基于图卷积半监督学习的论文作者同名消歧方法研究

盛晓光*① 王 颖② 钱 力②③ 王 颖①

①(中国科学院大学人工智能学院 北京 100049)

②(中国科学院文献情报中心 北京 100190)

③(中国科学院大学图书情报与档案管理系 北京 100190)

摘 要：为解决学者与成果的精确匹配问题，该文提出了一种基于图卷积半监督学习的论文作者同名消歧方法。该方法使用SciBERT预训练语言模型计算论文题目、关键字获得论文节点语义表示向量，利用论文的作者和机构信息获得论文的合作网络和机构关联网络邻接矩阵，并从论文合作网络中采集伪标签获得正样本集和负样本集，将这些作为输入利用图卷积神经网络进行半监督学习，获得论文节点嵌入表示进行论文节点向量聚类，实现对论文作者同名消歧。实验结果表明，与其他消歧方法相比，该方法在实验数据集上取得了更好的效果。

关键词：同名消歧；图卷积神经网络；BERT语言模型

中图分类号：TP391.1 **文献标识码**：A **文章编号**：1009-5896(2021)12-3442-09

DOI: 10.11999/JEIT200905

Author Name Disambiguation Based on Semi-supervised Learning with Graph Convolutional Network

SHENG Xiaoguang① WANG Ying② QIAN Li②③ WANG Ying①

①(*School of Artificial Intelligence*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

②(*National Science Library*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

③(*Department of Library*, *Information and Archives Management*, *University of Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: In order to solve the problem of exact matching between scholars and articles, a new method of author name disambiguation is proposed based on semi-supervised learning with graph convolutional network. In this method, the SciBERT pre-training language model is applied to calculating the semantic embedding vector of each paper with their title and keywords. Authors and organizations of papers are used to obtain the adjacency matrixes of the paper's co-author network and co-organization network. The pseudo labels are collected from the co-author network to obtain the positive and negative samples. The semantic embedding vector, adjacency matrixes and the positive and negative samples are used as input to be processed by Graph Convolution neural Network (GCN). In semi-supervised learning, the embedding vectors of papers are learned to be clustered in order to realize the name disambiguation of papers. The experimental results show that, compared with other disambiguation methods, this method achieves better results on the experimental dataset.

Key words: Name disambiguation; Graph Convolutional Network (GCN); BERT language model

1 引言

作者姓名歧义一直是国内外出版界和学术界的难点问题。近年来科学论文数量呈指数级增长，重名现象越来越严重，特别是名称缩写、拼音一音多字等问题影响着文献检索系统以及学术评价的准确性。为消除歧义，许多研究机构提出了人名标识系统以期通过唯一标识来区分作者，如开放研究者与贡献者身份识别码(Open Researcher and Contributor IDentifier, ORCID)[1]、Thomson Reuters的ResearchID[2]等。然而，人名标识系统的应用范围有限，大量科学出版物中并未明确标注作者身份识别码。因此，通过自动化方法解决论文中作者歧义问题仍然是同名消歧的主要手段，也是国内外学者

收稿日期：2020-10-23；改回日期：2021-09-23；网络出版：2021-11-10

*通信作者： 盛晓光 shengxiaoguang@ucas.ac.cn

基金项目：国家自然科学基金(61702038)，国家社会科学基金(15CTQ006)

Foundation Items: The National Natural Science Foundation of China (61702038), The National Social Science Foundation of China (15CTQ006)

的研究热点之一。常用的作者消歧方法往往将问题转化为机器学习的聚类问题或分类问题，如利用SVM[3]、层次聚类[4]、谱聚类[5]等机器学习算法进行处理。随着深度学习技术的发展，越来越多研究人员采用网络嵌入方法(Network Embedding)进行作者同名消歧[6,7]，从论文数据中抽取特征以便于聚类或分类任务。此外，具有表征学习能力的卷积神经网络(Convolutional Neural Networks, CNN)快速发展，在计算机视觉[8,9]、自然语言处理[10]等领域都取得了巨大成功，而图卷积神经网络(Graph Convolutional Network, GCN)由于能够有效处理具有丰富关系结构的任务，常用于处理图节点表示学习、图节点分类、边预测、图分类等问题[11–14]。鉴于此，本文提出了一种基于图卷积半监督学习的论文作者同名消歧方法，融合作者、机构、题目、关键词等论文属性信息，借助BERT语义表示方法和图卷积神经网络，探索作者消歧方法，以提高作者与成果的匹配效果。

2　相关研究

Zhang等人[6]将当前同名消歧的研究方法分为两类：基于特征的消歧方法和基于连接/图的消歧方法。

基于特征的消歧方法应用较早，根据文档的特征向量学习文档之间的距离函数，将相近的特征向量归入相同类别，实现同名消歧。Huang等人[15]提出了一个有效的综合框架来解决名称消歧问题，分别利用Blocking技术检索具有相似名称作者的候选类，使用在线主动选择支持向量机算法(LASVM)计算论文之间的距离度量进行DBSCAN聚类。Yoshida等人[16]提出一种基于bootstrapping的两阶段聚类算法来改善低查全率，其中第1阶段的聚类结果用于提取第2阶段聚类中使用的特征。Han等人[3]提出了基于SVM和贝叶斯网络的有监督消歧方法，利用论文合作者、题目出版物名称等特征对同名作者进行消歧。Zhu等人[17]使用多层聚类的方式进行同名消歧，如分别利用Email信息、论文合作者、论文题目等进行动态的作者聚类。

