文章编号:2095-6134(2022)03-0360-09

基于特征编码和图嵌入的姓名消歧方法*

马莹莹[1,2,3†],吴幼龙[1],唐华[1,2,3]

(1 上海科技大学信息科学与技术学院, 上海 201210; 2 中国科学院上海微系统与信息技术研究所, 上海 200050;
3 中国科学院大学, 北京 100049)

(2020 年 2 月 17 日收稿; 2020 年 4 月 3 日收修改稿)

Ma Y Y, Wu Y L, Tang H. Name disambiguation based on encoding attributes and graph topology[J]. Journal of University of Chinese Academy of Sciences, 2022, 39(3): 360-368. DOI: 10.7523/j.ucas.2020.0019.

摘 要 针对作者姓名歧义问题,提出基于特征编码和图嵌入的作者姓名消歧方法。该方法首先利用 word2vec 模型对文档的属性特征进行编码从而构建文档的表征向量,然后采用图自动编码器将文档关系编码至文档向量中,聚类相似文档。为进一步提升聚类结果的准确性,使用图嵌入的方法将文档关系网络和作者关系网络的拓扑结构信息引入文档向量,进一步聚集相关文档。该方法同时利用文档的属性特征以及多个关系网络的信息,通过无监督学习的方法寻找文档表征向量,实现良好的姓名消歧效果。在真实作者数据集 AMiner 上的测试结果表明,该方法显著优于目前几个其他基于图网络的方法。

关键词 姓名消歧;图神经网络;聚类方法;特征提取;图嵌入

中图分类号:TP391.1 **文献标志码**:A **DOI**:10.7523/j.ucas.2020.0019

Name disambiguation based on encoding attributes and graph topology

MA Yingying[1,2,3], WU Youlong[1], TANG Hua[1,2,3]

(1 *School of Information Science and Technology, ShanghaiTech University, Shanghai* 201210, *China*;
2 *Shanghai Institute of Microsystem & Information Technology, Chinese Academy of Sciences, Shanghai* 200050, *China*;
3 *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract Aiming at solving the problem of author name ambiguity, we propose a novel name disambiguation method based on encoding attributes and graph topology. A word2vec model is used to construct document representation vectors by encoding the attributes of documents. The relationship of documents is then encoded into the document embedding vectors by a graph auto-encoder and similar documents are aggregated. To further improve the accuracy of the clustering results, a graph embedding model is proposed to introduce the document-document network and author-author network topology into the document vectors afterword, thus related papers are moved closer. This method utilizes the information of document attributes and relationship networks at the same time, finds document representation vectors using an unsupervised model and improves the performance of name disambiguation. Experimental results on the real author dataset AMiner show

* 国家自然科学基金(61901267)资助

† 通信作者,E-mail: mayy1@shanghaitech.edu.cn

that our method is superior to several state-of-the-art graph-based solutions.

Keywords　name disambiguation; graph neural network; clustering method; feature extraction; graph embedding

近年来,随着数据信息化程度不断上升,网络数据库容量不断增加,如何在数据库中迅速地搜寻到准确的信息成为亟需解决的问题。由于自然语言具有多义性、复杂性和模糊性的多重特点,因此需要将文本中提到的实体与其知识库中的实体连接起来。实体链接主要是要解决实体间的歧义问题,在网络检索、信息提取和知识库填充等问题中有着广泛的应用。实体语义表达的模糊性和数据容量的日益增加,给实体歧义辨别带来很大的挑战。

实体歧义分为 2 种:一种是多词同义,指多个词语代表同一个意思;另一种是一词多义,是指一个实体名称可以指代多个不同的实体。作者姓名消歧是实体消歧中的一个重要应用,已知同名作者的所有文章集合,需要通过文章的一些属性特征对文章进行聚类,使每一个聚类仅包含一个作者的文章。作者姓名消歧任务在作者文献检索、学术画像分析中有着重要的价值。例如,在学术检索时,研究者需要在文献数据库中寻找名为“Charles”的学者的文献,但是由于“Charles”在数据库中对应着很多不同的实体,系统返回了所有名为“Charles”的作者撰写的文献,这会大大降低文献检索结果的有效性和准确性,从而降低网络搜索的性能。如果将搜索结果分组在一起,则搜索的有效性可以大大提高。另外,当计算学者影响力的时候,需要准确了解每一位学者的文章类型及数目。因此,作者姓名消歧问题是近年来研究者的研究热点之一。

目前,已经有一些文献研究作者姓名消歧问题。一些学者将作者姓名消歧视为分类任务,预测每篇论文的正确标签或预测 2 篇文章是否由同一作者撰写。分类任务需要大量标签,所以这类任务通常是有监督的。

例如,Wang 等[1]提出基于增强树的分类方法,通过文档的标题、作者、机构、摘要等属性判断 2 篇文章是否由同一作者撰写。深度神经网络模型[2]也被用于提取文档属性特征进行分类。其他一些方法利用了外部数据。如 Han 等[3]提出朴素贝叶斯概率生成模型和支持向量机模型并将这 2 种方法分别应用于从 Web 收集的数据和 DBLP 数据库。

另外一些工作采用无监督的聚类方法。无监督的姓名消歧任务是将文献分为几个簇,使得每个簇仅包括由一个作者所撰写的文献。

Cen 等[4]通过优化线性回归模型对成对文章相似性进行建模,提出一种具有自适应停止准则的层次聚类方法。基于 Dempster-Shafer 理论(DST)的分层聚类方法[5]将每个文档嵌入到低维向量空间中进行聚类,通过定义 2 个文档各个特征之间的相似度来计算它们文档之间的相似度,将相似度大于阈值的文档划分到同一个簇中。另外一些学者利用概率模型表示文档之间的相似性[6-8]。

监督方法需要大量的标记数据,而人工标记需要昂贵的人力和财力。但是对于无监督算法,要找到最佳数目的聚类或者合适的相似性阈值具有一定的挑战性。因此也有许多学者提出半监督算法。

