文章编号：2095-2163(2024)03-0054-07 中图分类号：TP391 文献标志码：A

基于图神经网络的姓名消歧算法

汤哲冲，方志坚，贾子杰
（浙江理工大学 信息科学与工程学院，杭州 310018）

摘 要：由于语言的不同，中国作者在发表外文文献时很容易出现作者重名的问题，导致许多重名学者发表的学术文献无法很好地区分开来。针对这一问题，本文提出了一种基于图神经网络的姓名消歧算法，解决外文文献中的中国作者同名问题。首先，基于待消歧文献的属性特征及其关系构建异质学术关系网络，对文献进行表示学习；然后再进行聚类消歧。由于文献属性特征之间具有强关联性，本文在原有文献关系的基础上引入了消歧特征对来丰富节点关系类型。实验结果表明，本文提出算法的性能明显优于其他对比方法，有更好的消歧性能。

关键词：姓名消歧；异质学术关系网络；消歧特征对

Graph neural network-based name disambiguation algorithm

TANG Zhechong, FANG Zhijian, JIA Zijie
(School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Due to the difference in languages, Chinese authors are more likely to share the same name to cause confusion when publishing foreign literature. It is a challenge to differentiate academic publications by multiple authors with the same name. To address the situation, this paper proposes the name disambiguation algorithm based on graph neural networks to solve the foreign literature's Chinese author name ambiguity problem. Specifically, the paper first constructs a heterogeneous academic relational network based on the attribute features of the documents to be disambiguated and their relationships to learn the representations of the documents, and then perform cluster disambiguation. Due to the significant association between document attribute features, the paper introduces disambiguation feature pairs based on the original document relationship to enrich the node relationship types. The experimental results show that the performance of the proposed algorithm in this paper is significantly better than other comparative methods with better disambiguation performance.

Key words: name disambiguation; heterogeneous academic relationship network; disambiguation feature pairs

0 引 言

随着科学技术的飞速发展，科研人员数量日渐增长，国内作者发表外文文献的数量也在不断增多。中国作者在发表外文文献的过程中，姓名主要以拼音的姓氏发表，这就导致不同作者重名现象日益增多。而学术文献数据库主要以人名作为重要的检索条件，作者重名现象严重影响了文献检索的效率和准确度，更加精确的姓名消歧算法对分析各个领域科研人员的组成、研究跨领域跨学科的科研人员以及搜索引擎的优化都有十分重要的意义，因此同名作者的消歧任务已经变得尤为迫切[1]。

普遍意义上的姓名歧义具有 2 种含义。一种是一人多名，即同一位作者可能会有多个姓名，如原名、曾用名、笔名、笔误等；二是一名多人，即不同的作者拥有相同的姓名，这种现象最为普遍，尤其是在作者发表的外文文献中，由于存在拼写不规范、复姓、双名连写、多音字等问题，中文姓名变换为英文后更容易重名。一名多人问题则涉及相同姓名、相同研究领域、甚至是相同研究机构中不同的人。消歧只能依赖于除姓名以外的其他属性特征，如邮箱、合作者等来作为消歧特征项，但海量文献数据中的信息往往不完整，存在特征稀疏、覆盖面不全的问题，降低了算法的精确度。因此一些专家学者提出

基金项目：浙江省科学技术厅“领雁”研发攻关计划项目(2022C01220)。

作者简介：汤哲冲(1997-)，男，硕士研究生，主要研究方向：数据挖掘与分析、智能信息处理；贾子杰(1998-)，男，硕士研究生，主要研究方向：智能信息处理、信息研究。

通讯作者：方志坚(1983-)，男，博士，工程师，主要研究方向：数据挖掘、信息研究。Email：hptnt@ zstu.edu.cn

收稿日期：2023-03-13

了利用图结构模型的网络拓扑结构和节点属性信息构建异质信息网络，解决同名作者消歧问题。而现有的基于图模型的方法主要考虑了文献之间的合著关系和引用关系等简单关系，但异质信息网络包含多种类型的节点和关系，简单的异质信息网络不能有效地捕捉文献数据中丰富的语义和结构信息，且文献的属性特征之间具有强关联性。例如，2 篇待消歧文献的作者姓名和研究机构相同且研究领域相似，那么几乎可以认为这 2 篇文献属于同一作者[2]。