基于连接/图的消歧方法利用图的拓扑结构或者聚合来自邻居节点的信息，例如Fan等人[18]提出了一种仅使用合作者关系的同名消歧框架GHOST，通过合作关系构造图，根据图中待排歧作者间有效路径的数目和长度计算相似度，再对相似度矩阵聚类实现同名消歧。Tang等人[19]利用隐马尔可夫随机域对统一概率框架下的节点特征和边特征进行建模。Zhang等人[7]提出一种基于网络嵌入的解决方案，构建作者-作者、作者-论文、论文-论文3个图，利用各种匿名网络的链接结构，将每个文档表示为低维向量空间，以解决名称消歧任务。Hermansson等人[20]提出了一种基于局部邻域结构的匿名图实体消歧方法，基于局部邻域结构利用Graph Kernels计算图中节点之间的相似度，并用SVM执行分类任务。Zhang等人[6]采用结合全局监督和局部上下文的表示学习方法，采用该技术的名称消歧模块应用在AMiner系统中能够高效处理十亿级规模的消歧问题。

本文结合两种消歧方法的优势，一方面利用论文文本属性信息如题目、关键词等计算语义特征向量，再通过合作关系和同机构关系构建论文网络，将卷积用于图结构进行半监督学习，达到作者消歧的目的。

3　基于图卷积半监督学习的作者同名消歧方法

图卷积神经网络是一种最为典型的图神经网络。图卷积半监督学习利用卷积操作将节点的特征向量和节点间的图结构结合在一起，节点的特征向量每经过1次图卷积操作，就通过图结构利用临近节点更新自己的特征向量，从而使相似的节点具有相似的特征向量[21]。此过程适用于作者同名消歧任务，待消歧论文通过相互关联构建网络并通过图卷积网络不断更新特征向量实现论文聚类任务。

基于这一思路，本文提出一种基于图卷积半监督学习的作者同名消歧方法框架如图1所示。首先，将论文的题目、关键字作为文本输入预先训练好的SciBERT模型得到每篇论文的语义表示向量；其次，利用论文的作者和机构信息构建论文合作网络与机构关联网络，分别获得邻接矩阵；然后，从论文合作网络中采集伪标签，获得正样本集和负样本集；将待消歧论文的BERT语义向量、论文合作网络和论文机构关联网络以及正、负样本集作为输入，利用图卷积神经网络进行半监督学习，获得论文最终节点向量；最后使用层次凝聚类算法将论文节点向量聚类划分，实现对论文作者同名消歧。

3.1　基于BERT预训练模型的论文语义表示

由于研究人员在一段时间内的研究方向相对稳定，论文的题目、关键词、摘要、出版物名称等文本特征也可用于表征作者的研究内容并用于区分从事不同研究的同名作者。目前，广泛使用的文本向量构建方法包括n-gram, NNLM, word2vec等。2018年Google发布了BERT预训练语言模型[22]，在自然语言处理的11个任务上大幅刷新了精度。随后，Beltagy等人[23]推出了专门为科学论文训练的SciBERT预训练语言模型，更适用于科学论文的

