文章编号:1008-1542(2020)03-0233-09

基于元路径异构网络嵌入的姓名实体消歧方法

王建霞,张玉璇,许云峰

(河北科技大学信息科学与工程学院,河北石家庄 050018)

摘 要:为了解决大型学术数据库中重名作者的歧义消解问题,提出了基于元路径异构网络嵌入的姓名实体消歧模型。使用大型在线学术搜索系统DBLP上的公开数据集,首先抽取学术出版物的作者信息、标题和会议期刊名称等特征属性,再利用word2vec模型工具生成的特征属性词嵌入输入到GRU网络中进行训练,构造出一个PHNet矩阵网络进行随机游走操作,从而捕捉不同类型节点之间的关系,最后进行相似节点的划分,完成姓名消歧工作。实验结果显示,新方法的精确度为0.865,召回率为0.792,F_1值为0.815。基于元路径的异构网络嵌入模型的精确度、召回率等指标都优于对比模型。因此,所提出的模型在提高大型学术数据库的消歧精准度方面具有良好的应用前景。

关键词:自然语言处理;计算机神经网络;实体消歧;网络嵌入;异构网络

中图分类号:TP311.13 **文献标识码**:A **doi**:10.7535/hbkd.2020yx03005

Disambiguation method of name entities embedded in meta-path heterogeneous networks

WANG Jianxia, ZHANG Yuxuan, XU Yunfeng

(School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China)

Abstract: In order to solve the problem of disambiguation of duplicate authors in large academic databases, a name entity disambiguation model based on meta-path heterogeneous network was proposed. Based on the public data of the large online academic search system DBLP, the author information, title, name of conference journal and other characteristic attributes of academic publications were extracted first. Then the characteristic attribute words generated by the word2vec model tool were embedded into the GRU network for training, so that a PHNet matrix network for random walk operation was constructed to capture the relationship between different types of nodes and finally similar nodes were divided to complete the name disambiguation. The experimental results show that the accuracy of the method is 0.865, the recall rate is 0.792, and the F_1 value is

收稿日期:2020-03-25;修回日期:2020-05-25;责任编辑:冯 民

基金项目:中国留学基金委地方合作项目(201808130283);中国教育部人工智能协同育人项目(201801003011);河北科技大学校立课题(82/1182108);河北科技大学雾霾与空气污染防治科研项目(82/1182169);河北省科技支撑计划项目(17210104D,18210109D);河北省高等学校科学技术研究项目(ZD2015099);河北省高层次人才资助项目(A2016002015)

第一作者简介:王建霞(1970—),女,河北临城人,教授,硕士,主要从事网络与数据库、图像处理方面的研究。

通讯作者:许云峰副教授。E-mail:hbkd_xyf@hebust.edu.cn

王建霞,张玉璇,许云峰.基于元路径异构网络嵌入的姓名实体消歧方法[J].河北科技大学学报,2020,41(3):233-241.

WANG Jianxia, ZHANG Yuxuan, XU Yunfeng.Disambiguation method of name entities embedded in meta-path heterogeneous networks [J].Journal of Hebei University of Science and Technology,2020,41(3):233-241.

0.815.The meta-path-based heterogeneous network embedding model is superior to the comparison model in terms of accuracy and recall rate. Therefore, the proposed model has a good application prospect in improving the accuracy of disambiguation of large academic databases.

Keywords: natural language processing; computer neural network; entity disambiguation; network embedding heterogeneous network

现今,人们检索学术论文主要依赖学术搜索引擎,如 Google Scholar、百度学术、DBLP(DataBase systems and logic programming)等。这些大型学术数据库共同面临的一个具有挑战性的问题是作者姓名的歧义消解,即通过作者的姓名来准确识别现实世界中的人。这一问题的解决对于 DBLP 这样的大型数据库图书馆尤为重要。DBLP 是 Schloss Dagstuhl-Leibniz 信息学中心和特里尔大学的联合服务机构。Schloss Dagstuhl 是一家"gemeinnutzige GmbH",是被德国法律所允许的一个非盈利慈善组织,是为了增进世界计算机科学界的学术信息交融而成立的。Schloss Dagstuhl 主要进行数字方法和论文书目元数据处理等研究。DBLP 在处理计算机科学数据的同时,还提供计算机学术论文所涉及到的论文作者的相关属性。除了公共领域所提供的论文数据外,DBLP 不会向任意第三方公开论文的私密数据,并且 DBLP 用户的行为也不会被系统跟踪,与此同时,DBLP 不会使用用户的任何数据进行广告宣传。总之,DBLP 就是一个仅仅提供计算机学术界科学会议和期刊论文出版记录的大型学术数据库。