本文提出了一种基于图神经网络的表示学习方法。首先，将每篇待消歧文献作为网络的节点；其次，依据文献特征之间的强关联性建立消歧特征对来构建边；最后，使用一个无监督的图自动编码器来进行文献的嵌入表示，进而通过聚类算法实现同名作者消歧。实验结果表明，本文的解决方案与对比方案具有明显的优越性。

1　相关工作

现有的姓名消歧方法主要分为 3 类：基于有监督、基于无监督和基于图的方法。基于有监督的方法主要依赖人为标注的训练集对同名作者文献进行分类，Tran 等学者[3]基于深度神经网络对文献特征进行编码，得到特征向量的低维嵌入。Yoshida 等学者[4]提出一种分步聚类的方法，第二阶段的聚类特征来源于第一阶段的聚类结果。Han 等学者[5]利用支持向量机（SVM）和朴素贝叶斯的方法进行同名作者消歧。基于无监督的方法主要通过人为提取文献特征来训练分类器进行分类，Zhang 等学者[6]提出一种结合全局和局部信息的表示学习方法，通过图自动编码器来增强文献节点的嵌入表示，但是忽略了文献与特征之间的强关联性。Kim 等学者[7]在不考虑语义特征的情况下提取结构感知特征和全局特征来解决成对的分类问题。基于图的方法主要通过构建图结构模型对文献进行表示学习。GHOST 提出一种名为 ghost 的姓名消歧框架[8]，基于文献的合著者信息构建无向图，并利用亲和传播算法进行聚类，但利用的有效特征过少。Zhang 等学者[9]采用一种新的表示学习模型，模型包括合作者合著网络（author-author）、文献-作者网络（author-paper）和文献网络（paper-paper）融合生成文献节点的低维向量表示。Yu 等学者[10]使用和 Zhang 等学者[9]相同的模型，利用 LINE[11]、DeepWalk[12]和成对相似性排序分别对 3 个网络进行表示学习。Qiao 等学者[13]基于文献之间的共同作者（CoAuthor）、共同出版物（CoVenue）和共同标题（CoTitle）关系构建图结构模型，并采用异质图卷积网络得到文献节点的低维嵌入。

上述方法或标注成本高、或严重依赖于分词精准度、或利用的有效特征过少、或只考虑了少量类型的节点和边。本文结合上述方法的优点，提出了一种节点和边类型更为丰富的网络表示方法。

2　准备工作

定义 1　同名作者消歧问题　给定待消歧姓名 a，$D^a=\{d_1^a,d_2^a,\cdots,d_n^a\}$ 表示一组重名作者 a 的待消歧文献集合。每篇待消歧文献可以由研究机构、出版物、关键词等长度不同的论文属性特征表示[14]。姓名消歧的目的是通过映射函数 $\phi(x)$ 将待消歧文献集合划分为若干个互不相交的同名作者簇，使得每个同名作者簇的文献都尽可能属于同一作者。即 $\phi(D^a)\rightarrow D^{ab}$，其中 $D^{ab}=(D_1^{ab},D_2^{ab},\cdots,D_n^{ab})$。

定义 2　异质信息网络　异质信息网络可以表示为 $G=\{V,E\}$，其中 V 和 E 分别表示网络中的节点和边，对应实体及其关系，如图 1 所示。异质信息网络包含多种类型的节点（对象）和边（关系），基于节点映射函数 $\phi(v)$ 和边映射函数 $\phi(e)$ 可以将节点及其复杂关系映射到一个低维空间，生成节点的嵌入表示。

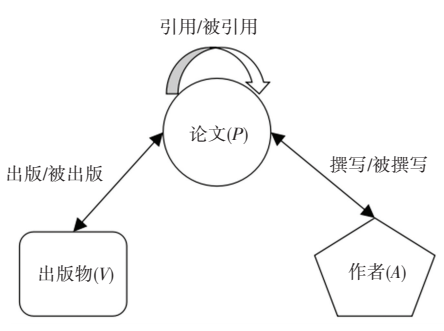

图 1　异质信息网络

Fig. 1　Heterogeneous information network

定义 3　异质学术关系网络　现有的异质文献信息网络常将文献作为网络节点，基于邮箱、合著者、标题、关键词、出版物等常见的节点关系类型构建边[15]。常见的节点关系类型定义如下。

