-keyword network, author-citation network. Later, for the same institutions, shared keywords and citation relationships, the framework of causal analysis was used to study their impact on collaboration. Then, we discuss the performance and efficiency of DeepWalk, LINE, HOPE, Node2vec, SDNE, NetMF and ProNE and other models for collaboration predictions. Finally, we integrate the Institution(I), Keyword(K) and Citation(C) features to network embedding models and use these schema to carry out actual network prediction task.

2.2 Research on Heterogeneous Information and Its Impact on collaboration

Heterogeneous information fusion is abbreviated as data fusion, which refers to the comprehensive analysis of different types of information sources or relational data through a specific method, and finally all the information can be used to jointly reveal the characteristics of the research object, making up for the single data type and the single relation type to reveal the research field insufficient associations between

entities to obtain more comprehensive and objective measurement results (Xu Haiyun et al. [14]).

Hua Berlin [15] put forward the fusion theory as one of the main methodologies of information science , and emphasized the important of data fusion, information fusion and knowledge fusion in information science. Later, he further discussed the influence of data fusion on intelligence. The importance of fusion of different types of multiple source information is analyzed. He also systematically explained the relevant theories and applications of information fusion from the perspective of information fusion's representation process, technical algorithms and models. Morris et al. [16] gave an overview of common measurement entities in scientific and technological literature, mainly including documents themselves, references, journals where documents are located, authors of documents, journals where references are located, authors of reference documents, subject headings etc. Xu Haiyun et al. [14] reviewed the multiple source data fusion methods in Scientometrics. The document-centered data fusion meta-path model given in the paper is shown in Figure 2(a). This article takes the author's collaboration as the research object, and studies the influence of the same institution, the same keywords, and the citation relationship on scientific research collaboration. On the basis of Figure 2(a), an author-centered scientific research collaboration network is constructed, as shown in Figure 2(b).



(a) Document-centered schema (b) Author-centered schema

Fig. 2. Heterogeneous Information Fusion Schema

Table. 1. Heterogeneous Information Network Construction

Information	Networks directly extract from documents	Uniform Network
collaboration	collaboration Network	collaboration Network
Institution	Author-Institution Bipartite Network	Same Institution Network
Keyword	Author-Document Bipartite Network Document-Keyword Bipartite Network	Author Co-word Network

Citation	Author-Document Bipartite Network	Author Citation Network
	Document Citation Network	

Table 1 shows several methods for constructing heterogeneous network information. The first column represents the fused heterogeneous information, the second column represents the feature network that can be directly extracted from the article, and the third column represents the mapping of the second type of network to the author dimension The formed network.

(1) *Author collaboration network.* The author collaboration network can be extracted directly from the author field of the article. The nodes in the network represent authors, the edges in the network represent collaboration relationships, and the weight of the edges represents the frequency of collaboration.

(2) *The same institution network*, extract the author institution bipartite network from the author-institution (C1) field of the article. An institution can contain multiple authors, and at the same time, an author can work part-time in multiple institutions, which can be mapped into Authors are in the same organization network. The nodes in the network represent authors. If two authors have the same organization to which they belong, the network is connected.

(3) *The author co-word network,* according to the downloaded documents, we can construct an article-author network, article-keyword network, and then form an author-keyword bipartite network. The weight d(a,k) in the network indicates that the the frequency of author a uses keywords k (number of articles), $\Gamma(a)$ represents the neighbor node of node a, that is, the keyword used by a. According to this network, we can deduce the author co-word network. The nodes in this network represent the author. The edge weight w(a,b) between author a and b is calculated by formula (1), D represents the total number of documents, and d(k) represents the frequency of keyword k.

$$\mathrm{w}(a,b)=\sum_{k\in\Gamma(a)\cap\Gamma(\mathrm{b})}\min(d(a,k),d(b,k))\times\log\frac{D}{d(k)} \tag{1}$$

(4) *Author citation network,* based on the downloaded documents, an author-article bipartite network can be constructed (the article written by author a in this network is denoted as $\Gamma(a)$), an article citation network, and then an author's citation network is formed. The weight c(a,b) in this network is the number of articles of author b cited in all articles of author a.

$$c(a,b)=\sum_{d\in\Gamma(a)}\sum_{c\in\Gamma(b)}\delta(d,c) \tag{2}$$

if document d cite document c， $\delta(d,c)=1$， Otherwise $\delta(d,c)=0$ .Unlike the previous two networks, the author cited network is a directed network.

In this way, we have constructed several heterogeneous information networks. However, the influence of different heterogeneous information on collaboration may be different. In order to measure the difference of this influence, we adopt the paradigm of causal reasoning. ACE (Average Casual Effect) is used to measure the

influence of treatment variable Y (same institution, co-word, citation) to variable X (collaboration).

$$\mathrm{ACE}(Y->X)=E(X\mid Y)-E(X\mid\sim Y) \tag{3}$$

The larger the ACE, the more obvious the causal effect of Y on X.

2.3 Network Embedding Algorithms for Cooperation Prediction and Evaluation Criteria

Network embedding is currently a hot method in the field of cooperative prediction. Scholars continue to incorporate new ideas and methods into network representation learning, and the accuracy and efficiency of the algorithm continue to improve. This article compares seven well-known and commonly used network representation learning algorithms and evaluates their accuracy and performance in scientific research collaboration prediction.