最近几年,DBLP 数据库成为一些学术研究的目标,研究内容包括分析社区结构[1]、预测未来学术界的发展[2],以及研究合著作者间的协作关系[3]、测试姓名消歧算法[4]或大数据管理[5-6]。为保证上述研究的现实意义,需要高质量的数据,即需要数据集与现实之间的差异达到最小化[7]。不难想象,如果数据集与现实数据有差距,那么当用户通过论文作者姓名搜索作者的相关论文记录时,就不能确保得到与目标作者相关的"唯一且所有"的检索结果。论文作者的姓名歧义问题对于临时用户不会产生很大的问题,因为他们可以通过手动、有目的地对检索结果进行有效的过滤输出,得到需要的内容。然而对于使用 DBLP 数据集模拟学术活动的研究人员来说,这就是一个亟待解决的问题。在大型学术检索数据库中,由于作者可能会以不同形式表示他们的名字,或由于同音异词情况的存在以及不同作者共享同一个名字的问题,导致通过论文作者姓名识别作者论文时出现混淆,在实验中使用这样的数据集将会降低实验研究的真实性和可靠性。论文作者的重名消歧,即将来自不同学术期刊的多个同名作者论文加以区分,这在知识网络研究和文献信息检索等任务中具有重要的作用[8]。在 DBLP 中通过作者姓名来搜索论文是最常用的操作,所以对于大型学术数据库来说,能够识别来自同一个作者人名的论文信息且生成准确完整的结果变得尤为重要。

1 相关工作

自从 Mark Newman 使用计算机技术研究大规模学术网络中的作者网络以来,许多学者都针对学术协作网络的模式和未来发展动态做出了贡献。GARFIELD[9]对于论文作者姓名消歧这一问题指出,以全名加首字母或全名中只有中间名提取首字母的 2 种情况选一的不定形式来表示作者姓名的方法是作者姓名歧义和混淆的根源。比如,当论文的出版记录发生歧义时可能会产生 2 个完全不同的作者碰巧拥有相同的姓氏和首字母的情况,就像是"Catherine Blake"和"Coope Blaker",有可能会被错误地合并为一个身份,也就是"C Blake"。这种错误的划分会对网络结构、实验结论以及构建的关于合著作者网络的模式和演化理论产生影响。

对于作者姓名的歧义问题,DBLP 管理团队也经常通过使用算法的方式或人力手工摘选的方式消除部分作者姓名的歧义。DBLP 研究者们从学术数字图书馆收集有歧义姓名和出版的论文记录,通过检查个人简历和个人网页上的信息来进行重名作者的区分,或者通过电子邮件直接询问作者,确定哪个姓名属于谁[10]。DBLP 团队每月都会公开发布数据集。这些开放数据不仅使其研究团队能够使用它们来测试 DBLP 的消歧性能,而且可用于帮助其他人再现研究的结果,并改进研究的评价方法[11]。

关于姓名消歧的方法,可以选择基于作者的特征属性方法,即利用论文的合著作者信息、论文主题相似度、隶属关系和论文标题等特征来处理作者姓名的模糊。例如文献[12]方法的突出贡献在于,它能够检测出即使有 2 个或多个作者改变了固定的隶属关系、合著作者和研究兴趣,也可以划分为同一个人。现有的姓名

消歧工作还考虑了监督、非监督和概率关系模型。基于监督的姓名消歧方法，HAN 等[13]提出利用朴素贝叶斯和支持向量机，在这些工作中，将一个独特的实体组视为一个类，其目的是将每一个实体作者记录归并为一个类。非监督的名称消歧，是将记录划分为几个集群，目标是获得一个分区，其中每个集群包含来自唯一实体的记录。例如，GILES 等[14]应用 K-way 谱聚类方法对书目数据中的名称进行消歧任务。另外还有利用隐马尔可夫随机场建立论文间关系[15]的模型。

现有的大多数论文作者姓名消歧任务解决方案要么使用特征属性，要么使用外部来源所收集的辅助特性。然而，试图提取个人的外部数据会带来侵犯隐私的风险。为了避免侵犯隐私问题，一些研究考虑使用匿名图来消除名称歧义，不使用节点属性。这些解决思路的中心思想是利用图的拓扑特征，通过论文属性的集合，在不侵犯用户隐私的情况下，解决姓名的歧义消解问题[16]。