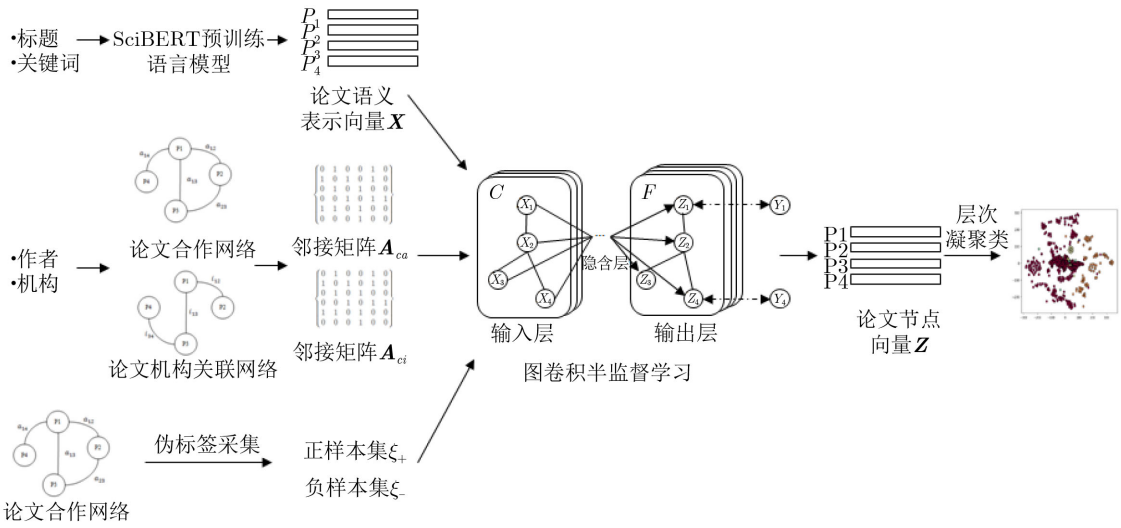


图 1 研究框架

自然语言处理任务。为充分利用论文文本特征，本文将论文的题目、关键词作为文本输入，利用SciBERT模型得到每篇论文的语义表示向量。

设每篇论文的题目和关键词拼接获得的句子输入为d，则BERT输入为[CLS, d, SEP]，CLS和SEP标识符分别作为句子的起始符和分隔符，经过分词获得句子的token序列$\{\text{tok}_1, \text{tok}_2, \cdots, \text{tok}_N\}$，依次输入到BERT模型中。BERT以双向Transformer的Encoder作为模型的基本组成单元(如图2中BERT层)，能够联合所有层中左右两个方向的上下文信息进行训练，利用多头注意力机制进行更多层面的特征提取，最后得到含有丰富语义特征的序列向量，即输出为该论文的语义表示向量，设为$\boldsymbol{d_S}$，向量维数为BERT的默认隐含元个数768，记为H。则待消歧同名作者论文集合的语义表示向量矩阵$\boldsymbol{X_{H\times K}} = (\boldsymbol{d_{S1}}, \boldsymbol{d_{S2}}, \cdots, \boldsymbol{d_{SK}})$，其中$\boldsymbol{d_{Si}}$为第$i$篇论文的语义表示向量，$K$为论文的数量。

3.2 论文合作网络和机构关联网络构建

为获得同名作者论文之间的关联信息，本文分别构建论文合作网络$\}_{ca}$和论文机构关联网络$\}_{ci}$，如图3所示。

定义1：论文合作网络$\}_{ca} =< P, \alpha >$表征论文节点之间的合作关系，其中P表示网络中节点集，节点类型为论文，α表示节点之间的合作关系边集合，如果论文节点P_1和P_2之间的待消歧作者的共同合作者数量大于1，则定义这两个论文节点在论文合作网络中存在边a_{12}。

定义2：论文机构关联网络$\}_{ci} =< P, i >$表征论文节点之间的机构关联关系，其中P表示网络中节点集，节点类型为论文，i表示节点之间的机构关联边集合，如果论文节点P_1和P_2的作者存在相同的所属单位则这两个论文节点在论文机构关联网络中存在边i_{12}。

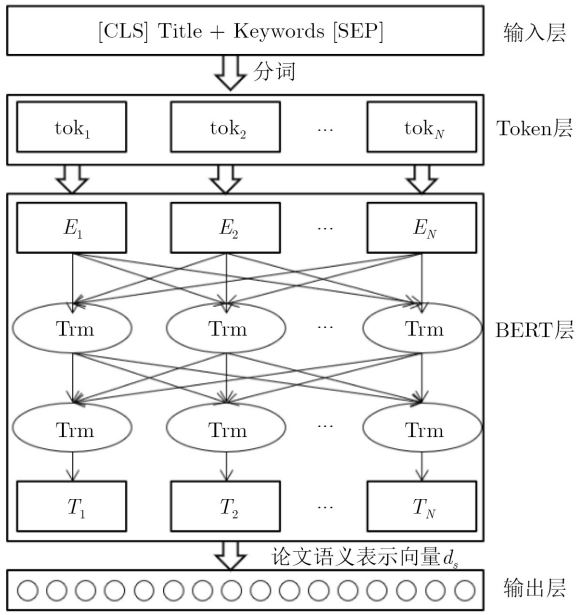
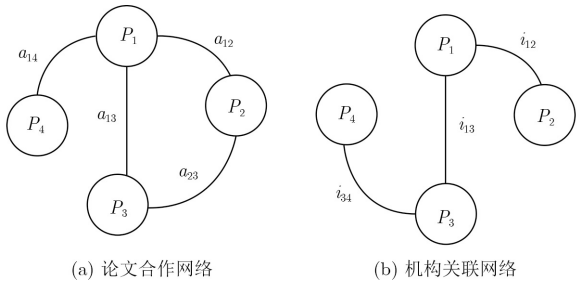


图 2 基于BERT预训练模型的论文语义表示



(a) 论文合作网络 (b) 机构关联网络

图 3 论文合作网络和机构关联网络

由此分别构建了论文之间的无权无向图$\mathfrak{g}_{ca}$和$\mathfrak{g}_{ci}$，用$\boldsymbol{A_{ca}}$和$\boldsymbol{A_{ci}}$分别表示论文合作网络和论文机构关联网络的邻接矩阵。

为了得到GCN训练的初始标签数据，本文利用已构建的论文网络采集伪标签。通过对比合作关系和机构关联关系，可以发现存在相同合作者的同名作者为同一人的概率相比同机构而言更大，为此从论文合作网络$\mathfrak{g}_{ca}$中采集伪标签。具体做法为定义集合$e_{ij}\in\xi_+$为图$\mathfrak{g}_{ca}$存在的边集合，即边e_{ij}在图$\mathfrak{g}_{ca}$的邻接矩阵中为1。同时随机采样同等数量不存在的边集合$e_{ij}\in\xi_$，即e_{ij}在图$\mathfrak{g}_{ca}$的邻接矩阵中为0。将ξ_+作为正样本集，ξ_-作为负样本集。

3.3 图卷积半监督节点表示学习

通过上述步骤分别获取了待消歧论文的BERT语义向量、论文合作网络和论文机构关联网络以及正、负样本集。在此基础上，使用图卷积半监督方法进一步学习每一个论文的节点表示。设图卷积神经网络的输入特征为BERT语义表示向量$\boldsymbol{X}$、论文合作网络邻接矩阵$\boldsymbol{A_{ca}}$、论文机构关联网络邻接矩阵$\boldsymbol{A_{ci}}$。