（1）CoEmail：由于升学或者工作的原因，一个学者可能具有多个邮箱，用 E 来表示学者的邮箱集。对于给定的 2 篇文献 p_i 和 p_j，如果含有相同的邮箱、即 $E_i\cap E_j\neq\varnothing$，那么认为 2 篇文献之间存在 CoEmail 关系。

（2）CoAuthor：每篇文献通常都不只有一位作者，将每篇文献 $p_i=\{a_i^o,a_i^1,\cdots,a_i^n\}$ 的第一作者 a_i^o

定义为主要作者,其他 $A_i = \{a_i^1, a_i^2, \cdots, a_i^n\}$ 定义为次要作者或合著者。如果 2 篇文献 p_i 和 p_j 含有相同的合著者 $A_i \cap A_j \neq \emptyset$,那么认为 2 篇文献之间存在 CoAuthor 关系。

(3)CoVenue:每个学者通常都会有固定的研究领域,有经常发表论文的出版物,如果 2 篇文献在不同的出版物上发表,有很大的概率属于不同的作者。对于给定的 2 篇文献 p_i 和 p_j,如果出版物相同 $V_i = V_j$,那么认为 2 篇文献之间存在 CoVenue 关系。

(4)CoKeyword:文献的关键词能在很大程度上反映该文献研究的主要内容。对于给定的 2 篇文献 p_i 和 p_j,如果 $K_i \cap K_j \neq \emptyset$,那么认为 2 篇文献之间存 CoKeyword 关系。

由于文献属性特征之间存在强关联性。本文基于以下假设:

(1)每个学者通常具有固定的学术关系网络或是合作关系圈,同一学术关系网络内几乎不存在同名作者[16]。由于同一研究机构的学者往往具有相同的学术关系网络,因此研究机构和合著者信息可以协助判断同名作者的 2 篇文献是否属于同一作者。

(2)每个学者具有固定的研究领域,同一位作者几乎不可能短时间内在不同的机构下以第一作者发表文章[17]。出版物和关键词能反映文献研究的大致方向和主要内容,所以出版物、关键词也可以协助判断 2 篇文献是否属于同一作者。

在现有的节点关系类型基础上引入消歧特征对来构建异质学术关系网络,异质学术关系网络如图 2 所示。消歧特征对包括 Co(name+org+CoAuthor)、Co(name+org+CoVenue)、Co(name+org+CoKeyword)。这里将给出研究定义分述如下。

(1)Co(name+org+coauthor):给定待消歧姓名 a 的两篇文献 p_i 和 p_j,如果地址信息相同,且含有 coauthor 关系,那么认为 2 篇文献之间存在 Co(name+org+coauthor)关系。如果 2 篇文献的作者姓名相同、地址信息相同且含有相同的合著者,那么几乎可以认为这 2 篇文献属于同一作者。

(2)Co(name+org+venue):给定同名作者的 2 篇文献 p_i 和 p_j,如果地址信息相同,且含有 CoVenue 关系,那么认为 2 篇文献之间存 Co(name+org+coauthor)关系。如果 2 篇文献的作者姓名相同、地址信息相同且发表在同一出版物上,那么几乎可以认为这 2 篇文献属于同一作者。

(3)Co(name+org+keyword):给定同名作者的两篇文献 p_i 和 p_j,如果地址信息相同,且含有 CoKeyword 关系,那么认为 2 篇文献之间存 Co(name+org+keyword)关系。如果 2 篇文献的作者姓名相同、地址信息相同且含有相同的关键词,那么几乎可以认为这 2 篇文献属于同一作者。