最近提出的网络嵌入方法如 DeepWalk[17]、LINE[18]、PTE[19]、Node2Vec[20]都具有很好的可扩展性，在节点分类和链路预测任务中表现出更好的性能。与以上方法所不同的是，为了提高姓名实体消歧任务的准确性，笔者使用网络嵌入的方法来学习论文的表示，从而完成姓名消歧任务。为每个需要进行消歧的人名数据集构建一个异构网络，结合单词嵌入的方法[21-22]，通过字符嵌入来挖掘语法特征，进而对异构网络中的结构以及节点特征编码进行表示学习。研究结果表明，结合深度学习模型的方法性能比基于特征属性的方法（如卷积神经网络（convolutional neural network，CNN）[23]、递归神经网络（neural network，RNN）[24]、LSTM 网络等）表现得更好。

2　研究方法

自然语言中存在大量的同音异词现象，这些词被称为歧义词。歧义词的消解工作是自然语言处理领域的基础研究，同时亦是核心研究，在机器翻译、语音识别、文本分类、信息检索等方面都具有十分重要的作用。针对 DBLP 这样一个存在作者姓名混淆的数据库，虽然其官网对作者姓名检索的重名问题有所解决（将重名作者的姓名后加以 4 位数的后缀区分），但是由于数据本身的复杂性以及不完整性，其区分并不准确，具有冗余以及缺失的情况，对于利用该数据集进行科学研究的人员来说，存在较大的弊端。

由于异构信息网络中的实体名称本质上是模糊的，使得学术作者网络中的不同作者可以有相同或相似的名称。在 DBLP 异构网络中，有 2 370 个带有消歧页面的高度歧义的作者名。例如，Yang Liu 这个名字指的是 33 名不同的研究人员，将每个人都连接到自己的论文，其中有 735 篇杨柳的论文没有被分配给这 33 位研究人员中的任何一位。当再有新的论文发表时，仅仅只是添加了 DBLP 中的论文出版记录，却没有增加消除歧义的过程。这种现象损害了实体的连接质量，从而影响了以现有数据库作为知识来源去完成其他任务的可能性。

本文针对 DBLP 数据库的重名作者消歧问题进行以下研究。

2.1　论文信息预处理

本文使用的 DBLP 数据信息包括论文的标题、作者、出版物名称、年份和 id 编号等信息。由于数据信息中存在噪音数据，所以首先需要进行预处理。预处理过程依次对论文信息进行去噪处理，包括去掉特殊字符串，去掉标点符号及特殊符号，去掉多余空格和换行符，去掉停用词等，然后提取需要的信息归纳到一起。

以歧义人名 Bo Liu（见图 1）为例，该人名下的出版物论文为 124 篇，根据论文标题的内容可知，Bo Liu 名下有研究神经网络的论文，也有研究基于图挖掘算法等研究方向的论文，再依据 organization 可粗略看出，有从属于清华大学、北京科技大学和暨南大学等的 Bo Liu，甚至很多 Bo Liu 并未显示其所属研究机构。这样有歧义的人名，本试验一共使用了 109 个，其中出版物数量最多的是 Wen Gao 数据集，其包含 484 条出版记录。

在预处理工作中，将 109 个 XML 格式的生数据集处理为 5 个 TXT 文件，分别为 paper_author.txt，paper_author1.txt，paper_conf.txt，paper_title.txt 和 paper_word.txt。图 2 为 paper_title.txt 部分文本内容，其中包含内容为出版物论文 id 以及论文标题，其中论文标题经过处理，将其统一使用小写字母表示，并且去掉了标题中的多种符号。对于论文标题的处理有助于后续生成 paper_word.txt 文档，该文档保留的内容如图 3 所示，即是论文 id 以及去掉预设的诸多停止词（例如，at，based，in 等）。每一词都另起一行，与论文 id 成行。另外 3 个文档内容不再赘述，都是与出版物论文 id 的结合。

```
<?xml version="1.0" encoding="UTF-8"?>
- <person>
    <personID>13520</personID>
    <FullName>Bo Liu</FullName>
    <FirstName>Bo</FirstName>
    <LastName>Liu</LastName>
  - <publication>
      <title>Rate-Distortion Optimized Progressive Geometry Compression</title>
      <year>2006</year>
      <authors>Wenbo Zhang,Bo Liu,Hongbin Zhang</authors>
      <jconf>CGIV</jconf>
      <id>8141</id>
      <label>0</label>
      <organization>null</organization>
    </publication>
  - <publication>
      <title>Multi-resolution Meshes Deformation Based on Pyramid Coordinates</title>
      <year>2007</year>
      <authors>Chenming Sha,Bo Liu,Zhanguo Ma,Hongbin Zhang</authors>
      <jconf>CGIV</jconf>
      <id>8149</id>
      <label>0</label>
      <organization>Beijing University of Technology, China</organization>
    </publication>
```