对于论文合作网络$\mathfrak{g}_{ca}$，记$\tilde{\boldsymbol{A}}_{\boldsymbol{ca}}$为正则化的邻接矩阵，两层的图卷积表示为

$$\boldsymbol{Z_{ca}}=\tilde{\boldsymbol{A}}_{ca}\sigma\left(\tilde{\boldsymbol{A}}_{ca}\boldsymbol{X}\boldsymbol{W}_1^{ca}\right)\boldsymbol{W}_2^{ca} \tag{1}$$

其中，$\boldsymbol{W}_1^{ca}$,$\boldsymbol{W}_2^{ca}$分别为第1层和第2层待学习参数，σ为ReLU激活函数，$\boldsymbol{Z_{ca}}$为论文合作网络方向的节点嵌入向量。

对于论文机构关联网络$\}_{ci}$，记$\tilde{\boldsymbol{A}}_{\boldsymbol{ci}}$为正则化的邻接矩阵，两层的图卷积表示为

$$\boldsymbol{Z_{ci}}=\tilde{\boldsymbol{A}}_{\boldsymbol{ci}}\sigma\left(\tilde{\boldsymbol{A}}_{\boldsymbol{ci}}\boldsymbol{X}\boldsymbol{W}_1^{ci}\right)\boldsymbol{W}_2^{ci} \tag{2}$$

其中，$\boldsymbol{W}_1^{ci}$,$\boldsymbol{W}_2^{ci}$分别为第1层和第2层待学习参数，$\boldsymbol{Z_{ci}}$为论文机构关联网络方向的节点嵌入向量。

为了更好地保留论文数据语义信息的作用，本文添加了一个全连接层，对BERT语义表示向量$\boldsymbol{X}$进行映射，获取其语义的表示向量

$$\boldsymbol{Z_{be}}=\boldsymbol{U}\boldsymbol{X} \tag{3}$$

其中，$\boldsymbol{U}$为全连接层的参数，$\boldsymbol{Z_{be}}$为基于BERT语义表示的节点嵌入表示。

由此获得了3个方向节点嵌入表示$\boldsymbol{Z_{ca}}$，$\boldsymbol{Z_{ci}}$和$\boldsymbol{Z_{be}}$，融合这3个方向的节点嵌入表示作为最终的论文节点向量进行半监督学习

$$\boldsymbol{Z}=(\beta_1\boldsymbol{Z_{ca}}+\beta_2\boldsymbol{Z_{ci}}+\beta_3\boldsymbol{Z_{be}})/(\beta_1+\beta_2+\beta_3) \tag{4}$$

其中，β_1，β_2和β_3分别是权重参数，本文选择权重分别为0.001，1，3。

半监督学习的目标是最小化正样本集中节点连接的距离，同时最大化负样本集中节点对的距离，损失函数定义为

$$\mathcal{L}=\mathrm{lam}\sum_{e_{ij}\in\xi_+}d(\boldsymbol{Z_i},\boldsymbol{Z_j})-(1-\mathrm{lam})\sum_{e_{ij}\in\xi_-}d(\boldsymbol{Z_i},\boldsymbol{Z_j}) \tag{5}$$

其中，$d(\cdot)$为距离函数，采用欧氏距离计算，lam为调和参数。

对上述获得整体的节点嵌入向量$\boldsymbol{Z}$，使用层次凝聚类算法对图中的n个论文进行聚类。层次凝聚类算法是一种凝聚型的聚类方法，相对于划分型的聚类方法更侧重于保留节点间已存在的相似性，而不至于忽略一些不够显著的关联，适用于本文提出的作者消歧方法。

3.4 算法流程

基于上述研究框架和具体方法，本文设计了如表1所示的实现算法，以同名作者的论文集合为输入，执行算法后输出这些论文的聚类集合。

4 实验结果

4.1 实验数据

由于学术论文来源不同，论文元数据信息往往存在缺少或不规范的情况。对于规模比较大的高校，不同分支机构甚至同一分支机构往往存在中英文同名学者。此外，部分学术数据服务商还将一些不同分支机构(如A大学计算机学院、A大学人工智能学院等)的数据统一处理为“A大学”，进一步

表 1　基于图卷积半监督学习的作者同名消歧算法

输入：同名作者论文集合P
输出：论文uuid序列和对应cluster_out列表

(1) 解析论文元数据，获得唯一标识符uuid、标题、关键词、摘要、出版物名称、作者列表、机构列表；
(2) 数据预处理如中英文转换、特殊字符处理等；
(3) 将标题和关键词的拼接文本d的列表作为BERT模型的输入，计算获得BERT语义表示向量$\boldsymbol{X}$；
(4) 遍历论文列表，构建论文网络$\}_{ca}$和$\}_{ci}$，建立合作关系和机构关联关系，获得邻接矩阵$\boldsymbol{A_{ca}}$和$\boldsymbol{A_{ci}}$；
(5) 从论文合作网络$\}_{ca}$中采集伪标签，获得正、负样本集ξ_+和ξ_-；
(6) 模型初始化，开始 GCN训练
(7) for epoch in range(nums_epoch):
(8) 根据式(1)执行图卷积；
(9) 根据式(2)执行图卷积；
(10) 根据式(3)执行全连接层运算；
(11) 根据式(4)获得节点向量$\boldsymbol{Z}$；
(12) 根据式(5)计算损失函数并反向传播梯度更新参数和节点向量；
(13) 反向传播更新参数；
(14) end for
(15) 利用训练后的最终论文节点向量$\boldsymbol{Z}$进行Agglomerative Clustering()聚类