Given a network G(V,E), V is the set of vertices, E is the set of edges, |V|=N,|E|=M, the adjacency matrix of the graph is W.

(1) DeepWalk, influenced by the well-known word2vec, Perozzi et al. proposed the DeepWalk [17] in KDD2014. The model uses random walks to generate vertex sequences, and each node sequence is analogous to a sentence in the language. Each node pair is analogous to a word, and the Skip-gram method is used for learning and training, and the vector representation of the node is obtained.

(2) LINE (Tang Jian et al., 2015) [18] defines the first-order similarity relationship (1st order proximity) and the second-order proximity relationship (2nd order proximity) in the graph in the algorithm. The distance is closer. The algorithm defines first-order and second-order optimization goals to characterize this relationship.

(3) Node2vec is a work published on KDD2016 by Grover A et al. [19]. They improved the random walk strategy in the DeepWalk algorithm, and also considered the direct neighbor relationship (homophily) and structural similarity relationship of the node. Using depth-first and breadth-first node traversal, a random walk method is designed. Assuming a random walk sequence from node u to node v, then for each v's neighbor node w, the next step is selected according to probability sampling with parameters p and q.

(4) SDNE (Wang D et al., 2016) [20], also published on KDD2016, SDNE model can be seen as an extension of LINE model, the essence of the method is a Graph AutoEncoder, making graph representation The reconstruction loss of is as small as possible, and the vector representation of nodes with connected edges is as close as possible.

(5) The HOPE [21] method depicts two different representations for each node, and focuses on preserving the asymmetry information in the original network. Different asymmetric relationship matrices are constructed, and then the JDGSVD algorithm is used for matrix reduction. Dimension gets the network representation of the node.

(6) NetMF, Tang J team unified deepwalk, line, node2vec and other algorithms into the framework of matrix decomposition, and proposed a NetMf representation

learning method that directly decomposes the target network [22]. After that, they performed the algorithm again. Improved, introduced sparse matrix factorization, and proposed NetSMF [23].

(7) ProNE, Jie Zhang from the team of Professor Tang Jie of Tsinghua University proposed a fast and scalable large-scale network representation learning algorithm ProNE[24] at IJCAI 2019. The algorithm is divided into two steps. 1)Sparse matrix factorization for fast embedding initialization and 2) Spectral propagation in the modulated networks for embedding enhancement. Compared with the classical random walk method, this method has an efficiency improvement of tens to hundreds of times.

The network embedding approaches and graph neural networks are moving from theory to application. In 2018, the Alimama team open sourced Euler, a distributed graph deep learning tool, DeepMind open sourced the graph_nets graph network tool, New York University researchers open sourced the graph neural network learning framework DGL, and in 2019 Facebook open sourced PyTorch-based graph representation learning and neural The network framework PyTorch-BigGraph [25]. In addition, recent graph representation learning tools include Tsinghua University's data mining team's CogDL[26] and Tang Jian's graph representation learning system GraphVite[27]. The emergence of these platform tools has lowered the application threshold of graph representation learning, allowing graph representation learning to be applied in a wider range.

In this paper, we use those network embedding for collaboration prediction. The collaboration prediction problem is transformed into a binary classification problem with/without links. According to the specific ratio, the edge set is divided into training set and test set respectively, then taking the training set as the baseline, we take the edges in the test set as positive samples, randomly generate the same number of negative samples to evaluate the model. We use the AUC, ROC_AUC and F1_Score indicators to evaluate the accuracy of the model. The AUC indicator is the most commonly used indicator to evaluate the link prediction problem. Provided the rank of all non-observed links, the AUC value can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link [28]. The AUC value ranges from 0.5 to 1. The AUC value of random allocation method is 0.5, and the AUC value of perfect prediction is 1. The closer the AUC value is to 1, the better the model effect is, and the ROC_AUC is defined as the area under the ROC curve in the binary classification problem, while the F1_Score is the balance of precision and recall in the binary classification problem. Moreover, we use execution time to measure the efficiency of the model.

2.4 Multiple Feature Fused collaboration Prediction Model

The method of fusing multiple features is shown in Figure 3. It extracts the basic collaboration relationship, the same organization relationship, the co-word relationship of articles, and the citation relationship between authors from the document collection, and constructs the collaboration network, the same organization network, the co-word network, and the citation network.

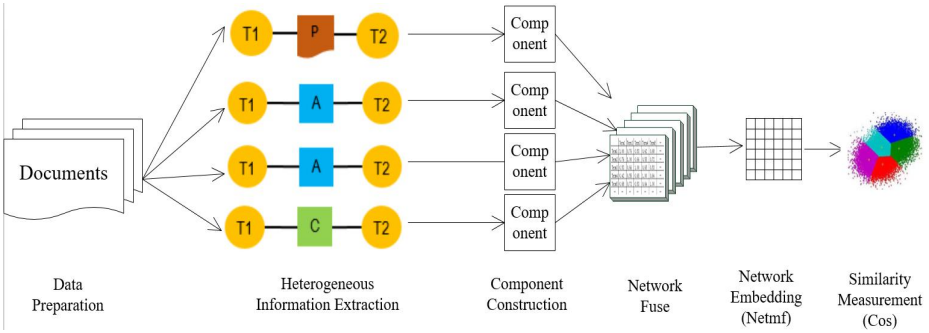


Fig. 3. Algorithm Flow

In this part, this article discusses the influence of several different feature combination methods on the prediction results of scientific research collaboration. The fusion features in the article can be divided into two types: single feature fusion and multiple feature fusion methods. Feature fusion is the result of fusing multiple sets of features, based on the analysis of the experimental results in the previous part. The same institution, the same keywords, and the citation relationship all have a positive relationship to collaboration. This section discusses the influence of several characteristics and their combinations on the performance of scientific research collaboration predictions. In this section, we also use the AUC, ROC_AUC and F1_Score indicators to evaluate the model.