图 1 Bo Liu 部分内容

Fig.1 Bo Liu partial content

```
i276300  mobility pattern based anomaly detection algorithm in mobile networks
i277587  detection of masquerade attacks on wireless sensor networks
i299548  dynamic access control framework based on events
i300057  towards bringing database management task in the realm of it non experts
i302056  enhanced business intelligence using erocs
i388794  hardware diagnosis as program debugging
i515636  efficient design of boltzmann machines
i545651  policy driven data administration
i596099  clustering short texts using wikipedia
i600181  liptus  associating structured and unstructured information in a banking environment
i675629  retaining personal expression for social search
```

图 2 paper_title.txt 部分内容

Fig.2 paper_title.txt partial content

2.2 训练基于 GRU 的编码器学习深层语义表示

该部分进行的是基于 GRU 的深度表示学习，应用 gensim 库中的 word2vec 模型生成出版物标题的词嵌入，训练单词向量时维数＝100。嵌入向量的维数定义 batch 大小为 128，嵌入大小为 64，学习率为 0.001。

GRU 即 Gated Recurrent Unit，是 LSTM 网络的一种的变体。试验发现使用 GRU 可以使训练成果得到提升。

更新门和重置门是 GRU 模型中仅有的 2 个门，具体结构如图 4 所示。

图 4 中的更新门用 z_t 表示，重置门用 r_t 表示。其中用于控制之前时刻的状态信息被带入到当前状态中的程度是更新门的任务，这个值越大，代表前一时刻带入的状态信息越多。重置门的作用是调控之前状态有多少信息被写入到当前的候选集 $\tilde{h}_t$，重置门的值越小，代表之前状态写入的信息越少。

根据图 4 的 GRU 模型图，网络的前向传播公式如式(1)—式(3)所示。

```
i39153   time
i39153   representation
i39153   prolog
i39153   circuit
i39153   modelling
i149270  optimal
i149270  deployment
i149270  detecting
i149270  events
i175996  web
i175996  page
i175996  ranking
i175996  events
i258611  integrated
```

图 3 paper_word.txt 部分内容

Fig.3 paper_word.txt partial content

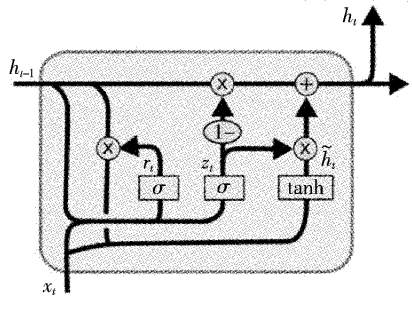


图 4 GRU 模型

Fig.4 GRU model

$$r_t=\sigma(W_r\cdot[h_{t-1},x_t]),\tag{1}$$

$$z_t=\sigma(W_z\cdot[h_{t-1},x_t]),\tag{2}$$

$$\tilde{h}_t=\tanh(W_{\tilde{h}}\cdot[r_t*h_{t-1},x_t])\ 。\tag{3}$$

先利用重置门控 r_t 来获得“重置”之后的数据 $h_{t-1}\cdot r_t$，再与输入 x_t 进行拼接，之后再经过一个 tanh 激活函数来处理数据，将其放缩到 $-1\sim1$ 的范围内。此时的 $\tilde{h}$ 包含了输入数据 x_t。式(3)对 $\tilde{h}_t$ 的操作与 LSTM 的选择记忆阶段类似，可以理解为记忆了当前时刻的状态。

在更新记忆阶段，使用了式(2)得到的更新门控 z_t 进行遗忘和记忆 2 个操作。更新表达式见式(4)。

$$h_t=(1-z_t)*h_{t-1}+z_t*\tilde{h}_t。\tag{4}$$

式中：z_t(门控信号)的区域是 0～1，若记忆下的数据越多，则门控信号越逼近 1，遗忘的数据越多则越逼近 0；$(1-z_t)*h_{t-1}$是对原本隐藏状态进行的选择性遗忘；$(1-z_t)$作为遗忘门，用来遗忘 h_{t-1}中一些不紧要的内容；$z_t*\tilde{h}_t$ 是对包含当前节点信息的$\tilde{h}_t$ 进行选择性“记忆”。

$$y_t=\sigma(W_o\cdot h_t)。\tag{5}$$

需要说明的是，[]用来代表有 2 个向量相连，* 是 Hadamard Product，代表操作矩阵中对应的元素相乘，此时要求 2 个相乘矩阵是同型的，+表示矩阵加法操作的进行，σ 为 sigmoid 函数，利用 sigmoid 函数能够将数据处理为 0～1 范围内的数值，从而来充当门控信号。激活函数 tanh 能够帮助调节流经网络的值，而且 tanh 函数的输出值一直在区间(−1,1)内。