Levin 等[9]提出一种结合分类和聚类方法的 2 阶段算法。在第 1 阶段,他们应用基于论文引用及其他的高精度规则自动生成用于有监督训练的标记数据。在第 2 阶段,将正例和负例用作有监督的分类器,该分类器用于预测 2 篇文章是否由同一作者撰写,最后将分类器的预测结果用作聚类中的相似性度量。Louppe 等[10]在此基础上提出用于预处理的区域策略,将很有可能属于同一作者的文献放置于同一区域。

随着近 2 年图网络研究的兴起,由于作者及其刊物可以自然地构建作者-作者网络和文档-文档网络,因此一些基于图的方法也被用于姓名消歧任务。谱聚类[11]可以将图划分为几个部分从而进行聚类。Zhang 等[12]提出结合全局度量学习和局部链接图模型,通过文档的属性特征学习文档的低维表征。Zhang 和 Hasan[13]将文章信息预处理为 3 个图网络:作者-作者图,文档-文档图和作者-文档图,并将文档数据投影到低维空间中。GHOST 模型[14]利用作者图来计算图节点对之间的相似度。除此之外,还有基于文章对的图网络(ADANA)[15]和基于标题与共同作者的图网络(GFAD)[16]。

当前研究方法存在一些问题:1)监督方法因为使用了标注信息,所以消歧性能一般会好于无监督方法。但是由于数据集规模通常较大,人工标注所有的标签会耗费大量的人力和时间。2)现有的大多数研究方法通常只基于文献的属性特征或者基于文献关系、作者关系的研究。利用文献属性特征的方法通常采用大量的属性特征并制定相应的规则,在数据有缺失的时候会导致规则失效。基于关系图的研究往往忽略文档的基础属性特征,降低了消歧的效果。3)目前作者姓名消歧问题中大多数研究方法都是应用于小规模数据集,通常只包含 10~20 个作者文献集,本研究希望将研究方法应用于更大规模的数据集。

本文针对更大规模的数据集(100 个待消歧作者姓名),提出一种基于文献属性特征和关系图网络的姓名消歧方法(如图 1 所示)。该方法同时考虑文档的属性特征以及多个关系网络的信息,通过无监督学习的方法寻找文档表征向量,使用簇数标签进行层次聚类,取得良好的姓名消歧效果。在作者数据集 AMiner 上的测试结果表明:该方法优于使用大量文档标签和簇数的半监督方法[12],也优于其他基于图网络的方法[13-14]。另外,本文通过可视化的方式增加了模型的可解释性。

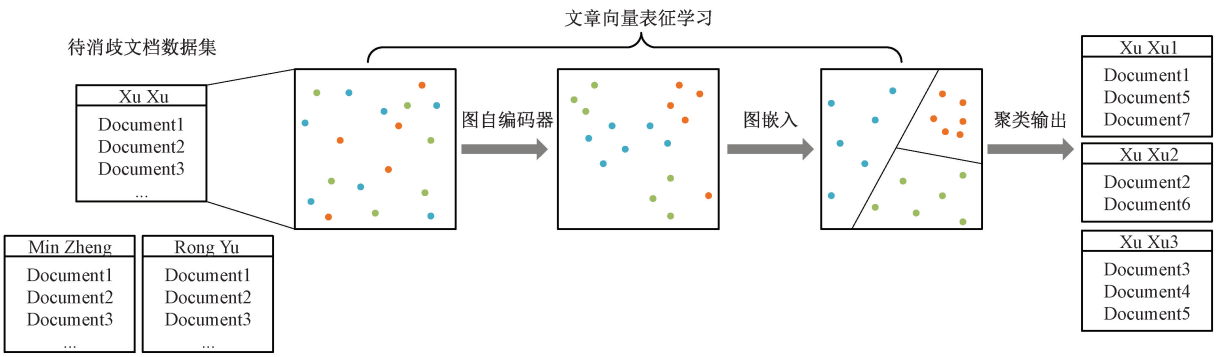


图 1 基于特征编码和图网络的姓名消歧方法

Fig. 1 Name disambiguation on encoding attributes and graph topology

1 问题形式化定义

假设 $\mathcal{D}$ 表示数据库中所有的文档集合,$|\mathcal{D}|$ 是所有文档的个数。对于需要进行消歧的作者姓名 i,$\mathcal{D}^i=\{\boldsymbol{d}_1^i,\boldsymbol{d}_2^i,\cdots,\boldsymbol{d}_N^i\}$ 代表与所有同名作者关联的所有 N 个文档的集合。每个文档都有一些属性特征,例如合作者、文档标题、发表刊物、作者隶属机构和文章关键词等。假设 $\boldsymbol{d}_n$ 代表其中的第 n 篇文档,那么 $\boldsymbol{d}_n$ 可以表示为 $\{f_1,f_2,\cdots,f_M\}$,其中 f 表示文档属性特征,M 表示属性特征的个数。将合作者集合表示为 $\mathcal{A}$,那么 $\mathcal{A}=\{a_1,a_2,\cdots,a_L\}$ 表示集合中有 L 个合作者。

在姓名消歧任务中,i 代表一个作者姓名。消歧任务就是找到合适的函数将与这个姓名相关的文档划分到不同的类别中,使得每个类别中仅包含同一作者的文档。给定一个文档集合 $\mathcal{D}^i$,任务是将文档划分为 K 个不相交的簇 $\mathcal{C}^i=\{C_1^i,\cdots,C_K^i\}$。其中,$C_k^i=\{d_j^i\mid\Phi(d_j^i)=p_k,d_j^i\in D^i\}$,$p_k$ 表示第 k 个作者。对于不同待消歧的作者姓名,这里 K 是不同的。用函数表示为

$$\Phi(\mathcal{D}^i)\rightarrow\mathcal{C}^i. \tag{1}$$

2 基于特征编码和图嵌入的姓名消歧

2.1 文档向量表征

Word2vec 模型被广泛用于单词表示学习中。本文利用 word2vec 的模型之一 CBOW[17] 用于学习文档的向量表示。

假设有一系列训练词 $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_T$,CBOW 模型通过某单词周围其他单词的出现频率预测这个单词的出现频率。该模型根据训练语料库中预定义上下文窗口内词的出现频率来学习单词向量。目标是最大化出现在预定义上下文窗口中的单词的共现概率,概率函数表示为

$$\frac{1}{T}\sum_{1}^{T}\sum_{-c\leqslant j\leqslant c,j\neq 0}\log p(\boldsymbol{w}_t\mid\boldsymbol{w}_{t+j}). \tag{2}$$