Table. 2. Various feature fusion methods

category	abbr.	description
Baseline	Raw	Raw Network
Single feature	Raw+I(ri)	Raw Network+Same Institution Network
	Raw+K(rk)	Raw Network+Author Coword Network
	Raw+C(rc)	Raw Network+Author Citation Network
Multiple feature	Raw+I+K(rik)	Raw Network+ Same Institution Network+ Author Citation Network
	Raw+K+C(rkc)	Raw Network+ Author Coword Network+ Author Citation Network
	Raw+I+C(ric)	Raw Network+ Same Institution Network+ Author Citation Network
	Raw+I+K+C(rikc)	Raw Network+ Same Institution Network+ Author Coword Network+ Author Citation Network

3 Experiments

3.1 Field selection and exploratory analysis

Research field selection

Stem cell and regenerative medicine research has brought revolutionary changes to the treatment of cancer and other diseases. It has been selected as one of the top ten

scientific and technological advances in the US "Science" magazine for 9 times. The project has also laid out related projects many times, so the author selects the field of stem cells for empirical research. Search in ISI Web of Knowledge with the search formula (TI=Stem Cells). The search time was May 2019. The search yielded 433 469 articles. The number of articles in different years in the search results is shown in Figure 4, which shows that the number of articles is presented The trend of slow growth to rapid growth and then to saturation growth may have declined in 2019 due to incomplete data collection.

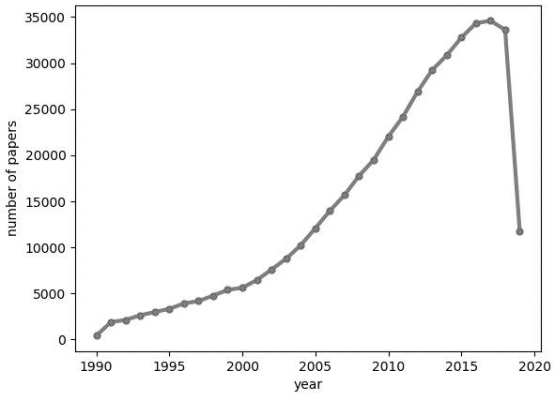


Fig. 4. Number of articles per year

Heterogeneous information extraction

Analyzed from the ISI database, DE author keywords, C1 field represents the author institution, the citation relationship between articles is obtained from the CR field, and the author and institution information is split from the author institution field. The institution information only intercepts information such as university hospitals.

After splitting, 1,682,654 authors were obtained, and 1,461,721 authors were merged. More than 40,000 authors were selected as the first author to publish articles. Considering calculation performance, we selected 5,403 of these authors who have collaborated with other authors more than 2 times. The total number of collaboration between these authors is 4,818, forming Table 3 shows the basic statistical characteristics of the benchmark cooperative network by year.

Table. 3. Annual statistical characteristics of the collaboration network

Year	#Nodes	#Edges	Density
2007	65	43	2.0673E-02
2008	402	296	3.6724E-03
2009	529	367	2.6279E-03
2010	705	486	1.9584E-03

2011	844	593	1.6669E-03
2012	1044	723	1.3280E-03
2013	1255	878	1.1158E-03
2014	1355	960	1.0465E-03
2015	1515	1136	9.9053E-04
2016	1384	969	1.0125E-03
2017	1259	847	1.0696E-03
2018	1052	745	1.3476E-03
2019	392	255	3.3274E-03

The predictability of collaboration networks

Predictability is an important research problem in link prediction. The predictability of the network represents the upper limit of prediction. Random networks are completely unpredictable. Any link prediction algorithm will not get better results on completely random networks. The predictability of the network is related to the characteristics and evolution of the network itself. Two articles by Newman et al [29,30] studied the structural path and other properties of the scientific research collaboration network. Xu Xiaoke [31], Tan Suoyi [32] and others have conducted special research on the predictability of the network. Studies have shown that in networks with good predictability, the largest eigenvalue of the adjacency matrix of the network is much larger than the second largest eigenvalue.

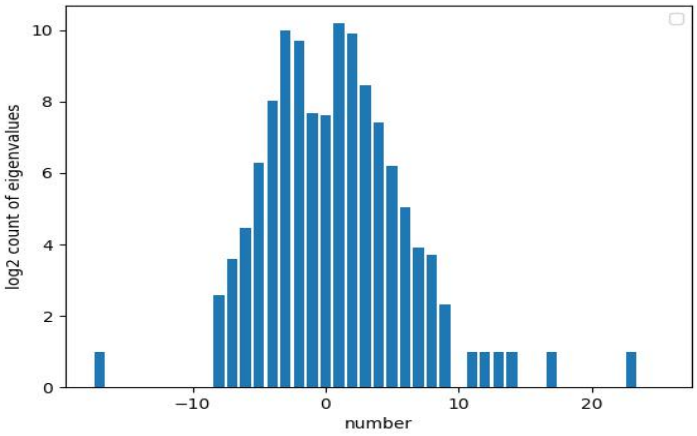


Fig.5. Eigenvalue distribution of adjacency matrix

Figure 6 shows the eigenvalue distribution of the critical matrix of the studied network. The largest eigenvalue is about 23.37. There are obvious gaps on the right and left sides of the figure, and the network will have better predictability.