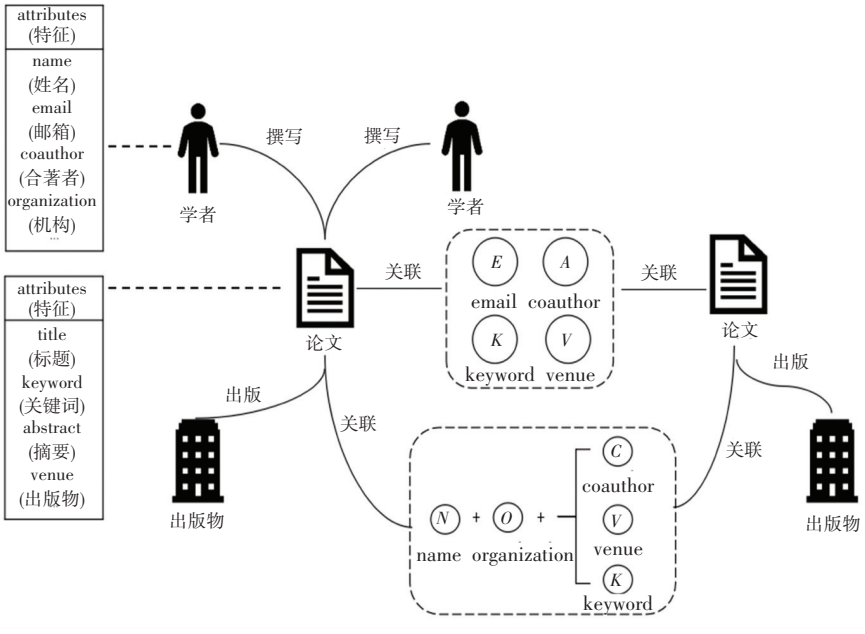

图 2 异质学术关系网络

Fig. 2 Heterogeneous academic network

3 基于关系网络的图模型

姓名消歧模型整体框架如图 3 所示,主要由一人多名问题解决、特征内容嵌入、关系网络构建、关系网络学习这 4 个部分组成。

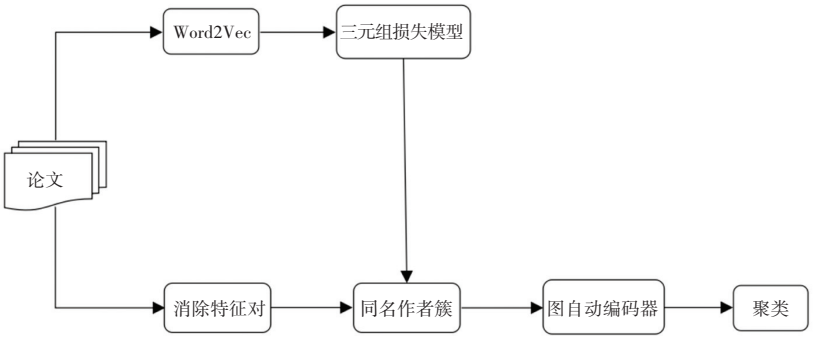

图 3　姓名消歧模型整体框架

Fig. 3　Name disambiguation model

3.1　一人多名问题解决

由于采用了不同的写法顺序和缩写规则，中国作者在发表外文文献时存在一人多名的问题，即同一作者姓名有多种不同的写法[18-19]。对于一人多名问题，现有的方法主要通过判断别名是否为作者姓名的最长公共子序列。但是当作者姓名只存在声母上的区别时，例如“Zeng Qianwen”和“Zheng Qianwen”时，现有方法就无法有效解决该问题。本文提出了一种拼音声母算法来解决此问题，算法的流程如下。

输入　待消歧论文集合 $D^n = \{D^1, D^2, \cdots, D^n\}$

输出　同名作者文献集合 D^i

步骤 1　将每篇文献的作者姓名分别看做一类构成集合 $A = \{a_1, a_2, \ldots, a_n\}$；

步骤 2　统一去除作者姓名的大小写和特殊符号，如逗号、分号、连接符等；

步骤 3　将每个姓名的拼音用一个唯一的汉字对应，如“Zeng”对应“曾”、“Zheng”对应“郑”；

步骤 4　分析作者姓名是拼音全称、还是声母简写，并将拼音全称解析为拼音、拼音对应的声母和拼音对应的汉字；

步骤 5　如果类 a_1 和类 a_2 的作者姓名都为全称，且对应的汉字相同，或者类 a_1 和类 a_2 的作者姓名中含有声母简写，且对应的汉字相同，那么将 a_1 和 a_2 合并为 a_{12}，并把 a_{12} 添加到集合 A 中，同时去除 a_1 和 a_2，否则跳到步骤 7；

步骤 6　如果类集合中类的个数大于 1，则重复步骤 4，步骤 5；

步骤 7　结束聚类。

3.2　特征内容嵌入

给定一篇作者姓名含有歧义的文献 D_i，将 D_i 用长度不同的特征集合表示 $D_i = \{x_1, x_2, \cdots, x_n\}$，其中包括作者姓名、邮箱、合著者、研究机构等；基于 Word2Vec 生成每项特征的嵌入表示，再通过逆文本频率（TF-IDF）依照消歧特征性的强弱对每项特征的嵌入向量进行加权融合，得到每篇待消歧文献的整体嵌入，其数学模型如下：

$$X_i = \sum_{x_m \in D_i} f_m \, x_m \tag{1}$$

其中，x_m 表示文献的属性特征，f_m 表示特征对应的权重系数。

通常情况下 2 篇文献的特征向量越相似，就越有可能是属于同一个作者，但是同一个研究人员往往可能有几种不同的研究方向，而且这些研究方向之间的差异可能很大，那么基于特征向量相似性的方法就无法很好地将这些不同研究领域的文献区分开来。针对这一问题，本文通过标注的数据训练了一个三元组损失网络模型对文献的特征嵌入进行调整，如图 4 所示。