其中 c 代表预定义窗口大小。

本文将文档属性特征编码至文档向量中。将文献属性特征分割为单个词语并保留词干,用词干特征 $\hat{\boldsymbol{f}}$ 表示,其中属性特征包括文献标题、合作者、所属机构。对于一个特定的文档 $\boldsymbol{d}_n$,可以用

$\{\hat{\boldsymbol{f}}_1,\hat{\boldsymbol{f}}_2,\cdots,\hat{\boldsymbol{f}}_{M'}\}$ 表示。词干特征 $\hat{\boldsymbol{f}}$ 使用 one-hot 编码,M' 为词干特征的个数。将特征输入 CBOW 模型后训练,得到每个词干特征的低维向量表示 $\tilde{\boldsymbol{f}}$。文档的向量表征可以通过这些低维特征向量求和得出

$$\boldsymbol{d}_n = \sum_{i=1}^{M'} \alpha_i \tilde{\boldsymbol{f}}_i. \tag{3}$$

其中:α_i 是 $\hat{\boldsymbol{f}}_i$ 的逆文档频率(idf 值),用于减少语料库中常见单词(例如介词)的权重。待消歧文献集合表示为矩阵 $\boldsymbol{D}^i = [\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_N]$,该矩阵利用特征出现频率表示文档间的相似性。

2.2　变分图自动编码器

对于每一个待消歧的作者姓名 i,$\mathcal{D}$ 表示其对应的待消歧的文献集合。首先构建 i 对应的文档图网络 $\mathcal{G} = (\mathcal{D},\varepsilon)$,文档 $\boldsymbol{d} \in \mathcal{D}$ 可以表示网络节点,利用 2.1 节中的文档向量表征构建,ε 用于表示节点之间是否存在边,本文用邻接矩阵 $\boldsymbol{A}$ 表示。

在本文中,邻接矩阵由 2 种方法构建:一是计算两两文档之间的特征相似性,二是利用高精确度的规则。2 个文档相似度由 2 个文档的共同特征进行计算。假设 2 个文档的共同特征为 $\tilde{\boldsymbol{f}}_1,\cdots,\tilde{\boldsymbol{f}}_O$,相似度可以由 $\sum_{i=1}^{O}\alpha_i\tilde{\boldsymbol{f}}_i$ 计算得出,如果文档 $\boldsymbol{d}_i$ 和 $\boldsymbol{d}_j$ 的相似度大于阈值,则在这 2 个文档所表示的节点之间构造一条边,在邻接矩阵中表示为 $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji} = 1$。对于相似度小于阈值的 2 个文档,邻接矩阵中的相关值为 0。由于缺少标签数据,本文使用基于合作者和隶属机构的高精度规则来构建正例,表示 2 篇文档是由相同的作者撰写。通常情况下,同名作者在同一个机构工作并且他们的合作者也在同一个机构工作的可能性很小,这种情况下,在邻接矩阵中 2 个节点间构建一条边。一个作者的合作者在一段时间内的合作者也是相对固定的,所以当 2 个文档有很多合作者重合,在这 2 个文档之间也构建一条边。

如图 2 所示,本文使用变分图自动编码器用于增强模型的泛化能力。假设 $\boldsymbol{X} = [\boldsymbol{d}_1^T,\boldsymbol{d}_2^T,\cdots]$ 代表待消歧文档集合的矩阵。编码器是 2 层的图卷积神经网络(graph convolutional network, GCN):

$$q(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{A}) = \prod_{i=1}^{N} q(\boldsymbol{z}_i\mid\boldsymbol{X},\boldsymbol{A}). \tag{4}$$

其中

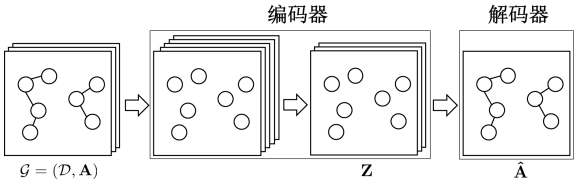


图 2　变分图自动编码器

Fig. 2　Variational graph auto-encoder

$$q(\boldsymbol{z}_i\mid\boldsymbol{X},\boldsymbol{A}) = \mathcal{N}(\boldsymbol{z}_i\mid\boldsymbol{\mu}_i,\mathrm{diag}(\boldsymbol{\sigma}_i^2)). \tag{5}$$

$\boldsymbol{\mu} = [\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots] = \mathrm{GCN}_\mu(\boldsymbol{X},\boldsymbol{A})$ 是图卷积网络输出各个向量的平均值构成的矩阵,$\boldsymbol{\sigma} = [\log\boldsymbol{\sigma}_1,\log\boldsymbol{\sigma}_2,\cdots] = \mathrm{GCN}_\sigma(\boldsymbol{X},\boldsymbol{A})$ 代表标准差矩阵。2 层卷积神经网络可以表示为

$$\mathrm{GCN}(\boldsymbol{X},\boldsymbol{A}) = \tilde{\boldsymbol{A}}\mathrm{Relu}(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}_0)\boldsymbol{W}_1. \tag{6}$$