3.2 The impact of features on collaboration from the perspective of causal

In this part, we use the method introduced in section 2.2 to compare the effects of the same institution, co-word, and citation on the probability of collaboration by year.

Institutional Information

First, we studied whether the same institution has a causal effect on scientific research collaboration in a large network composed of the first authors of all cooperative articles, and calculated separately the same institution-there is collaboration, the same institution-no collaboration, and different institutions-the side of collaboration. The number, and then subtract the first few numbers from all possible edges to get the frequency of different organizations-no collaboration. The results are shown in Table 4, suppose that event **X**: there is collaboration between authors, and **Y**: the authors have the same organization.

Table. 4. Same Institution and its impact of collaboration in the big network

	X	~X
Y	16205	1908127
~Y	24937	984907682

P(X|Y)=16205/(16205+1908127)=8.421e-3

P(X|~Y)=24937/(24937+984907682)=2.5318e-5

The average causal effect of event Y on event X:

ACE(Y—>X)=P(X|Y)-P(X|~Y)=8.396e-3

The probability of collaboration is increased (**PI**) by 331.62.

Next, we disassemble the collaboration network of 5403 nodes and 4818 edges according to the year to discuss the causal effect of scientific research collaboration with the institution each year. The results are shown in Table 5.

Table.5. The impact of the same institution on collaboration by year

Year	X&Y	X &~Y	~X& Y	~X&~Y	P(X|~Y)	P(X|Y)	PI
2007	12	21	4	2043	1.0174E-02	7.5000E-01	72.71
2008	79	139	240	80143	1.7314E-03	2.4765E-01	142.03
2009	116	159	473	138908	1.1433E-03	1.9694E-01	171.25
2010	171	192	1030	246767	7.7746E-04	1.4238E-01	182.14
2011	197	265	1318	353966	7.4810E-04	1.3003E-01	172.82
2012	246	310	1667	542223	5.7139E-04	1.2859E-01	224.05
2013	287	397	2177	784024	5.0611E-04	1.1648E-01	229.14
2014	321	421	2559	914034	4.6038E-04	1.1146E-01	241.10
2015	341	557	2515	1143442	4.8689E-04	1.1940E-01	244.23
2016	332	409	2091	954204	4.2845E-04	1.3702E-01	318.81
2017	288	355	1622	789646	4.4937E-04	1.5079E-01	334.55
2018	247	310	1251	551018	5.6228E-04	1.6489E-01	292.25
2019	80	116	116	76324	1.5175E-03	4.0816E-01	267.97

It can be seen from Figure 5 that for each time slice of the collaboration network, the same institution has a very obvious causal effect on scientific research

collaboration, and the probability of scientific research collaboration with the same institution is increased by 200-300 times.

Keyword information

This part discusses whether the authors have common keywords that affect scientific research collaboration. Let event **X**: there is collaboration between authors, and **Y**: there are common keywords used by authors. The method similar to section 3.2.1 is used for annual calculation. There is a causal effect of common use of keywords on scientific research collaboration. The results are shown in Table 6.

Table. 6. The impact of the co-word on collaboration by year

Year	X&Y	X &~Y	~X& Y	~X&~Y	P(X|~Y)	P(X|Y)	PI
2007	0	33	0	2047	1.5865E-02	0	-1
2008	1	217	8	80375	2.6926E-03	1.1111E-01	40.27
2009	34	241	1052	138329	1.7392E-03	3.1308E-02	17.00
2010	95	268	3727	244070	1.0968E-03	2.4856E-02	21.66
2011	139	323	8327	346957	9.3009E-04	1.6419E-02	16.65
2012	189	367	20044	523846	7.0010E-04	9.3412E-03	12.34
2013	228	456	26193	760008	5.9963E-04	8.6295E-03	13.39
2014	290	452	42131	874462	5.1662E-04	6.8362E-03	12.23
2015	352	546	58446	1087511	5.0181E-04	5.9866E-03	10.93
2016	341	400	53734	902561	4.4299E-04	6.3061E-03	13.24
2017	313	330	58653	732615	4.5024E-04	5.3081E-03	10.79
2018	349	208	58258	494011	4.2087E-04	5.9549E-03	13.15
2019	138	58	9827	66613	8.6994E-04	1.3848E-02	14.92

As can be seen from Table 6, for each time slice of the collaboration network, the use of the same keyword has a causal effect on collaboration, and the co-word increases the probability of collaboration by about 10-20 times. Not as obvious as the causal effect of the same organization. This may also be due to the fact that the authors tend to use some common keywords in the small areas of our research. These common keywords that are used too frequently have weakened the impact of keyword co-occurrence on collaboration.