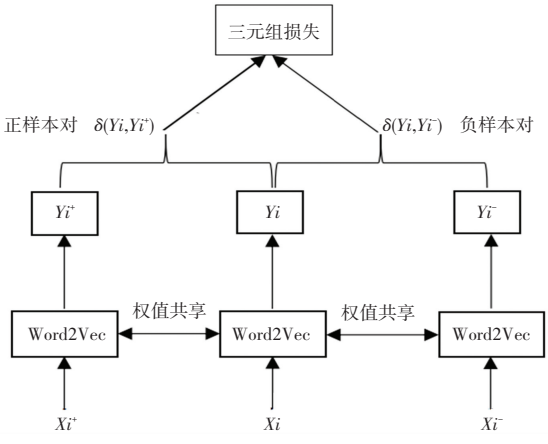

图 4　三元组损失模型

Fig. 4　Model with triplet loss

三元组损失的目的是找到一个准确的边界距离阈值 m 来区分正样本对和负样本对，可以让正样本对之间的距离愈发接近，负样本对之间的距离愈发远离。

这样就可以有效区分同一作者的不同类型文献，其损失函数 $\zeta_{d_{ij}}$ 可以定义为：

$$\zeta_{d_{ij}} = \sum_{(i,j,k),y_{ij}=1,y_{ik}=0} [d_{ij} - d_{ik} + m]_+ \quad (2)$$

其中，d_{ij} 表示文献 i 与文献 j 之间的距离，通常采用欧式距离计算；$y_{ij}=1$ 表示 2 篇文献属于同一作者，即是一对正样本对；$y_{ik}=0$ 表示 2 篇文献属于不同作者，即是一对负样本对；m 是一个固定的边界距离常量。

3.3 关系网络构建

由于邮箱具有强消歧特征性，具有相同邮箱的同名作者指向同一人物实体。本文基于文献的共同通讯作者来构建学术关系网络，构建消歧特征对来对同一学术关系网络内的同名作者文献进行消歧，将高度相似的文献聚集起来，缩小所构建图的规模。算法的伪代码如下。