$\mathrm{GCN}_\mu(\boldsymbol{X},\boldsymbol{A})$ 和 $\mathrm{GCN}_\sigma(\boldsymbol{X},\boldsymbol{A})$ 共享权重矩阵 $\boldsymbol{W}_0$,权重矩阵 $\boldsymbol{W}_1$ 不同。$\tilde{\boldsymbol{A}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{\frac{1}{2}}$,$\boldsymbol{D}$ 是度数矩阵(degree matrix)。

解码器通过线性层重构邻接矩阵 $\hat{\boldsymbol{A}}$:

$$p(\boldsymbol{A}\mid\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(\boldsymbol{A}_{ij}\mid\boldsymbol{z}_i,\boldsymbol{z}_j). \tag{7}$$

其中

$$p(\boldsymbol{A}_{ij}\mid\boldsymbol{z}_i,\boldsymbol{z}_j) = \sigma(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z}_j), \tag{8}$$

$\sigma(\cdot)$ 是 sigmoid 函数。自动编码器的优化目标是使得重构的邻接矩阵 $\hat{\boldsymbol{A}}$ 与原始邻接矩阵 $\boldsymbol{A}$ 尽可能地接近。损失函数表示为

$$\mathcal{L} = E_{q(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{A})}[\log p(\boldsymbol{A}\mid\boldsymbol{Z})] - \mathrm{KL}[q(\boldsymbol{Z}\mid\boldsymbol{X},\boldsymbol{A})\,\|\,p(\boldsymbol{Z})]. \tag{9}$$

等式右边第 1 项表示交叉熵函数,第 2 项中 $p(\boldsymbol{Z}) = \prod p(\boldsymbol{z}_i)$,$p(\boldsymbol{z}_i) = \mathcal{N}(0,1)$ 是标准正态分布,相对熵 $\mathrm{KL}[q(\cdot)\,\|\,p(\cdot)]$ 表示分布 p 和分布 q 之间的相似程度。

将编码器输出矩阵 $\boldsymbol{Z} = [\boldsymbol{z}_1^{\mathrm{T}},\boldsymbol{z}_2^{\mathrm{T}},\cdots]$ 中的每一项作为待消歧文档的新的向量表示。通过引入节点间的关系信息,输出矩阵具有更强的表征性能。

2.3　图网络嵌入

如果仅利用图自动编码器,在引入节点关系的时候只考虑到文档特征之间的相关性。当出现表 1 所示情况时,无法判断文档 1 和文档 2 是否由同一作者所撰写。引入文档 3 和前 2 个文档之间的关系后,因为 2 篇文章的所有合作者都是文档 3 的作者,所以可以判断出 2 篇文章属于同一作者。

本文研究希望通过利用合作者关系进一步推

表 1 合作者相关联的文档

Table 1 Related documents refer to one-hop co-author

文章	作者
1	张军燕,刘占军,李凡庆
2	张军燕,左广汉,季明
3	刘占军,李凡庆,左广汉,季明

断文档相似性,并将合作者关系网络信息嵌入文档表征向量 $\boldsymbol{Z}$ 中。提取网络结构信息的方法有 Deepwalk[18]、GCN[19]等。本文通过构建作者-作者网络、文档-作者网络和文档-文档网络,聚集有关系的作者和文档向量。

对于待消歧作者姓名 i, 构建作者-作者网络 $\mathcal{G}_{\text{pp}}=(\mathcal{A},\boldsymbol{\varepsilon}_{\text{pp}})$。$\mathcal{A}$ 是所有合作者构成的集合,$\boldsymbol{e}_{ij}\in\boldsymbol{\varepsilon}_{\text{pp}}$ 是作者 a_i 和作者 a_j 之间的边。$\boldsymbol{e}_{ij}$ 的权重定义为 2 个作者合作文章的数目。

文档之间的关系可以由文档-文档网络 $\mathcal{G}_{\text{dd}}=(\mathcal{D},\boldsymbol{\varepsilon}_{\text{dd}})$ 表示。$\boldsymbol{\varepsilon}_{\text{dd}}$ 包含表 1 所示的合作者信息,当 2 个文档有多个合作者有共同合作的第 3 篇文档时,在这 2 个文档所表示的节点之间建立一条边。为了在关系图中引入文档特征信息,将文档节点用 $\boldsymbol{Z}$ 表示, $\boldsymbol{Z}$ 是图自编码器中待消歧文档集合 $\mathcal{D}$ 的输出矩阵。

为了将合作者关系引入文档关系中,使待消歧文档可以更好地被聚类,本文构建图网络 $\mathcal{G}_{\text{pd}}=(\mathcal{A}\cup\mathcal{D},\boldsymbol{\varepsilon}_{\text{pd}})$ 用于连接文档和作者, $\mathcal{D}$ 为待消歧文档集合, $\mathcal{A}$ 是这些文档对应的合作者集合。如果作者 a_i 是文档 $\boldsymbol{d}_j$ 的作者之一, $\boldsymbol{e}_{ij}\in\boldsymbol{\varepsilon}_{\text{pd}}$ 为 1;反之, $\boldsymbol{e}_{ij}$ 为 0。如果 2 个文档的合作者集合有很大的重合,那么它们所对应的合作者在作者-作者图中也很接近,这 2 个文档相对也较近。

文档间相似性度量定义为 $S_{ij}^{\text{dd}}=\boldsymbol{z}_i^T\boldsymbol{z}_j$。当(文档 i, 文档 j)为正例,图网络嵌入模型希望最大化 S_{ij}^{dd} 从而使余弦相似度越接近于 1;(文档 i, 文档 j)为负例时,最小化 S_{ij}^{dd}。保留节点对顺序的概率为

$$p(S_{ij}\geqslant S_{it}\mid\boldsymbol{z}_i,\boldsymbol{z}_j,\boldsymbol{z}_t)=\sigma(S_{ij}^{\text{dd}}-S_{it}^{\text{dd}}). \quad (10)$$