Reference information

This part discusses the citations on research collaboration. Let event **X**: there is collaboration between authors and **Y**: there is citation between authors,that is author a cite some documents of author b or author b cite some documents of author a. Using the method similar to section 3.2.1, the causal effect of citations on scientific research collaboration is calculated annually. . The results are shown in Table 7.

Table. 7. The impact of the citations on collaboration by year

Year	X&Y	X &~Y	~X& Y	~X^~Y	P(X|~Y)	P(X|Y)	PI
2007	0	33	0	2047	1.5865E-02	0	-1
2008	0	218	0	80383	2.7047E-03	0	-1
2009	3	272	32	139349	1.9481E-03	8.5714E-02	43.00
2010	44	319	228	247569	1.2869E-03	1.6176E-01	124.70
2011	115	347	825	354459	9.7800E-04	1.2234E-01	124.09

2012	145	411	1537	542353	7.5724E-04	8.6207E-02	112.84
2013	213	471	2518	783683	6.0065E-04	7.7993E-02	128.85
2014	259	483	4132	912461	5.2906E-04	5.8984E-02	110.49
2015	312	586	5165	1140792	5.1341E-04	5.6965E-02	109.95
2016	325	416	4527	951768	4.3689E-04	6.6983E-02	152.32
2017	294	349	5209	786059	4.4379E-04	5.3425E-02	119.38
2018	320	237	5521	546748	4.3328E-04	5.4785E-02	125.44
2019	126	70	728	75712	9.2370E-04	1.4754E-01	158.73

It can be seen from Figure 6 that for each time slice of the collaboration network, the causal effect of citations on collaboration is strong, and the citations increase the probability on collaboration by about 100 times. Not as obvious as the causal effect of the same institution, but obviously stronger than the effect of keyword.

3.3 Research on performance and efficiency of network embedding based collaboration prediction algorithms

Divide the data set into training set and test set according to the ratio of 80%-20%, 60%-40%, and 40%-60%, and discuss the performance of the seven algorithms introduced in section 2.3. Each algorithm takes the same embedding dimension, and the 4 numbers in each cell represent the (ROC_AUC, AUC, F1_Score, Run time) introduced in Section 2.3.

Table. 8. Performance and efficiency of embedding algorithms

Algorithm	dataset1(80% training set,20% test set)	dataset2(60% training set，40% test set)	dataset3(40% training set，60% test set)
ProNE	(0.8548, 0.7638, 0.7248, 16.9s)	(0.8181, 0.6891, 0.6459, 16.13s)	(0.8079, 0.6600, 0.6278, 16.68s)
NetMF	(0.8686, 0.7554, 0.7166, 19.3s)	(0.8370, 0.7079 , 0.6683, 21.89s)	(0.8152, 0.6550, 0.6270, 23.27s)
Hope	(0.4760, 0.3014, 0.2643, 33.7s)	(0.5559, 0.3355, 0.2998, 26.26s)	(0.5910, 0.3760, 0.3562, 24.31s)
LINE	(0.8833, 0.7906, 0.7439, 2203s)	(0.8376, 0.7179, 0.6715, 1788s	(0.8150, 0.6679, 0.6358, 1297s)
Node2vec	(0.8049 ,0.4355, 0.3379, 404.6s)	(0.7592 , 0.3895, 0.3238, 326.5s)	(0.7576, 0.3563, 0.3099, 244.0s)
Deepwalk	(0.8062, 0.4431, 0.3379, 340.4s)	(0.7628, 0.3960, 0.3174, 257.0s)	(0.7528, 0.3548, 0.3083, 219.0s)
SDNE	(0.5096, 0.3357, 0.2888, 1661s)	(0.5184, 0.3517, 0.3110, 974.0s)	(0.5541, 0.3789, 0.3514, 609.8s)

As can be seen from Table 8 ,the larger the proportion of the training set, the higher the effect of various algorithms. Algorithmic comparison. It seems that the LINE model has the best accuracy, with the highest ROC_AUC, AUC, and F1_Score values, but the running time of the LINE model is too long, which is hundreds of times the running time of fast models such as ProNE and NetMF. The time efficiency of classic random walk algorithms such as Deepwalk and Node2vec is moderate, but the accuracy is slightly worse in this example, which may be related to parameter

selection. The accuracy of matrix factorization models such as ProNE and NetMF is not much different from the LINE model, but the time efficiency is improved by hundreds of times. Especially for the NetMF model, the accuracy is very small and the time efficiency is very high. Therefore, the following experiments choose NetMF as the network representation learning model.

3.4 Results of the multiple features fused cooperation prediction methods based on NetMF

In this part, we select the collaboration data of a certain year as the positive examples of the test data, randomly generate the same number of negative examples as the test positive examples, and select the collaboration data, institutional data, and author co-word data before that year (excluding the year). The cross-citation data is fused as a training set. Several types of feature fusion methods discussed in Section 2.4 are used for experiments. Table 9 shows the results of single feature fusion, Table 10 shows the results of multiple features fusion, and the three numbers in the cell in the table indicate (ROC_AUC, F1_Score, AUC), the best result in each row is expressed in bold.