输入 同名作者文献集合 D_n

输出 同一作者的文献集合 D_{mn}

对于 D_n 中的任意 2 篇文献 a_1 a_2：

If 邮箱相同：

判定为同一人，并分裂到更小的聚类 D_{mn}

Else if 地址机构相同：

If 出版物、合作者、关键词至少有一个相同：

判定为同一人，并分裂到更小的聚类 D_{mn}

Else：

继续聚类

Else：

If 合著者、出版物、关键词都相同：

判定为同一人，并分裂到更小的聚类 D_{mn}

Else：

继续聚类

3.4 关系网络学习

本文使用一个基于无监督的图自动编码器来学习异质文献关系网络中节点的分布式表示，然后对节点之间的链接关系进行预测。图自动编码器的模型结构如图 5 所示。

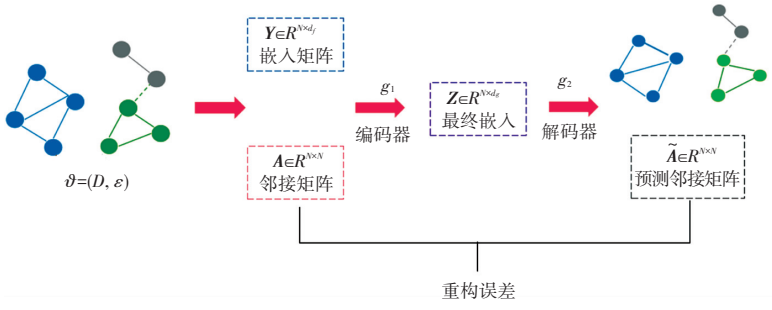

图 5 图自动编码器模型结构

Fig. 5 Graph autoencoders structure

图自动编码器由节点编码器模型 $Z = g_1(\boldsymbol{Y}, \boldsymbol{A})$ 和边解码器模型 $\tilde{\boldsymbol{A}} = g_2(Z)$ 两部分组成，其中 $\boldsymbol{Y} = [\boldsymbol{y}_1^{\mathrm{T}}, \boldsymbol{y}_2^{\mathrm{T}}, \cdots, \boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}}$ 表示节点 D 的嵌入矩阵，$\boldsymbol{A} \in R^{N \times N}$ 是图 G 的邻接矩阵，主要用来表示节点之间的关系，$\tilde{\boldsymbol{A}}$ 是模型预测的邻接矩阵，$\boldsymbol{Z} = [\boldsymbol{Z}_1^{\mathrm{T}}, \boldsymbol{Z}_2^{\mathrm{T}}, \cdots, \boldsymbol{Z}_n^{\mathrm{T}}]^{\mathrm{T}}$ 是节点嵌入矩阵。目标是使预测的邻接矩阵 $\tilde{\boldsymbol{A}}$ 与原始的邻接矩阵 $\boldsymbol{A}$ 之间的重构误差最小。

编码部分，图自动编码器使用了一个 2 层的图卷积神经网路（GCN）作为编码器来得到节点的嵌入表示。编码器 g_1 可以定义，对此可写为：

$$g_1(\boldsymbol{Y}, \boldsymbol{A}) = \boldsymbol{A} Relu(\boldsymbol{A}\boldsymbol{Y}\boldsymbol{W}_0)\ \boldsymbol{W}_1 \quad (3)$$

其中，$\boldsymbol{A}$ 表示对称归一化的邻接矩阵，即 $\boldsymbol{A} = \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$，$\boldsymbol{D}$ 表示图 G 的节点度矩阵 $Relu(.) = \max(0, \cdot)$，$\boldsymbol{W}_0$ 和 $\boldsymbol{W}_1$ 是图神经网络第一层和第二层的参数。

解码部分，图自动编码器采用了内积（inner-product）的方式来重构原始图的结构信息。解码器 g_2 可以定义为：

$$g_2(\boldsymbol{Z}) = sigmoid(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}) \quad (4)$$

节点 D_i 和 D_j 之间存在边的概率为：

$$P(\tilde{\boldsymbol{A}}_{ij} = 1 \mid \boldsymbol{Z}_i, \boldsymbol{Z}_j) = sigmoid(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{Z}_j) \quad (5)$$

采用交叉熵作为损失函数，可由式（6）进行描述：

$$\zeta_g = -\sum_{D_i, D_j} \boldsymbol{A}_{ij} \log_p \tilde{\boldsymbol{A}}_{ij} \quad (6)$$

最后，基于图自动编码器可以得到包含文献属性特征信息和文献间关系信息的潜在变量 $\boldsymbol{Z} = [z_1, z_2, \cdots, z_n]$，将其作为文献新的嵌入表示。新的嵌入表示为：

$$R = \sum_{i \in p} N_i \cdot S_i \quad (7)$$

4　实验

4.1　数据集

本文使用 Aminer 数据集作为实验的测试集，该数据集包括 12 798 位作者的 70 285 篇文献，每篇文献记录中包含了文献的 id、标题、关键词、合作者等特征信息，文献数据格式见表 1。本文选择其中的 100 个待消歧名称进行实验测试，数据集样例见表 2。表 2 中给出了实验数据集中 10 个待消歧作者的文献数量和真实作者数量。

表 1　文献数据格式

Table 1　Document data format

特征	类型	含义	示例
Id	String	论文 ID	5b5433eae1cd8e4e15039b23
Title	String	标题	Data mining: concepts and techniques
Authors.name	String	作者姓名	Yanjun Zhang
Authors.org	String	作者单位	College of Chemical Engineering
Authors.id	String	作者 ID	5b5433f1e1cd8e4e1511294d
Venue	String	会议/期刊	Polymer Journal
Year	Int	发表年份	2018
Keywords	List of string	关键词	["The fore-device reading the data", "The remote recorder"]
Abstract	String	摘要	Our ability to generate...