在输出层中，计算 loss 使用的是 softmax 的交叉熵(labels 和 logits)+平均值。

2.3　构造一个 PHNet 并生成随机游走

使用基于元路径的随机游走操作来捕捉不同节点间的关系，即通过论文标题、论文作者、论文发表期刊，构建 PHNet(异构网络)矩阵。本文所构建的异构网络中的节点类型只有论文一种，关系类型为 3 种(合著作者、共同标题、共同发表期刊)。在一个 PHNet 中，2 个论文节点之间可以通过多个无向关系进行连接，由这些无向关系连接的节点序列可以看作是从论文到论文的表述。受网络嵌入 DeepWalk 和 Metapath2Vec 方法的启发，利用随机游走策略和跳跃图模型学习网络节点表示。本文提出了一种元路径和关系权值引导的随机游走策略，用于加权异构网络上的采样路径。

元路径通过异构关系捕获节点间的相关性，在异构网络嵌入中得到了广泛的应用。本文在采样路径上考虑了 PHNet 中关系的权值，从直观上看，两个节点之间的关系值越大，它们之间的相似性就越大。在每一步游走中，当游走到一个邻居时，连接当前节点到邻居节点的关系值越高，就越有可能对该邻居进行采样。具体来说，本文依次选择 PHNet 中的一个论文节点作为路径的第一个节点，生成一个长度为 100 的元路径，然后选择最后一个节点作为另一条元路径的第一个节点。每个随机递归采样网络中的节点，都会生成一条由论文节点引导的长路径，直到满足固定长度，最后生成的结果输入到 WMRW.txt 文档中，如图 5 所示。

```
1  i96347 i1212386 i1323265 i1323265 i176218 i176718 i96145 i176904 i176783 i539689 i677790
2  i575075 i575075 i224475 i223846 i1324718 i659453 i659453 i659453 i408893 i409001 i1126126
3  i1214718 i120860 i1095864 i1092660 i1092660 i998486 i1000406 i605656 i605656 i1025139 i47
4  i96043 i1214791 i1214785 i1214686 i1214686 i1214740 i630505 i18145 i325752 i319197 i42333
5  i1237536 i1237536 i1492306 i1492306 i921863 i921863 i336488 i1343060 i1095815 i625959 i10
6  i97245 i289824 i595994 i1213967 i1213967 i1655596 i1655596 i1655596 i233075 i233168 i2331
7  i550354 i860876 i226312 i226542 i226518 i550199 i550067 i550067 i226124 i226121 i226121 i
8  i482356 i482356 i705807 i703183 i16932 i16932 i16785 i16932 i973533 i974207 i412667 i1122
9  i50575 i276204 i532920 i532803 i100291 i106105 i106105 i106105 i1126001 i1126452 i1304542
10 i430562 i189873 i411139 i410241 i997534 i1266579 i1266579 i1266579 i1143424 i1143463 i109
11 i8679 i445645 i989921 i988488 i483274 i233864 i234062 i234062 i1312773 i1312773 i1312773
12 i1071208 i1183343 i1305255 i1305255 i1241308 i142578 i195202 i941984 i942573 i942573 i113
13 i517996 i517996 i767013 i1754865 i962404 i860876 i860874 i860876 i813356 i814767 i273469
14 i861023 i861023 i403086 i403308 i446141 i884949 i885080 i1098153 i227665 i225559 i225565
15 i1230129 i1123199 i834546 i833718 i649717 i835113 i1400792 i1400792 i402523 i400791 i4088
16 i569826 i569826 i1210353 i335370 i335370 i849249 i854561 i854561 i1301918 i730702 i124319
17 i158817 i8253 i8253 i8623 i8623 i8623 i8762 i445645 i35970 i35250 i35250 i1655596 i165559
18 i1385434 i1385434 i27604 i27565 i303513 i225406 i225406 i225406 i300057 i300028 i300028 i
19 i1213105 i1213105 i733419 i733154 i733154 i722537 i722455 i1365556 i1750633 i1747152 i174
20 i113945 i113945 i1168025 i1168001 i1168001 i1168001 i1168040 i792009 i122627 i123166 i138
21 i325752 i325752 i1388852 i1390589 i464050 i1390589 i1390589 i464050 i474744 i476958 i4769
22 i489968 i489968 i302262 i302092 i1319502 i307748 i307812 i307812 i265762 i265114 i265114
23 i365316 i1430494 i147056 i146271 i146271 i896946 i1224452 i1224452 i1108984 i1109012 i110
24 i1358126 i819862 i822629 i160918 i1163164 i1163178 i314892 i880831 i879903 i879903 i44564
```