Table. 9. Single feature fusion collaboration prediction results

Year	Raw Network	Raw+Institution	Raw+Keyword	Raw+Citation
2009	(0.7053, 0.6757, 0.8176)	**(0.8551, 0.8093, 0.9143)**	(0.7047, 0.6785, 0.8157)	(0.7287, 0.7084, 0.8311)
2010	(0.7579, 0.6975, 0.8279)	**(0.8603, 0.8086, 0.9018)**	(0.7278, 0.6872, 0.8085)	(0.7794, 0.7469, 0.8461)
2011	(0.7644, 0.688, 0.8091)	**(0.8767, 0.7926, 0.8922)**	(0.7497, 0.6897, 0.8018)	(0.785, 0.7099, 0.8216)
2012	(0.7389, 0.657, 0.7694)	**(0.854, 0.7746, 0.869)**	(0.718, 0.6515, 0.7564)	(0.7808, 0.7206, 0.8089)
2013	(0.7471, 0.6674, 0.7708)	**(0.8683, 0.7813, 0.8744)**	(0.7476, 0.6834, 0.7754)	(0.8232, 0.7528, 0.8341)
2014	(0.7231, 0.6385, 0.7214)	**(0.8684, 0.7719, 0.8616)**	(0.7255, 0.6531, 0.7408)	(0.8273, 0.749, 0.8307)
2015	(0.7301, 0.6347, 0.7151)	**(0.8630, 0.7861, 0.8648)**	(0.7554, 0.6655, 0.7549)	(0.8371, 0.7456, 0.8296)
2016	(0.749, 0.6615, 0.7011)	**(0.8977, 0.8225, 0.8766)**	(0.8024, 0.7059, 0.7742)	(0.8688, 0.7833, 0.8463)
2017	(0.756, 0.6635, 0.6905)	**(0.9041, 0.8276, 0.8729)**	(0.7938, 0.6989, 0.7574)	(0.8801, 0.7804, 0.8429)
2018	(0.8072, 0.698, 0.7212)	(0.9132, 0.8268, 0.8743)	(0.8489, 0.7463, 0.7988)	**(0.9225, 0.8282, 0.8807)**
2019	(0.8471, 0.6196, 0.6761)	(0.9501, **0.8275**, 0.8343)	(0.9073, 0.7373, 0.7815)	**(0.9678,** 0.8118**, 0.882)**

Table. 10. Multi-feature fusion collaboration prediction results

Year	Raw+I+K	Raw+K+C	Raw+I+C	Raw+I+K+C
2009	(0.8374, 0.7820, 0.9064)	(0.7301, 0.7003, 0.8286)	(**0.8550**, 0.8038, **0.9154**)	(0.8547, **0.8147**, 0.9150)
2010	(0.8627, 0.7963, 0.9012)	(0.7547, 0.7181, 0.8277)	(0.8652, 0.8107, 0.9041)	**(0.8826, 0.8292, 0.9140)**
2011	(0.8542, 0.7723, 0.8782)	(0.7936, 0.7184, 0.8312)	(0.8833, 0.8078, 0.9009)	**(0.8842, 0.8111, 0.8992)**
2012	(0.8496, 0.7621, 0.8637)	(0.7697, 0.6888, 0.7920)	(0.8792, 0.8147, 0.8924)	**(0.8804, 0.7911, 0.8883)**
2013	(0.8726, 0.7745, 0.8696)	(0.8198, 0.7472, 0.8293)	**(0.903, 0.8269, 0.9035)**	(0.8908, 0.8132, 0.8904)
2014	(0.8375, 0.7375, 0.8316)	(0.804, 0.7000, 0.7995)	**(0.9028, 0.826, 0.8953)**	(0.8962, 0.8021, 0.8829)
2015	(0.8541, 0.7368, 0.8359)	(0.8194, 0.7201, 0.8063)	**(0.9166, 0.8257, 0.9011)**	(0.8965, 0.7923, 0.8788)
2016	(0.8887, 0.7781, 0.8536)	(0.8495, 0.7472, 0.8161)	**(0.9406, 0.8648, 0.9194)**	(0.9187, 0.8328, 0.8894)
2017	(0.8854, 0.7627, 0.8361)	(0.8685, 0.7438, 0.8139)	**(0.9355, 0.8524, 0.9041)**	(0.9301, 0.8076, 0.8801)
2018	(0.9003, 0.7678, 0.8357)	(0.9032, 0.7893, 0.8455)	**(0.9583, 0.8644, 0.9193)**	(0.9388, 0.8148, 0.8836)
2019	(0.9344, 0.7333, 0.7811)	(0.9329, 0.7373, 0.7862)	**(0.9657, 0.8196, 0.8739)**	(0.9438, 0.7529, 0.8034)

Combining Table 9 and Table 10, it can be clearly seen that (1) Almost all feature fusion cooperative prediction methods are larger than the original network in ROC_AUC, F1_Score, and AUC, indicating that the feature fusion method can improve the accuracy of cooperative prediction. (2) Same institution (I)> reference relationship (K)> same keyword (C), which is exactly the same as the causal effect sequence of the several types of relationships discussed in Section 3.2. (3) The accuracy comparison of multiple features fusion methods is generally I+C> I+K+C> I+K> K+C. In the case of simple network nodes and relationships, the I+K+C three-feature fusion method is better than I+C two features. Multiple features can bring more relationships and improve the prediction effect. There are relatively many data nodes later, and the prediction result is close to the upper limit of the predictability of the network. Adding keyword co-occurrence features may cause more confusion in the network due to frequently occurring words, and the prediction effect will not be further improved.