加大了同名消歧的难度。为此，本节以从公开学术数据库获得的某高校论文数据为实验数据，对本节提出的方法进行实验验证。通过与该高校教育管理数据对比，待消歧的作者姓名数据集中，包括856个校内同名导师姓名(实际对应2285名导师)、52个同分支机构同名导师姓名(实际对应108名导师)。从中随机选取20个待消歧作者(实际对应96名导师)作为测试集(表2)。

20个待消歧作者的论文量共计3753篇，包括中文论文2473篇，英文论文1280篇，部分中文论文包含英文元数据，具有英文题目的论文合计2921篇。在实验前，对待消歧论文进行预处理，将中文姓名统一为英文姓名，利用机构规范库将中文机构统一为英文机构名，区分中英文题目、摘要、关键词、出版物名称，使用百度通用翻译API[1)]将中文翻译为英文，并统一处理缩写、停用词、特殊字符等。

4.2 实验结果与分析

本文实验开发环境为Python3.6, CUDA 10.01，使用PyTorch 1.1.0, Transformers 2.1.1, Gensim 3.8.1, Numpy 1.18.1等工具。硬件环境为Intel Xeon十核处理器、64GB内存、NVIDIA Geforce RTX 2080Ti显卡。

4.2.1 与其他方法对比分析

为综合评价本文方法，分别与匿名图网络嵌入消歧方法[7]、多维网络嵌入消歧方法[24]以及基于合作者和共现关键词等规则的基础消歧方法进行比较。选用常用的Pairwise Precision, Recall, F1-score作为评估指标，实验结果如表3所示。

从表3可以看出，本文方法在20个待消歧作者的11个子任务中都取得了最优效果，并且在所有子任务的平均指标上也取得了最优效果，F1值相比其他3种方法分别提升了3.57, 2.7和32.98。匿名图网络嵌入消歧方法在子任务(Jia Liu, Jie Liu, Jun Liu, Yunshan Wang, Xu Zhao)中消歧效果更优。多维网络嵌入消歧方法在子任务(Wei Li, Bin Wang, Lin Wang, Ming Zhu)中消歧效果更优。

表 2 待消歧作者测试集

姓名	论文数	真实作者数	姓名	论文数	真实作者数
Tao Huang	167	2	Yunshan Wang	46	2
Haibo Li	132	3	Liang Wang	119	6
Ming Li	312	10	Lin Wang	56	2
Wei Li	27	2	Gang Xiong	151	2
Jia Liu	29	2	Jun Yang	395	7
Jie Liu	30	2	Peng Zhang	237	10
Jing Liu	277	6	Tao Zhang	939	9
Jun Liu	228	7	Xu Zhao	131	3
Yun Liu	169	5	Feng Zhao	122	6
Bin Wang	201	8	Ming Zhu	31	2

[1)] https://api.fanyi.baidu.com/

从待消歧论文规模上比较，本文方法在论文量较大的任务(如Tao Zhang, Jun Yang, Ming Li)上效果好于其他方法。而在论文量较小的任务 (Wei Li, Jia Liu, Jie Liu, Yunshan Wang, Lin Wang, Xu Zhao, Ming Zhu)上匿名图网络嵌入消歧方法和多维网络嵌入消歧方法的效果稍优，本文方法次之。

从待消歧类别上比较，歧义类别较多的几组任务Ming Li, Peng Zhang, Tao Zhang上，本文方法较优，而匿名图网络嵌入消歧方法在消歧类别较少任务Jia Liu, Jie Liu, Yunshan Wang上更优，多维网络嵌入消歧方法在消歧类别较少任务Wei Li, Lin Wang, Ming Zhu上更优，本文方法在消歧类别较少的Tao Huang子任务上效果更优。

综合比较论文规模和消歧类别上各任务的消歧效果，本文方法的适应性更强，所以综合表现最好，表明其具有良好的细粒度区分能力和数据规模处理能力。

4.2.2 组件贡献分析

为评估本文使用模型各组成部分在聚类中的作用，分别仅利用BERT模型计算论文节点的语义表示向量进行聚类，设定语义向量为$\boldsymbol{0}$并且仅使用图卷积网络计算合作者和机构关系进行聚类，以及综合使用两个组件，即利用图卷积网络对节点向量优化后进行论文聚类，对比结果如表4所示。

表4结果显示如果仅用文本语义表示向量表示论文，聚类结果平均F1值为57.03，而利用图卷积网络利用合作关系和机构相似关系进行优化后，平均F1值提升了24.51。如果仅使用图卷积网络计算合作者和机构相似网络进行消歧，平均F1值为75.76，相比仅使用文本语义表示向量表示论文，提升效果显著，这说明联合使用合作关系和机构关联进行图卷积学习对于作者同名消歧的贡献度高于论文本身的文本特征。

4.2.3 论文文本语义表示分析

为评估采用不同语言模型进行论文文本语义表示的消歧效果，本文在实验数据集上分别使用Word2Vec模型、Google的BERT-base-uncased基础预训练模型和BERT-base-multilangual-uncased多语言预训练模型、哈工大的中文BERT-wwm-Chinese预训练模型以及科学论文SciBERT模型开展实验，对比结果如表5所示。从Word2Vec, BERT-base-multilangual-uncased, BERT-wwm-

表 3　对比实验结果(%)