表 2　数据集样例

Table 2　Datasets samples

作者姓名	文献数量	真实作者数量
Xu Xu	555	28
Rong Yu	273	9
Yong Tian	262	16
Lu Han	366	24
Lin Huang	553	31
Kexin Xu	203	3
Wei Quan	174	7
Tao Deng	306	15
Hongbin Li	286	15
Hua Bai	351	11

4.2　实验结果及分析

为了验证算法的性能，将本文算法与 Zhang 等学者[9]、GHOST[8]、Louppe 等学者[20]、Zhang 等学者[6]提出的几种主要的姓名消歧算法进行对比实验。为了公平地与基线方法做比较，聚类个数（即每个待消歧名字的实际作者人数）均使用预先指定的真实值。算法的评价指标为成对准确率（*pre*）、成对召回率（*Rec*）和成对 *F*1 值，实验结果见表 3。

实验结果表明，在给定真实正确的聚类个数的前提下，本文提出的方法在 *F*1 值上优于其他基线方法，验证了本文提出的基于图神经网络的姓名消歧算法的有效性。

表 3　算法性能对比

Table 3　The algorithms performance comparison

姓名	Zhang 等学者[6]			Zhang[9]			GHOST[8]			Louppe 等学者[20]			Our method		
	Prec	*Rec*	*F*1	*Prec*	*Rec*	*F*1	*Prec*	*Rec*	*F*1	*Prec*	*Rec*	*F*1	*Prec*	*Rec*	*F*1
XuXu	74.18	45.86	56.68	48.16	41.87	44.80	61.34	21.79	32.15	22.55	64.40	33.40	71.44	58.89	64.56
RongYu	89.13	46.51	61.12	65.48	40.85	50.32	92.00	36.41	52.17	38.85	91.43	54.53	86.91	47.71	61.61
Yong Tian	76.32	51.95	61.82	70.74	56.85	63.04	86.94	54.58	67.06	32.08	63.71	42.67	68.81	50.64	58.34
Lu Han	51.78	28.05	36.39	47.88	20.62	28.82	69.72	17.39	27.84	30.25	46.65	36.70	60.53	30.36	40.45
Lin Huang	77.10	32.87	46.09	71.84	34.17	46.31	86.15	17.25	28.74	24.86	71.32	36.87	86.71	53.32	66.04
Kexin Xu	91.37	98.64	94.87	90.02	82.47	86.08	92.90	28.52	43.64	91.26	98.35	94.67	91.05	98.62	93.78
Wei Quan	53.88	39.02	45.26	64.45	47.66	54.77	86.42	27.80	42.07	37.86	63.41	47.41	45.38	34.16	39.01
Tao Deng	81.63	43.62	56.86	53.04	29.89	38.23	73.33	24.50	36.73	40.46	51.38	45.27	86.23	45.77	59.80
Hongbin Li	77.20	69.21	72.99	54.66	53.05	53.84	56.29	29.12	38.39	19.48	85.96	31.77	66.93	86.08	75.31
Hua Bai	71.49	39.73	51.08	58.58	35.90	44.52	83.06	29.54	43.58	36.39	41.33	38.70	69.05	33.19	44.83
平均值	77.96	63.03	67.79	70.63	59.53	62.81	81.62	40.43	50.23	57.09	77.22	63.10	78.03	63.56	**70.06**

4.3 消歧特征对分析

为了研究消歧特征对对模型性能的影响,本文将消歧特征对 Co(name+org+coauthor)、Co(name+org+venue)和 Co(name+org+keyword)依次添加到模型进行实验,同时对模型的性能进行评估,实验结果如图 6 所示。由图 6 可见与 Aminer(基线)方法相比,不同的消歧特征对对模型的性能都会有提升,在 3 个消歧特征对中,Co(name+org+coauthor)对模型性能提升得最多。

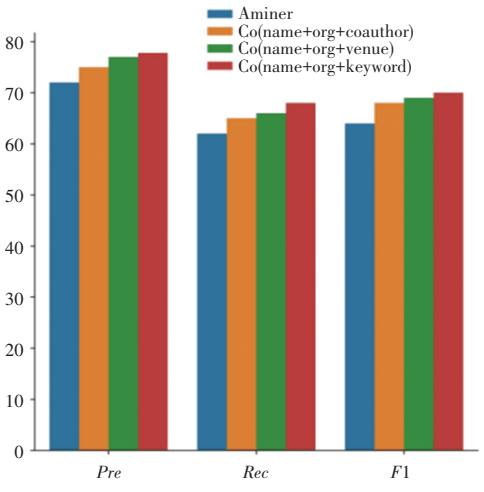

图 6 不同消歧特征对模型性能的贡献

Fig. 6 Contribution of different disambiguation features

5 结束语

本文基于图神经网络提出了一种新的姓名消歧算法。首先,将每篇待消歧文献作为图网络中的节点;基于文献属性特征之间的强关联性建立消歧特征对来构建关系网络;再通过图自动编码器来学习文献的分布式表示,有效利用异质文献网络丰富的语义和结构信息,得到更为准确的文献嵌入表示。实验结果表明,本文提出的方法比目前几种最主要的姓名消歧方法的性能要好。