对于待消歧文档集合 $\mathcal{D}$, 这个概率为

$$p(\mathcal{D})=\prod_{\substack{(\boldsymbol{d}_i,\boldsymbol{d}_j)\in\mathcal{P}_{\mathcal{G}_{\text{dd}}}\\(\boldsymbol{d}_i,\boldsymbol{d}_t)\in\mathcal{N}_{\mathcal{G}_{\text{dd}}}}}\sigma(S_{ij}^{\text{dd}}-S_{it}^{\text{dd}}). \quad (11)$$

对于文档-文档网络,希望这个概率更大,所以需要最大化这个概率

$$\mathcal{L}_{\text{dd}}=\min_{\mathcal{D}}-\ln p(\mathcal{D}). \quad (12)$$

类似地,对于作者-作者网络和作者-文档网络:

$$\mathcal{L}_{\text{pp}}=\min_{\mathcal{A}}-\ln p(\mathcal{A}), \quad (13)$$

$$\mathcal{L}_{\text{pd}}=\min_{\mathcal{A},\mathcal{D}}-\ln p(\mathcal{A},\mathcal{D}). \quad (14)$$

目标是将 3 个网络的拓扑结构信息嵌入文档表征向量中,优化函数为

$$\mathcal{L}=\min_{\mathcal{A},\mathcal{D}}\mathcal{L}_{\text{dd}}+\mathcal{L}_{\text{pp}}+\mathcal{L}_{\text{pd}}+\lambda(\|\boldsymbol{A}\|_{\text{F}}^2+\|\boldsymbol{Z}\|_F^2). \quad (15)$$

其中: $\boldsymbol{A}$ 为合作者构成的矩阵, $\boldsymbol{Z}$ 代表微调后的待消歧文档矩阵。通过训练图嵌入模型,文档表征向量中包含了文献属性特征及 3 个图网络的拓扑信息。

2.4 聚类

对图网络嵌入模型训练后的文档表征向量应用层次聚类算法[20]。该算法将训练样本中的每一个数据点都当作一个簇,然后计算每 2 个样本点之间的距离并合并距离最近的簇,直到满足终止条件。本文将终止条件设置为簇个数等于真实聚类个数。

3 仿真实验

3.1 仿真设置

本文使用在线学术搜索和数据挖掘系统 AMiner[21]上采样的 100 个作者姓名数据集,每个姓名都对应着与这个姓名相关联的文档,采样数据集共包含 27 128 篇文献和 1 066 个真实作者。

超参数设置上,CBOW 模型中,文档表征向量维度设置为 100,预定义上下文窗口为 5。变分图自编码器中,逆文档频率的阈值为 25,第 1 层图卷积网络输出维数为 200,第 2 层图卷积网络输出维度设置为 100,学习率为 0.01,迭代 200 次。图网络嵌入模型中,学习率为 0.05,正则化参数为 0.01。

3.2 性能比较

在仿真实验中,对比本文方法与其他几个基于图网络的姓名消歧方法。Zhang 等[12]提出一种合并全局表示学习和局部嵌入学习的方法(Aminer)。在全局表示学习中,需要引入标签信息构建正负样本。在局部嵌入学习方法中使用图自动编码器。Zhang 和 Hasan[13]将作者-作者、作者-文档、文档-文档网络信息压缩至低维空间。GHOST 模型[14]只考虑作者合著关系,在每个合作者间建立网络,通过选择有效路径计算作者节点之间的相关性划分作者聚类。并查集方法通过合作者和隶属机构的严格匹配在文档间建立图连接,将所有有连接关系的文档节点构成一个集群。

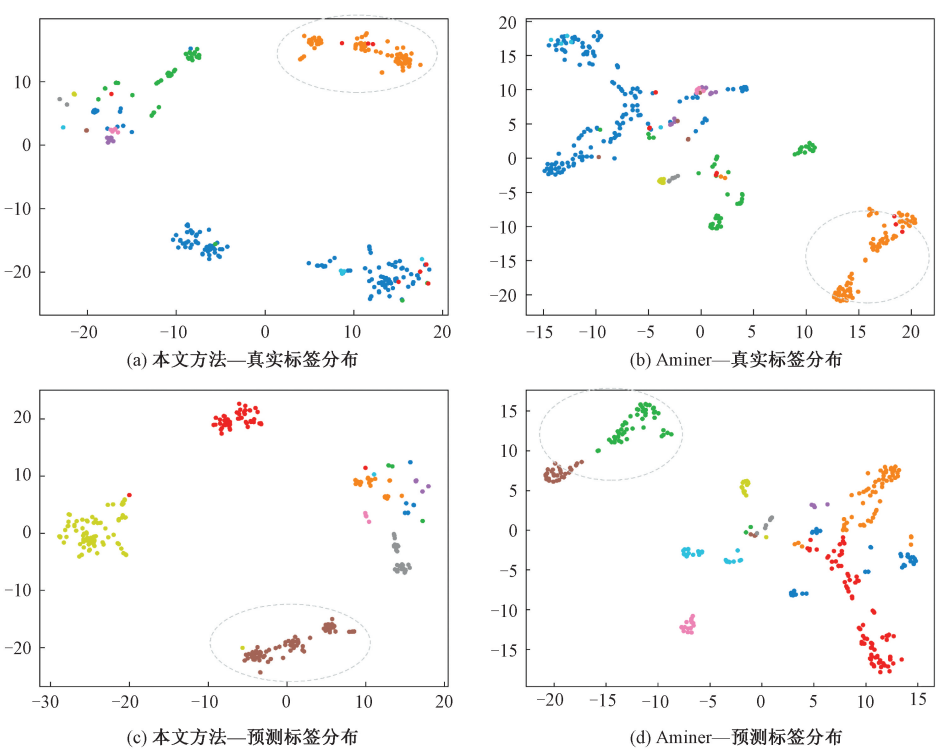
本文使用 pairwise Precision、Recall 和 F_1 值[22]对模型进行性能比较。对 100 个消歧作者数据集计算每个指标的平均值。表 2 显示不同的消歧方法在 AMiner 数据集上的仿真结果。可以看到,本文提出的方法在表中 15 个姓名中有 11 个都表现最佳,平均 F_1 值比 Aminer 算法[12]提高 3.87%,比 Zhang 和 Hasan[13]的算法高 25%,比 GHOST 模型[14]高 33.85%。