姓名	本文方法			匿名图网络嵌入[7]			多维网络嵌入[24]			基于规则的方法		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
Tao Huang	**98.29**	**97.20**	**97.74**	87.32	82.90	85.05	90.12	86.20	88.46	73.53	29.33	41.93
Haibo Li	**87.57**	**93.77**	**90.56**	86.58	87.22	86.89	80.06	78.68	79.82	42.77	58.06	49.25
Ming Li	**92.25**	**83.30**	**87.55**	60.01	62.95	61.44	74.64	70.03	72.65	15.78	79.62	26.33
Wei Li	73.37	65.61	69.27	78.70	70.37	74.30	**80.14**	77.35	**78.24**	52.16	**83.07**	64.08
Jia Liu	56.82	84.03	67.8	**92.59**	**84.03**	**88.11**	88.23	78.65	82.78	58.62	60.01	73.91
Jie Liu	74.29	53.61	62.28	**100.0**	**100.0**	**100.0**	80.12	70.30	76.24	72.20	64.26	68.0
Jing Liu	**92.63**	**66.91**	**77.69**	76.96	54.26	63.65	78.11	56.47	64.94	35.47	60.16	44.63
Jun Liu	91.45	74.57	82.15	**98.12**	95.07	**96.61**	96.42	80.20	88.23	59.72	**96.63**	73.81
Yun Liu	**99.33**	**99.72**	**99.52**	97.30	87.35	92.06	91.84	82.42	86.21	48.39	62.36	54.49
Bin Wang	91.13	47.09	62.09	81.16	31.51	45.39	**94.39**	49.20	**64.69**	49.38	**61.51**	54.78
Yunshan Wang	85.8	81.92	83.82	**93.01**	**90.21**	**91.59**	87.01	84.18	86.53	51.30	60.01	57.81
LiangWang	**82.92**	**76.06**	**79.34**	50.77	57.01	53.71	62.65	60.27	61.38	20.74	66.77	31.65
Lin Wang	62.75	82.53	71.30	63.73	86.54	73.41	**66.19**	88.20	**76.69**	64.23	**90.42**	75.11
Gang Xiong	**99.00**	**89.21**	**93.85**	98.43	84.30	90.82	94.33	89.21	92.30	83.70	42.35	56.24
Jun Yang	**94.42**	**83.46**	**88.60**	74.35	71.84	73.08	78.81	75.25	77.59	20.37	32.61	25.07
PengZhang	**75.36**	**70.76**	**72.99**	48.74	40.60	44.30	56.30	58.43	57.38	16.09	62.81	25.62
Tao Zhang	**99.02**	**89.50**	**94.02**	80.12	73.04	76.41	88.23	86.52	87.11	42.99	29.06	34.67
Xu Zhao	89.97	66.54	76.50	**99.22**	**95.54**	**97.35**	90.67	86.15	88.46	61.38	81.91	70.18
Feng Zhao	**92.25**	**89**	**90.59**	86.14	78.16	81.96	83.92	80.54	82.78	27.49	62.33	38.15
Ming Zhu	81.57	84.90	83.20	81.57	84.90	83.20	**83.12**	**85.29**	**84.22**	58.29	41.63	48.57
平均	**86.01**	**78.98**	**81.54**	81.75	75.88	77.97	82.27	76.18	78.84	47.73	59.29	48.96

表 4　组件聚类结果对比(%)

	Avg-Pre	Avg-Rec	Avg-F1
对论文文本语义表示向量聚类	52.20	66.26	57.03
图卷积网络计算合作者/机构关系进行聚类	76.49	78.38	75.76
综合	86.01	78.98	81.54

Chinese 3个模型针对原始中英文论文题目和关键词进行语义表示的实验可以看出BERT-base-multilangual的执行效果较好。而对比BERT-base-uncased和SciBERT模型针对论文原有和翻译后的英文题目、关键词的实验显示SciBERT模型的执行效果较好，并且整体性能最优。

此外，为评估论文文本属性对消歧效果的影响，分别针对题目、关键词、摘要、出版物名称的联合使用进行实验，实验结果如表6所示。从实验结果可以发现利用题目和关键词的消歧效果要高于其他，尽管摘要存在更多的信息，但也同时带来了干扰，而题目和关键词含有的领域主题词密度更高，因而效果更好。

4.2.4　特征权重分析

在图卷积神经网络训练中，最终节点嵌入表示向量融合了合作关系、机构关联、语义表示3个方向的节点嵌入表示，如公式(4)。其中每一个方向的向量使用权重参数β_1，β_2和β_3。通过组合不同的特征权重进行对比实验，选择一组最优组合，实验

表 5　使用不同文本语义表示模型的消歧结果对比(%)

模型	Avg-Pre	Avg-Rec	Avg-F1	文本语言
Word2Vec	47.73	65.24	51.22	中英文混合
BERT-base-multilangual-uncased	45.68	67.07	54.35	中英文混合
BERT-wwm-Chinese	48.07	63.88	51.66	中英文混合
BERT-base-uncased	46.60	71.58	55.34	英文
SciBERT	52.20	66.26	57.03	英文

结果如图4所示，当$\beta_1 = 0.001$，$\beta_2 = 1$和$\beta_3 = 3$时消歧效果最好。

在调参过程中可以发现β_1相比其他两个权重对性能的影响更大。如图5所示，在Liang Wang, Tao Zhang, Ming Li和Feng Zhao 4个子任务，β_1从1下降到0.01过程中查准率提升明显，并且在0.001时达到最优。β_1参数跨度较大的原因在于同一个人的合作者较少或存在缩写名称相同的合作者，查准率降低。而式(5)中损失函数调和参数lam的对比实验结果如图6所示，当lam = 0.6时性能最优。