参考文献

[1] 吴柯烨,闵超,孙建军,等. 面向特定科研任务的著者姓名消歧方法[J]. 情报学报,2021,40(7):734-744.

[2] 展金梅,陈君涛. 基于聚类的人名消歧研究综述[J]. 现代信息科技,2019,3(10):88-91.

[3] TRAN H N, HUYNH T, DO T. Author name disambiguation by using deep neural network[C]// Asian Conference on Intelligent Information and Database Systems. Cham: Springer, 2014: 123-132.

[4] YOSHIDA M,IKEDA M,ONO S,et al. Person name disambiguation by boot strapping [C]//Proceedings of the 33rd International ACMSIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland:ACM, 2010:10-17.

[5] HAN Hui, LEE G, ZHA Hongyuan, et al. Two supervised learning approaches for name disambiguation in author citations [C]// ACM/IEEE Joint Conference on Digital Libraries. Tucson, AZ, USA:IEEE,2004: 296-305.

[6] ZHANG Yutao, ZHANG Fanjin, YAO Peiran, et al. Name disambiguation in AMiner: Clustering, maintenance, and human in the loop [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining.New York:ACM, 2018: 1002-1011.

[7] KIM K, ROHATGI S, GILES C L. Hybrid deep pairwise classification for author name disambiguation[C]//The 28th ACM International Conference. Beijing, China: ACM, 2019: 2369-2372.

[8] FAN Xiaoming, WANG Jianyong, XU Pu, et al. On graph-based name disambiguation[J]. Journal of Data and Information Quality, 2011, 2(2): 10.

[9] ZHANG Baichuan, MOHAMMAD A H. Name disambiguation in anonymized graphs using network embedding[C]//Conference on Information and Knowledge Management. Singapore: ACM, 2017: 1239-1248.

[10] YU Chuanming, YUNCI Z, AOCHEN L, et al. Author name disambiguation with network embedding[J]. Data Analysis and Knowledge Discovery, 2020, 4(2): 48-59.

[11] TANG Jian, QU Meng, WANG Mingzhe, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. New York, USA: ACM, 2015: 1067-1077.

[12] PEROZZI B, AI-RFOU R, SKIENA S. Deepwalk: Online learning of social representations [C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA:ACM,2014: 701-710.

[13] QIAO Ziyue, DU Yi, FU Yanjie, et al. Unsupervised author disambiguation using heterogeneous graph convolutional network embedding[C]//2019 IEEE International Conference on Big Data. Los Angeles, USA: IEEE, 2019: 910-919.

[14] 张立志. 基于图模型和规则的同名作者消歧研究[D]. 呼和浩特:内蒙古大学, 2020.

[15] 沈喆, 王毅, 姚毅凡, 等. 面向学术文献的作者名消歧方法研究综述[J]. 数据分析与知识发现, 2020, 4(8): 15-27.

[16] 邓启平, 陈卫静, 嵇灵, 等. 一种基于异质信息网络的学术文献作者重名消歧方法[J]. 数据分析与知识发现, 2022, 6(4): 60-68.

[17] WANG Xiao, JI Houye, SHI Chuan, et al. Heterogeneous graph attention network[C]//The World Wide Web Conference. New York:ACM, 2019: 2022-2032.

[18] 杨南. 谈中国人名汉语拼音的书写方式[J]. 编辑学报, 1995, 7(3): 156-157.

[19] 闫蓉. 浅析中英文姓名互译中的混乱现象[J]. 长春理工大学学报:社会科学版, 2014, 27(3):127-129.

[20] LOUPPE G, AI-NATSHEH H T, SUSIK M, et al. Ethnicity sensitive author disambiguation using semi-supervised learning [C]// International Conference on Knowledge Engineering and the Semantic Web. Cham: Springer, 2016: 272-287.