图 3 是一个待消歧文档数据集通过本文方法与 Aminer 学习后的文档表征向量的 2 维空间可视化,图 3(a)、3(b)中不同的颜色表示不同的真实集群。图 3(c)、3(d)为预测集群分布。在此

表 2　几种基于图网络的姓名消歧方法的聚类结果

Table 2　Clustering results of different graph-based name disambiguation methods

姓名	Our			Aminer 方法[12]			Zhang 和 Hasan[13]			GHOST 模型[14]			并查集		
	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1	Prec	Rec	F_1
Xu Xu	0.7073	0.7006	**0.7039**	0.7418	0.4586	0.5668	0.4773	0.3998	0.4351	0.6134	0.2179	0.3215	0.0722	0.6629	0.1302
Rong Yu	0.7307	0.4281	0.5399	0.8913	0.4651	**0.6112**	0.6653	0.3690	0.4747	0.9200	0.3641	0.5217	0.1600	0.4122	0.2305
Yong Tian	0.7812	0.6103	**0.6853**	0.7632	0.5195	0.6182	0.7318	0.5634	0.6366	0.8694	0.5458	0.6706	0.1078	0.9450	0.1936
Lu Han	0.6002	0.3529	**0.4445**	0.5178	0.2805	0.3639	0.4605	0.1795	0.2583	0.6972	0.1739	0.2784	0.1510	0.9630	0.2610
Lin Huang	0.7472	0.5043	**0.6021**	0.7710	0.3287	0.4609	0.6943	0.3313	0.4486	0.8615	0.1725	0.2874	0.0591	0.3376	0.1006
Kexin Xu	0.9109	0.8965	0.9037	0.9137	0.9864	**0.9487**	0.8574	0.4413	0.5827	0.9290	0.2852	0.4364	0.7763	0.8362	0.8051
Wei Quan	0.7723	0.4886	**0.5985**	0.5388	0.3902	0.4526	0.7441	0.3394	0.4662	0.8642	0.2780	0.4207	0.3716	0.9657	0.5367
Tao Deng	0.7638	0.5382	**0.6315**	0.8163	0.4362	0.5686	0.5525	0.2793	0.3711	0.7333	0.2450	0.3673	0.1255	0.6476	0.2103
Hongbin Li	0.7009	0.7848	**0.7405**	0.7720	0.6921	0.7299	0.6579	0.5286	0.5862	0.5629	0.2912	0.3839	0.1292	0.9459	0.2273
Hua Bai	0.6586	0.4194	**0.5125**	0.7149	0.3973	0.5108	0.5493	0.3597	0.4347	0.8306	0.2954	0.4358	0.2208	0.9323	0.3571
Meiling Chen	0.7960	0.4784	**0.5976**	0.7493	0.4470	0.5599	0.7922	0.2515	0.3818	0.8611	0.2385	0.3735	0.2483	0.6692	0.3622
Yanqing Wang	0.4900	0.6649	0.5642	0.7152	0.7533	**0.7337**	0.7273	0.4262	0.5374	0.8079	0.4039	0.5386	0.2412	0.6695	0.3546
Xudong Zhang	0.7455	0.2224	**0.3426**	0.6240	0.2254	0.3312	0.5563	0.0811	0.1416	0.8575	0.0723	0.1334	0.6512	0.4736	0.5484
Qiang Shi	0.5338	0.5036	**0.5183**	0.5220	0.3615	0.4272	0.4333	0.3799	0.4049	0.5372	0.2680	0.3576	0.1811	0.8637	0.2994
Min Zheng	0.5985	0.2076	0.3082	0.5765	0.2235	**0.3221**	0.5362	0.1763	0.2654	0.8050	0.1521	0.2558	0.1195	0.7448	0.2060
Average	0.7810	0.6747	**0.7240**	0.7796	0.6303	0.6970	0.7022	0.4872	0.5753	0.8172	0.4043	0.5409	0.4078	0.7652	0.5320



(a) 本文方法—真实标签分布　(b) Aminer—真实标签分布

(c) 本文方法—预测标签分布　(d) Aminer—预测标签分布

图 3　不同姓名消歧方法聚类结果的可视化

Fig. 3　Visualization of clustering results of different name disambiguation methods

数据集上，本文提出的方法的 F_1 值为 0.633 8，Aminer 方法仅为 0.538 2。从图 3(d) 虚线内的样本可以看出，Aminer 方法学习后的文档表征向量在向量空间中距离较远，样本并没有被正确归类。而本文提出的方法通过将相似的表征向量聚集在一起，如图 3(a) 的橙色散点表示，输出的文档表征向量更加接近，图 3(c) 中并没有将这些散点划分错误，从而实现了更好的聚类效果。

从表 2 中可以看到对于其中的 4 个姓名 Aminer 方法更好，为进一步分析其中的原因，本文选取作者姓名为“Rong Yu”的文档集合并对本文方法与 Aminer 模型的聚类结果进行比较。

图 4 为在这个文档数据集上 2 种方法聚类结果的可视化对比。从图 4(b) 中的蓝色散点可以看出，通过 Aminer 方法学习文档向量表征后，属于这个作者的文档向量主要集中在 2 个区域，而本文的方法将更多的点集中到左侧椭圆虚线框内，如图 4(a) 所示，这意味着本文方法将更多的文档划分到了正确的类中。但是因为本文方法将很多文档向量从右侧虚线框内移出，导致其余的文档向量在向量空间中太过分散，从图 4(c) 中可以看出，这些文档向量被划分为 3 个不同的类。在图 4(d) 中，这些文档向量虽然也被划分到另一个类别中，但是根据聚类方法中 pairwise F_1 值的计算方法可知，这些文档向量组成的两两文档对在预测集和真实集中仍然都属于同一个类别，仍算作 True-Positive 文档对。因此在作者姓名为“Rong Yu”的文档数据集上，Aminer 的 F_1 值高于本文提出的方法。