5 结束语

本文提出一种基于图卷积半监督学习的论文作者同名消歧方法，利用图卷积神经网络在图半监督学习方面的优势解决作者同名消歧问题。该方法一方面利用了论文中表示研究主题的题目和关键字计算论文语义表示向量，另一方面利用论文的作者和机构信息构建论文之间关系网络，将论文语义表示向量和网络邻接矩阵作为图卷积神经网络的输入开展半监督学习，通过合作网络中采集的伪标签获得正样本集和负样本集计算每次训练的损失函数，经过深度学习获得论文节点的嵌入表示向量进行聚类。

通过对比实验可以发现本文方法相比其他方法可以取得更好的消歧效果，在不同论文规模和消歧类别上的适应能力和处理能力更强。本文还对比了文本特征语义向量计算、论文合作/机构关系网络图卷积学习两个组件的贡献，不同语义表示模型、文本元素以及特征权重对消歧效果的影响，探究本文方法各个组成部分的影响，以期为进一步研究和改进提供参考。

本文研究也存在一些不足：(1)由于本文研究面向科教大数据融合的具体应用，尚未在常用测试集如DBLP, Arnetminer, CiteSeerX等上开展实验；(2)本文方法仅从合作网络随机采集伪标签，伪标签的学习以及影响分析有待于进一步研究和实验；(3)本文方法的执行效率有待进一步优化，以实际应用于大规模数据融合中。这几方面也成为接下来研究的重点。

表 6 针对不同文本内容的消歧结果对比(%)

	Avg-Pre	Avg-Rec	Avg-F1
题目、关键词	52.20	66.26	57.03
题目、关键词、出版物名称	49.43	67.25	55.16
题目、关键词、摘要	50.75	58.87	52.98

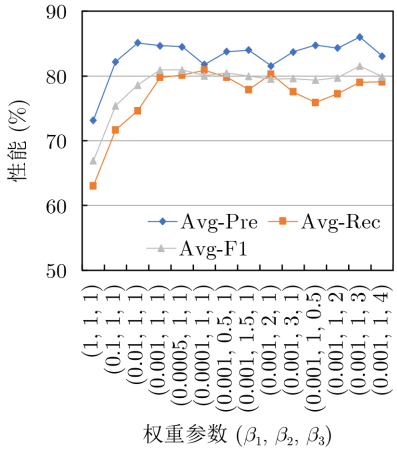


图 4 权重组合性能对比

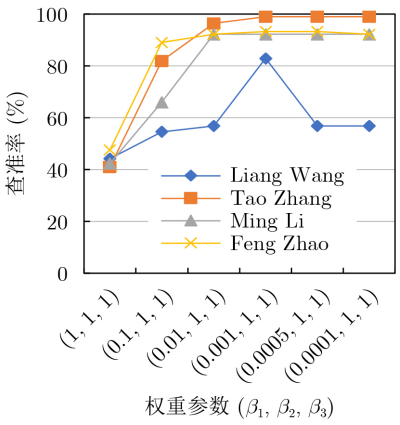


图 5 β_1权重调节查准率对比

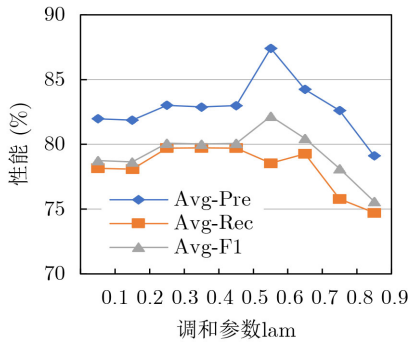


图 6 调和参数lam对比实验结果

参 考 文 献

[1] ORCID. What is ORCID[EB/OL]. https://www.lanl.gov/library/scholarly/orcid.php.

[2] Thomson Reuters Company. What is ResearcherID?[EB/OL]. https://libanswers.lib.xjtlu.edu.cn/faq/240918, 2020.

[3] HAN Hui, GILES L, ZHA Hongyuan, *et al*. Two supervised learning approaches for name disambiguation in author citations[C]. Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tuscon, USA, 2014: 296–305. doi: 10.1145/996350.996419

[4] MALIN B. Unsupervised name disambiguation via social

network similarity[C]. Proceedings of the SIAM Workshop on Link Analysis, Counterterrorism, and Security, Newport Beach, USA, 2005: 93–102.

[5] HAN Hui, ZHA Hongyuan, and GILES C L. Name disambiguation in author citations using a K-way spectral clustering method[C]. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, USA, 2005: 334–343. doi: 10.1145/1065385.1065462.

[6] ZHANG Yutao, ZHANG Fanjin, YAO Peiran, *et al*. Name disambiguation in aminer: Clustering, maintenance, and human in the loop[C]. The Twenty-Forth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 2018: 1002–1011. doi: 10.1145/3219819.3219859.

[7] ZHANG Baichuan and AL HASAN M. Name disambiguation in anonymized graphs using network embedding[C]. The 2017 ACM on Conference on Information and Knowledge Management, Singapore, 2017: 1239–1248. doi: 10.1145/3132847.3132873.

[8] 盖杉, 鲍中运. 基于改进深度卷积神经网络的纸币识别研究[J]. 电子与信息学报, 2019, 41(8): 1992–2000. doi: 10.11999/JEIT181097.