3.5 Collaboration prediction results

We randomly select scientific researchers and predict collaborators for him/her. In this example, we selected the researchers Lin Mingyan of Albert Einstein Coll Med, and used the NetMF of Raw+I+C feature fusion with better results in the

experiment in Section 3.4. Table 11 lists the top 20 authors with the highest probability of collaboration in the future.

Table. 11. Authors with the top 20 collaboration probability

Author	Institution	sim	Author	Institution.	Sim
Zheng, Deyou	Albert Einstein Coll Med	0.9977	Delahaye, Fabien	Albert Einstein Coll Med	0.9783
Pedrosa, Erika	Albert Einstein Coll Med	0.9967	Rockowitz, Shira	Albert Einstein Coll Med	0.9768
Chen, Jian	Albert Einstein Coll Med	0.9965	Wijetunga, N. Ari	Albert Einstein Coll Med	0.9643
Zhao, Dejian	Albert Einstein Coll Med	0.9937	Pal, Rajarshi	Manipal Univ Branch Campus	0.9590
Wang, Ping	Albert Einstein Coll Med	0.9928	Carromeu, Cassiano	Univ Calif San Diego	0.9240
Xue, E.	Albert Einstein Coll Med	0.9924	Marchetto, Maria C. N.	Salk Inst Biol Studies	0.9225
Sharma, V. P.	Albert Einstein Coll Med	0.9924	Zhou, Li	Albert Einstein Coll Med	0.9148
Abrajano, Joseph J.	Albert Einstein Coll Med	0.9904	Jaffe, Andrew E.	Lieber Inst Brain Dev	0.8058
Guo, Xingyi	Albert Einstein Coll Med	0.9891	Lei, Mingxing	Univ So Calif	0.7802
Qureshi, Irfan A.	Albert Einstein Coll Med	0.9851	Will, Britta	Albert Einstein Coll Med	0.7758

The first few authors in the results seem to be the first few authors who have worked closely with the author, and the chances of continuing to cooperate in the future are also very high. In the results of the collaboration recommendation, the proportion of the same institution as the author is very large, accounting for 75%, which is also consistent the law of scientific research collaboration.

4 Conclusion and Discussion

This paper constructs the collaboration prediction method based on heterogeneous information fused network embedding, and conducts an empirical analysis in the field of stem cells.

(1) Construct a author-centered heterogeneous information fusion schema, based on information fusion theory. The predictability of the scientific research collaboration network and the effects of institutions, co-word, and citation information on collaboration are discussed. Experiments show that the they have an impact on collaboration. The average causal effect analysis shows that the influence order of the three factors is as follows(same institution> citation> keywords).

(2) The accuracy and efficiency of the network representation learning methods for collaboration are discussed. Experiments show that the accuracy and

computational efficiency of the comprehensive method, and the graph representation learning method based on the new matrix factorization (like NetMF) has achieved good results.

(3) Construct a prediction method for scientific research collaboration based on heterogeneous information fusion, and conduct empirical analysis in a yearly network. Experiments show that the method of multiple features fusion can greatly improve the accuracy of collaboration prediction. In terms of feature combination, The combination of the same institution + citation relationship has achieved outstanding results.

Future research will build some more complicated causal diagram under the author-centered of information fusion framework, explore the causal effects of feature combinations; explore more detailed methods for selecting relationships within features, and continuously explore relationships in various information, and improve the identification effect. Expand the framework of information fusion, and introduce other invisible features such as research fields, research topics, and writing styles into the framework of collaboration prediction, and continuously enrich the connotation of the methods.

References

1. Newman M.E.J. Coauthorship Networks and Patterns of Scientific Collaboration. Proceedings of the National Academy of the United States of America,(101) : 5200-5205 (2004)
2. Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. Journal of the American society for information science and technology, 58(7): 1019-1031(2007).
3. Lü Linyuan. Link Prediction in Complex Networks. Journal of University of Electronic Science and Technology of China, 39(05):651-661(2010).
   (吕琳媛.复杂网络链路预测.电子科技大学学报,39(05):651-661.(2010))
4. Guns R, Rousseau R. Recommending research collaborations using link prediction and random forest classifiers. Scientometrics, 101(2): 1461-1473(2014).
5. Yan E, Guns R. Predicting and recommending collaborations: An author-, institution-, and country-level analysis. Journal of Informetrics, 8(2): 295-309(2014).
6. Shan Songyan, Wu Zhenxin. Review on the author similarity algorithm in the field of author name disambiguation and research collaboration prediction .Journal of Northeast Normal University(Natural Science Edition), ,51(02):71-80(2019).
   (单嵩岩,吴振新.面向作者消歧和合作预测领域的作者相似度算法述评.东北师大学报(自然科学版),51(02):71-80(2019).)
7. Zhang Jinzhu, Yu Wenqian, Liu Jingjie, Wang Yue. Predicting Research Collaborations Based on Network Embedding.. Journal of the China Society for Scientific and Technical Information,37(02): 132- 139 (2018).
   (张金柱,于文倩,刘菁婕,王玥.基于网络表示学习的科研合作预测研究[J].情报学报, 37(02): 132 -139,(2018)).