图 5 为使用 word2vec 构建文档向量表征后直接对该文档集合中的文档向量进行聚类的结果可视化。从蓝色散点可以看出，进行文档向量表征后属于同一作者的文档向量就被划分到了向量空间中不相连的 2 个区域中，从文档属性特征分析，代表这个作者的文章有 2 个强属性特征，他的大部分文章都与其中一个属性相关，例如他可能有 2 个不同的研究方向，这 2 部分文章的特征词并不相关，所以在特征编码后与他相关的文档向量分布在 2 个区域。而本文方法在引入关系信息后使得模型能够区分出其中一部分文档。但是由于并不能覆盖到所有的文档，在属性特征关系弱的数据集中，文档向量分布较为分散，本文的方法会导致一部分文章被划分到多个不同的类别中，而 Aminer 方法虽然也没有将这些文档划分到正确的类别中，但是保留了它们彼此之间的联系，使得这些文档被划分为同一个类别，所以本文方法的聚类结果的 F_1 值相对较低。

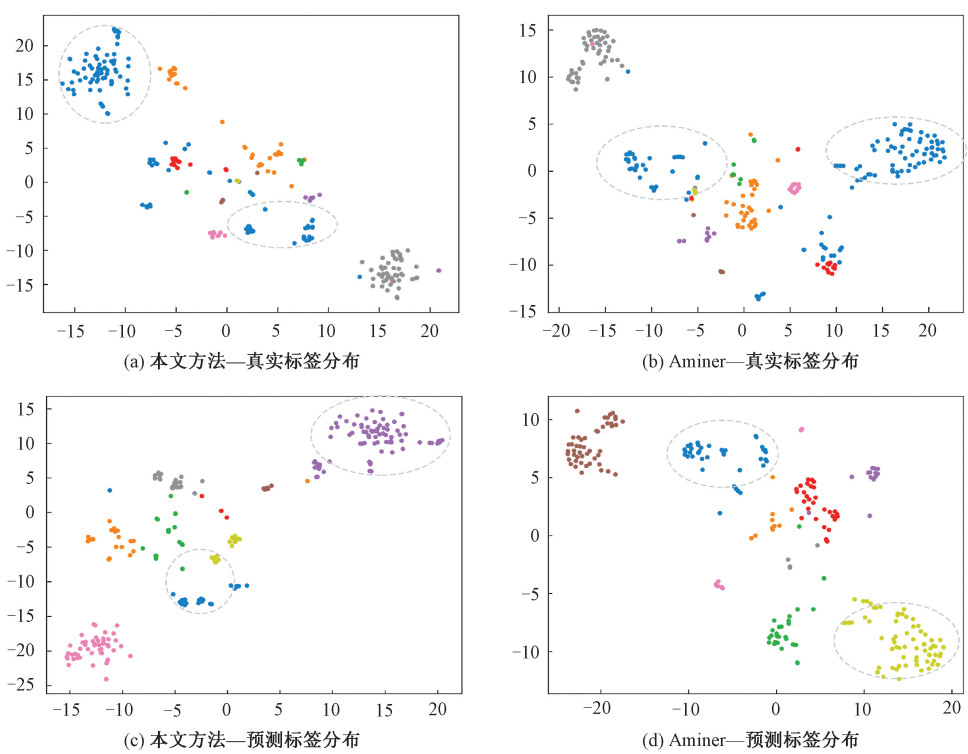


图 4 Rong Yu 文档集合上的聚类结果对比

Fig. 4 Comparison of clustering results on the document set of Rong Yu

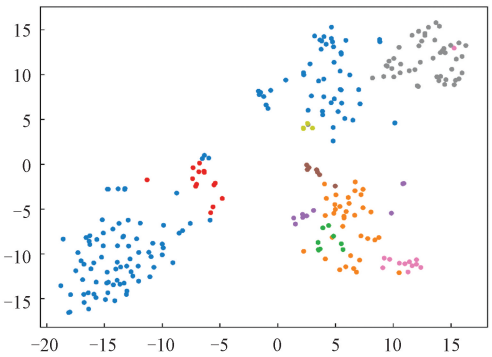


图 5　文档向量表征后的聚类结果

Fig. 5　Clustering results after document representation

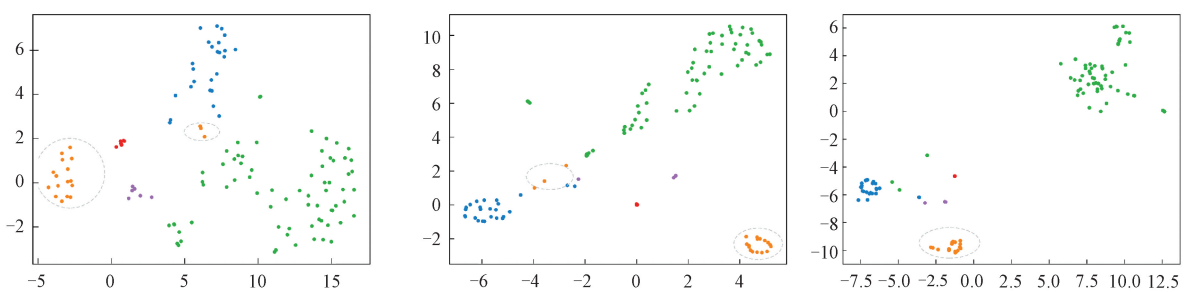
3.3　组件性能分析

为了展示本文方法中文档向量表征、变分图自编码器、图嵌入模型各自的作用与聚类效果,本节将每个组件分开评估。图自动编码器和图网络嵌入模型建立于构建了文档向量表征之后。如表 3 所示,图自编码器和图网络嵌入分别将模型的 F_1 值提高了 0.064 1 和 0.048 3。而本文提出的综合方法取得了最高的准确率和召回率。图 6 为每个子模型训练后学习的文档向量的低维可视化,这里使用真实标签在文档表征空间的分布,不同颜色代表不同作者所撰写的文档向量。由图 6 可以看出,图自编码器将绿色点和蓝色点聚集在了一起,而图网络嵌入使这些点更加接近使得模型可以更准确地聚类。同时,图网络嵌入模型将离群的黄色点移动到了正确的区域,所以本文的模型对异常值有一定效果。