图 5　WMRW.txt 部分内容

Fig.5　WMRW.txt partial content

2.4 基于元路径异构网络嵌入

当前进行网络研究应用较多的是同构网络。若要把基于同构信息网络的方法用在异构信息网络中，需要将异构网络映射为同构网络，或者忽略节点间的连接信息，只是上述这 2 种方法都将会产生信息丢失的情况。因此，直接在异构信息网络上进行数据挖掘的方法是非常必要的。由于在异构信息网络中节点的连接是通过不同的语义意义，从而提出最好充分利用异构信息网络的网络模式期盼。网络模式即是了解信息网络的元结构，能够对网络的检索和数据挖掘进行指导，对于分析和理解网络中对象和关系的语义意义大有帮助。简单而言，就是一种基于元路径的方法。元路径就是在网络模式上加以定义的路径，代表了在 2 个对象类型之间的关系，同时能够定义实体之间新的或现存的关系。

现实世界中普遍存在着异构信息网络，本文选用的 DBLP 数据集是非常经典的异构网络，包含了 4 类实体：Paper，Venue，Author，Term。对于每篇论文，它都有一组 4 类实体的连接。此网络也包含了一些论文的信息，即论文之间有论文引用的论文集合。图 6—图 8 为学术网络与元路径示意图。

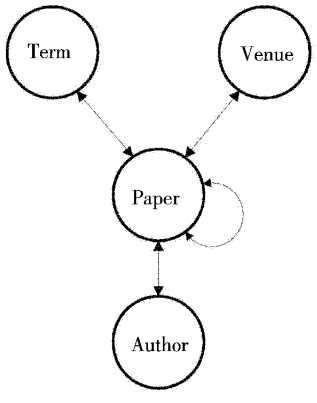


图 6 学术网络

Fig.6 Academic network

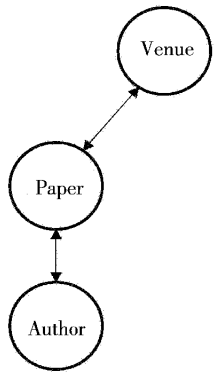


图 7 元路径：APV

Fig.7 Meta-path: APV

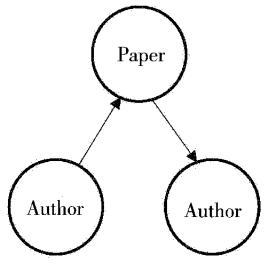


图 8 元路径：APA

Fig.8 Meta-path: APA

为了将异构网络结构合并到 skip-gram 中，提出了在异构网络中基于元路径的随机游走。与传统的方法相比，潜在空间表示学习的优势在于即使没有连接元路径，也能够对节点之间的相似性进行建模。在嵌入时定义每次扫描的数据大小为 128，嵌入向量的维数为 64，上下文取得词的个数为 2，每次移动的窗口大小为 1，负样本的个数为 5，定义完毕后度量当前词向量与其他词向量的相似度，采用余弦定理计算，完成重名作者的歧义消解工作。

2.5 评估结果

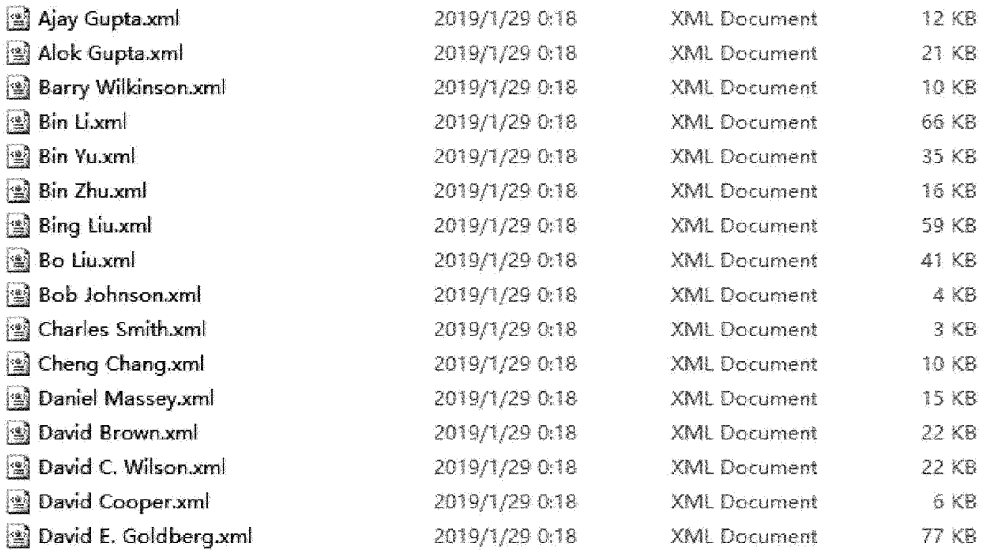
评估指标为精确度、召回率、F_1 值，其中精确度和召回率中对 TP，TP_FP 和 TP_FN 的定义是：TP 为正确预测到同一作者的配对，TP_FP 为对同一作者的预测总对数，TP_FN 为同一作者的总对数。