8. Yu Chuanming, Lin Aochen, Zhong Yunci, An Lu. Scientific Collaboration Recommendation Based on Network Embedding. Journal of the China Society for Scientific and Technical Information.38(05): 500-511(2019).
(余传明,林奥琛,钟韵辞,安璐.基于网络表示学习的科研合作推荐研究[J].情报学报 38(05): 500 - 511(2019)).
9. Wang Zhibing, Han wenmin, Sun Zhumei, Pan xuelian.Research on Scientific Collaboration Prediction Based on the Combination of Network Topology and Node Attributes. Information Studies: Theory & Application, (08):116-120+109(2019).
(汪志兵,韩文民,孙竹梅,潘雪莲.基于网络拓扑结构与节点属性特征融合的科研合作预测研究.情报理论与实践,(08):116-120+109(2019)).
10. Liu P, Zheng K , Zou D. Research on Recommendation S&T colleboration based on LDA model. Information Studies: Theory &Application.38(9): 79-85(2015).
(刘萍, 郑凯伦, 邹德安. 基于LDA模型的科研合作推荐研究.情报理论与实践, 38(9): 79-85(2015)).
11. Yu C, Gong Y,Zhao S,et al. Collaboration Recommendation of Finance Research Based on Multi-feature Fusion. Data Analysis and Knowledge Discovery,(8): 39-47(2017).
(余传明, 龚雨田, 赵晓莉, 等. 基于多特征融合的金融领域科研合作推荐研究. 数据分析与知识发现, (8): 39-47(2017)).
12. Lin Y,Wang K,Liu H,et al. Application of Network Representation Learning in the Prediction of Scholar Academic collaboration.Journal of the China Society for Scientific and Technical Information, 39(04):367-373(2020).
(林原,王凯巧,刘海峰,许侃,丁堃,孙晓玲.网络表示学习在学者科研合作预测中的应用研究.情报学报, 39(04):367-373(2020)).
13. Zhang X,Wen Y,Xu H. A Fusion Model of Network Representation Learning and Topic Model for Author collaboration Prediction. Data Analysis and Knowledge Discovery (2021).
(张鑫,文奕,许海云.一种融合表示学习与主题表征的作者合作预测模型.数据分析与知识发现: 1-19(2021)).
14. Xu H, Dong K, Wei L et al. Research on Multi-source Data Fusion Method in Scientometrics. Journal of the China Society for Scientific and Technical Information,37(03):318- 328(2018).
(许海云,董坤,隗玲,王超,岳增慧.科学计量中多源数据融合方法研究述评.情报学报, 37(03):318- 328(2018)).
15. Hua B,Li G. Discussion on Theory and Application of Multi-Source Information Fusion in Big Data Environment. Library and Information Service ,59(16):5-10 (2015).
(化柏林,李广建.大数据环境下多源信息融合的理论与应用探讨.图书情报工作, 2015,59(16):5-10.)
16. Morris S A, Yen G G. Construction of bipartite and unipartite weighted networks from collections of journal papers. Physics, (2005).
17. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM:701-710(2014).
18. Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding//Proceedings of the 24th international conference on world wide web. International World Wide Web Conferences Steering Committee,1067-1077(2015).

19. Grover A, Leskovec J. node2vec: Scalable feature learning for networks//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 855-864(2016).
20. Wang D, Peng C, Zhu W. Structural Deep Network Embedding// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. (2016).
21. Ou M, Cui P, Pei J, et al. Asymmetric transitivity preserving graph embedding//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM: 1105-1114(2016).
22. Qiu, Jiezhong , et al. "Network Embedding as Matrix Factorization: Unifying DeepWalk, LINE, PTE, and node2vec." the Eleventh ACM International Conference ACM, (2018).
23. Qiu J, Dong Y, Ma H, et al. Netsmf: Large-scale network embedding as sparse matrix factorization//The World Wide Web Conference. ACM: 1509-1520(2019).
24. Jie Zhang, Yuxiao Dong, Yan Wang,et al. ProNE: Fast and Scalable Network Representation Learning.//In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19),(2019)
25. Lerer A, Wu L, Shen J, et al. PyTorch-BigGraph: A Large-scale Graph Embedding System// Proceedings of the 2nd SysML Conference,(2019).
26. Fey M, Lenssen J E. Fast graph representation learning with PyTorch Geometric. arXiv preprint arXiv:1903.02428, (2019).
27. Zhu Z, Xu S, Tang J, et al. GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding//The World Wide Web Conference. ACM, 2019: 2494-2504.
28. Lü L. “Link Prediction in Complex Networks”. Journal of University of Electronic Science and Technology of China,39(05) , pp .651-661(2010).
29. Newman, M. E J. Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E , 64(1):016131(2001).
30. Newman, M. E J . Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E Statal Nonlinear & Soft Matter Physics, 64(1):016132 (2001).
31. Xu X, Xu S,Zhu Y,et al,Link Predictability in Complex Networks. Complex Systems and Complexity Science,11(01): 41-47(2014).
(许小可,许爽,朱郁筱,张千明.复杂网络中链路的可预测性.复杂系统与复杂性科学, 11(01):41-47(2014)).
32. Tan S,Qi M,Wu J et al. Link predictability of complex network from spectrum perspective. Acta Physica Sinica,69(08):188-197(2020).
(谭索怡,祁明泽,吴俊,吕欣.复杂网络链路可预测性:基于特征谱视角.物理学报,69(08):188-197.(2020)).