表 3　组件性能分析

Table 3　Clustering results of each component

	Prec	Rec	F_1
文档向量表征	0.722 9	0.501 4	0.592 1
图自动编码器	0.755 3	0.580 1	0.656 2
图网络嵌入	0.777 1	0.544 6	0.640 4
综合	0.781 0	0.674 7	0.724 0



(a) 文档向量表征　(b) 图自编码器　(c) 图网络嵌入

图 6　各组件聚类结果可视化

Fig. 6　Visualization of clustering results of each component

4　结论

本文基于图网络提出一种新的作者姓名消歧方法,该方法通过文档表征、图自动编码器和图嵌入模型来编码所有论文的属性特征和作者及论文的关系图拓扑结构。采样于数据挖掘系统 AMiner 的数据集被用于验证本文提出的图网络姓名消歧方法,仿真结果证明本文提出的模型优于目前其他几种基于图网络的姓名消歧方法。

参考文献

[1] Wang J, Berzins K, Hicks D, et al. A boosted trees method for name disambiguation[J]. Scientometrics, 2012, 93(2): 391-411. DOI:10.1007/s11192-012-0681-1.

[2] Tran H N, Huynh T, Do T. Author name disambiguation by using deep neural network[C]//Intelligent Information and Database Systems. ACIIDS 2014. Lecture Notes in Computer Science, Spvinger, Cham. 2014, 8397: 123-132. DOI:10.1007/978-3-319-05476-6_13.

[3] Han H, Giles L, Zha H, et al. Two supervised learning approaches for name disambiguation in author citations[C]// Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004. Tucson, AZ, USA. IEEE, 2004: 296-305. DOI:10.1145/996350.996419.

[4] Cen L, Dragut E C, Si L, et al. Author disambiguation by hierarchical agglomerative clustering with adaptive stopping criterion[C]//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin Ireland. New York, NY, USA: ACM, 2013: 741-744. DOI:10.1145/2484028.2484157.

[5] Wu H, Li B, Pei Y J, et al. Unsupervised author disambiguation using Dempster-Shafer theory [J]. Scientometrics, 2014, 101(3): 1955-1972. DOI:10.1007/s11192-014-1283-x.

[6] Song Y, Huang J, Councill I G, et al. Efficient topic-based unsupervised name disambiguation[C]//Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. June 18-23, 2007. Vancouver, BC, Canada. New York: ACM Press, 2007: 342-351. DOI: 10.1145/1255175.1255243.

[7] Tang J, Fong A C M, Wang B, 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.

[8] Tang J, Qu M, Mei Q Z. PTE: predictive text embedding through large-scale heterogeneous text networks[C]// Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 1165-1174. DOI:10.1145/2783258.2783307.

[9] Levin M, Krawczyk S, Bethard S, et al. Citation-based bootstrapping for large-scale author disambiguation[J]. Journal of the American Society for Information Science and Technology, 2012, 63(5): 1030-1047. DOI:10.1002/asi.22621.

[10] Louppe G, Al-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. Springer, Cham, 2016: 272-287. DOI: 10.1007/978-3-319-45880-9_21.

[11] Han H Y, Zha H, 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 (JCDL'05). June 7-11, 2005, Denver, CO, USA. IEEE, 2005: 334-343. DOI: 10.1145/1065385.1065462.

[12] Zhang Y T, Zhang F J, Yao P R, et al. Name disambiguation in AMiner: clustering, maintenance, and human in the loop[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1002-1011. DOI:10.1145/3219819.3219859.

[13] Zhang B C, Hasan M A. Name disambiguation in anonymized graphs using network embedding[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore, Singapore. New York, NY, USA: ACM, 2017: 1239-1248. DOI:10.1145/3132847.3132873.

[14] Fan X M, Wang J Y, Pu X, et al. On graph-based name disambiguation[J]. Journal of Data and Information Quality, 2011, 2(2): 1-23. DOI:10.1145/1891879.1891883.

[15] Wang X Z, Tang J, Cheng H, et al. ADANA: active name disambiguation[C]// 2011 IEEE 11th International Conference on Data Mining. December 11-14, 2011, Vancouver, BC, Canada. IEEE, 2011: 794-803. DOI:10.1109/ICDM.2011.19.

[16] Shin D, Kim T, Choi J, et al. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information[J]. Scientometrics, 2014, 100(1): 15-50. DOI:10.1007/s11192-014-1289-4.

[17] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems (NIPS). 2013: 3111-3119.

[18] Perozzi B, Al-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, NY, USA: ACM, 2014: 701-710. DOI:10.1145/2623330.2623732.

[19] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[EB/OL]. arXiv:1609.02907. (2016-11-03)[2020-02-10]. https://arxiv.org/abs/1609.02907.

[20] Ward Jr J H. Hierarchical grouping to optimize an objective function[J]. Journal of the American statistical association, 1963, 58(301): 236-244. DOI:10.1080/01621459.1963.10500845.

[21] Tang J, Zhang J, Yao L M, et al. ArnetMiner: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 24-27, 2008. Las Vegas, Nevada, USA. New York, NY, USA: ACM Press, 2008: 990-998. DOI:10.1145/1401890.1402008.

[22] Menestrina D, Whang S E, Garcia-Molina H. Evaluating entity resolution results[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 208-219. DOI: 10.14778/1920841.1920871.