精确度 precision=TP/TP_FP

召回率 recall=TP/TP_FN

F_1 f1=(2 * precision * recall)/(precision+recall)。

实验结果部分截图如图 9、图 10 所示。图 10 中 name 一列为实验数据集中歧义作者名，可与图 9 生数据集相对照，每一个有歧义的人名归结为一个 XML 文档。



Ajay Gupta.xml	2019/1/29 0:18	XML Document	12 KB
Alok Gupta.xml	2019/1/29 0:18	XML Document	21 KB
Barry Wilkinson.xml	2019/1/29 0:18	XML Document	10 KB
Bin Li.xml	2019/1/29 0:18	XML Document	66 KB
Bin Yu.xml	2019/1/29 0:18	XML Document	35 KB
Bin Zhu.xml	2019/1/29 0:18	XML Document	16 KB
Bing Liu.xml	2019/1/29 0:18	XML Document	59 KB
Bo Liu.xml	2019/1/29 0:18	XML Document	41 KB
Bob Johnson.xml	2019/1/29 0:18	XML Document	4 KB
Charles Smith.xml	2019/1/29 0:18	XML Document	3 KB
Cheng Chang.xml	2019/1/29 0:18	XML Document	10 KB
Daniel Massey.xml	2019/1/29 0:18	XML Document	15 KB
David Brown.xml	2019/1/29 0:18	XML Document	22 KB
David C. Wilson.xml	2019/1/29 0:18	XML Document	22 KB
David Cooper.xml	2019/1/29 0:18	XML Document	6 KB
David E. Goldberg.xml	2019/1/29 0:18	XML Document	77 KB

图 9　生数据集

Fig.9　Raw data set

name	Prec	Rec	F1	Actual	GHAC
Avg	0.865164	0.791973	0.814994	0	0
Ajay Gupt	0.532468	0.689076	0.600733	9	9
Alok Gupt	0.91762	0.954762	0.935823	2	2
Barry Wil	1	1	1	1	1
Bin Li	0.907767	0.685819	0.781337	60	60
Bin Yu	0.926267	0.648387	0.762808	17	17
Bin Zhu	0.92	0.807018	0.859813	15	15
Bing Liu	0.9607	0.52138	0.675928	18	18
Bo Liu	0.891304	0.501222	0.641628	47	47
Bob Johns	0.6	0.6	0.6	7	7
Charles S	1	1	1	4	4
Cheng Cha	0.804878	0.814815	0.809816	5	5
Daniel Ma	0.997446	0.907085	0.950122	2	2
David Bro	0.986301	0.977376	0.981818	25	25
David C.	0.99466	0.899758	0.944832	5	5
David Coo	0.9	0.9	0.9	7	7
David E.	0.991266	0.991266	0.991266	3	3

图 10　实验结果部分截图

Fig.10　Screenshots of experimental results

3　实验结果分析

本文使用 DBLP 数据集进行实验，有歧义的人名为 101 个，论文出版物有 7 585 篇，其中包含的节点特征有作者 id，作者名以及出版物的详细信息。详细信息包含：论文标题、出版年份、作者（论文所有的作者）、出版期刊、出版物 id、作者所属单位。因较多人的所属单位信息为空白，所以该特征属性在本次消歧任务中不作为侧重点。本次实验整理数据侧重于利用论文标题、作者集合、出版物期刊名称、出版年份和 id 编号等特征属性进行消歧操作。

为了验证本文所提出方法的消歧性能，将其与另外 4 种方法进行比较，这 4 种方法包括：DeepWalk，LINE，Node2Vec 和 PTE，都是目前最先进的顶点嵌入方法。为了公平起见，所有这些方法都使用相同的数据来实现姓名消歧。

DeepWalk:DeepWalk 是近期所提出的一种网络嵌入方法。在给定论文合作关系的情况下用来捕获与关联文档集合中的一对人员之间的协作,并采用均匀随机游走的方法来获取其邻域的上下文信息进行文档嵌入。