GAI Shan and BAO Zhongyun. Banknote recognition research based on improved deep convolutional neural network[J]. *Journal of Electronics & Information Technology*, 2019, 41(8): 1992–2000. doi: 10.11999/JEIT181097.

[9] 卢俊言, 贾宏光, 高放, 等. 语义分割网络重建单视图遥感影像数字表面模型[J]. 电子与信息学报, 2021, 43(4): 974–981. doi: 10.11999/JEIT200031.

LU Junyan, JIA Hongguang, GAO Fang, *et al*. Reconstruction of digital surface model of single-view remote sensing image by semantic segmentation network[J]. *Journal of Electronics & Information Technology*, 2021, 43(4): 974–981. doi: 10.11999/JEIT200031.

[10] 孙晓, 彭晓琪, 胡敏, 等. 基于多维扩展特征与深度学习的微博短文本情感分析[J]. 电子与信息学报, 2017, 39(9): 2048–2055. doi: 10.11999/JEIT160975.

SUN Xiao, PENG Xiaoqi, HU Min, *et al*. Extended multi-modality features and deep learning based microblog short text sentiment analysis[J]. *Journal of Electronics & Information Technology*, 2017, 39(9): 2048–2055. doi: 10.11999/JEIT160975.

[11] 郑睿刚, 陈伟福, 冯国灿. 图卷积算法的研究进展[J]. 中山大学学报:自然科学版, 2020, 59(2): 1–14. doi: 10.13471/j.cnki.acta.snus.2020.02.001.

ZHENG Ruigang, CHEN Weifu, and FENG Guocan. A concise survey on graph convolutional networks[J]. *Acta Scientiarum Naturalium Universitatis Sunyatseni*, 2020, 59(2): 1–14. doi: 10.13471/j.cnki.acta.snus.2020.02.001.

[12] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5): 755–780. doi: 10.11897/SP.J.1016.2020.00755.

XU Bingbing, CEN Keting, HUANG Junjie, *et al*. A survey on graph convolutional neural network[J]. *Chinese Journal of Computers*, 2020, 43(5): 755–780. doi: 10.11897/SP.J.1016.2020.00755.

[13] 葛尧, 陈松灿. 面向推荐系统的图卷积网络[J]. 软件学报, 2020, 31(4): 1101–1112. doi: 10.3969/j.issn.1000-9825.2020.04.016.

GE Yao and CHEN Songcan. Graph convolutional network for recommender systems[J]. *Journal of Software*, 2020, 31(4): 1101–1112. doi: 10.3969/j.issn.1000-9825.2020.04.016.

[14] 王鑫, 李可, 宁晨, 等. 基于深度卷积神经网络和多核学习的遥感图像分类方法[J]. 电子与信息学报, 2019, 41(5): 1098–1105. doi: 10.11999/JEIT180628.

WANG Xin, LI Ke, NING Chen, *et al*. Remote sensing image classification method based on deep convolution neural network and multi-kernel learning[J]. *Journal of Electronics & Information Technology*, 2019, 41(5): 1098–1105. doi: 10.11999/JEIT180628.

[15] HUANG Jian, ERTEKIN S, and GILES C L. Efficient name disambiguation for large-scale databases[C]. 10th European Conference on Principles and Practice of Knowledge Discovery, Berlin, Germany, 2006: 536–544. doi: 10.1007/11871637_53. doi: 3.

[16] YOSHIDA M, IKEDA M, ONO S, *et al*. Person name disambiguation by bootstrapping[C]. The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, Switzerland, 2010: 10–17. doi: 10.1145/1835449.1835454.

[17] ZHU Jia, WU Xingcheng, LIN Xueqin, *et al*. A novel multiple layers name disambiguation framework for digital libraries using dynamic clustering[J]. *Scientometrics*, 2018, 114(3): 781–794. doi: 10.1007/s11192-017-2611-8.

[18] FAN Xiaoming, WANG Jianyong, PU Xu, *et al*. On graph-based name disambiguation[J]. *Journal of Data and Information Quality*, 2011, 2(2): 10. doi: 10.1145/1891879.1891883.

[19] TANG Jie, FONG A C M, WANG Bo, *et al*. A unified probabilistic framework for name disambiguation in digital library[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2012, 24(6): 975–987. doi: 10.1109/TKDE.2011.13.

[20] HERMANSSON L, KEROLA T, JOHANSSON F, *et al*. Entity disambiguation in anonymized graphs using graph kernels[C]. The 22nd ACM International Conference on Information & Knowledge Management, San Francisco,

USA, 2013: 1037–1046. doi: 10.1145/2505515.2505565.

[21] KIPF T N and WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609.02907, 2016.

[22] DEVLIN J, CHANG Mingwei, LEE K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. https://arxiv.org/pdf/1810.04805.pdf, 2019.

[23] BELTAGY I, LO K, and COHAN A. SciBERT: A pretrained language model for scientific text[C]. The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kang, China, 2019: 3615–3620.

[24] XU Jun, SHEN Siqi, LI Dongsheng, *et al*. A network-embedding based method for author disambiguation[C]. The 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, 2018: 1735–1738. doi: 10.1145/3269206.3269272.

盛晓光：男，1989年生，博士生，研究方向为教育数据挖掘、人工智能.

王 颖：女，1982年生，副研究馆员，研究方向为知识组织与知识挖掘.

钱 力：男，1981年生，研究馆员，研究方向为大数据与机器智能.

王 颖：女，1969年生，教授，研究方向为数字信号处理、教育数据挖掘.

责任编辑：陈 倩