LINE:LINE 不再采用随机游走的方法,它在图上定义一阶相似度和二阶相似度,对节点的信息进行了补充,从而得到更丰富的节点嵌入。

Node2Vec:和 DeepWalk 近似,Node2Vec 为实现文档嵌入设计了一个有偏差的随机游走过程。

PTE:预测性文本嵌入框架的目标是捕获词-词、词-文档和词标签之间的关系。可是,该种方式不能捕捉文档间的连接信息。

表 1 显示了本论文所提出的方法与对比方法在处理多个不同人名姓名歧义消除方面的性能(表 1 用于 DBLP 数据集)。在表 1 中,列 1 为需要消歧的作者姓名,第 3 列—第 6 列为各种方法的 F_1 值。F_1 值表示各种方法给定姓名数据集下的消歧性能。最后一列显示了本文所提出的方法相较于对比方法的改进水平。

表 1 F_1 值比较

Tab.1 F_1 value comparison

姓名	F_1					
	本文方法	Deepwalk	LINE	Node2Vec	PTE	提高/%
Bin Yu	0.763	0.610	0.643	0.531	0.399	12.0
Rakesh Kumar	0.926	0.617	0.641	0.372	0.219	28.5
Lei Wang	0.697	0.419	0.639	0.263	0.447	5.8
Bin Li	0.781	0.392	0.641	0.186	0.349	14.0
Yang Wang	0.704	0.640	0.623	0.331	0.444	6.4
Yu Zhang	0.669	0.454	0.658	0.196	0.385	1.1
David Brown	0.981	0.494	1.000	0.221	0.575	40.6
Wei Xu	0.825	0.228	0.599	0.136	0.236	22.6

表 1 表明,本文方法相较于对比方法的总体改进比较大。PTE 的表现很差,因为它没有将相关的结构信息整合到实验中。DeepWalk 的方法忽略了边缘权值,这一点恰恰在异构学术网络中是非常重要的。这几种基于嵌入的对比方法都不能利用多个网络信息来处理消歧任务,本论文的模型利用了这一点,提出了基于元路径异构网络嵌入实现姓名消歧的方法,这可能是该方法优于现有的基于网络嵌入方法的一个重要原因。

4 结 语

笔者提出了一个有效解决作者姓名消歧问题的框架。该框架对 DBLP 数据集中有待消解歧义的作者姓名的数据集进行了预处理操作,利用 word2vec 模型进行嵌入,再输入到 GRU 网络中进行训练,根据节点间的关系构造了 PHNET 网络,最后基于元路径异构网络嵌入实现姓名消歧。该方法所提出的表示学习方案比其他现有的网络嵌入方法能更有效地将属于同名作者的文档进行消歧处理。实验结果验证了该方法的可行性和有效性。

本研究虽实现了预期目标,但是在组合不同类型的特征属性(如利用文本信息的语义信息和离散特征)来学习有待消歧作者论文的有效表示方面仍有进步空间。在未来的工作中,将尝试把此方法应用于分布式计算系统,进一步提高大型学术数据库的消歧速度和效果。

参考文献/References:

[1] DENG H, KING I, LYU M R. Formal models for expert finding on DBLP bibliography data[C]//Eighth IEEE International Conference on Data Mining. [S.l.]: [s.n.], 2008: 163-172.

[2] HUANG Zhixing, YAN Yan, QIU Yuhui, et al. Exploring emergent semantic communities from DBLP bibliography database[C]// International Conference on Advances in Social Network Analysis and Mining. [S.l.]: [s.n.], 2009: 219-224.

[3] FRANCESCHET M. Collaboration in computer science: A network science approach[J]. Journal of the American Society for Information Science and Technology, 2011, 62(10): 1992-2012.

[4] KIM J, KIM H, DIESNER J. The impact of name ambiguity on properties of coauthorship networks[J]. Journal of Information Science Theory and Practice, 2014, 2(2): 6-15.

[5] CAVERO J M, VELA B, CACERES P. Computer science research: More production, less productivity[J]. Scientometrics, 2014, 98(3): 2103-2111.

[6] SHI Quan, XU Bo, XU Xiaomin, et al. Diversity of social ties in scientific collaboration networks[J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(23/24): 4627-4635.

[7] REITZ F, HOFFMANN O. Learning from the past: An analysis of person name corrections in the DBLP collection and social network properties of affected entities[J]. Social Network Analysis and Mining, 2013,6: 427-453.

[8] 余传明，林奥琛，钟韵辞，等.基于网络表示学习的科研合作推荐研究[J]. 情报学报，2019，38(5)：500-511.
YU Chuanming, LIN Aochen, ZHONG Yunci, et al. Research of author name disambiguation based on network embedding[J]. Journal of the China Society for Scientific and Technical Information, 2019, 38(5): 500-511.

[9] GARFIELD E. British quest for uniqueness versus American egocentrism[J]. Nature, 1969, 223(5207): 763-763.

[10] LEY M. DBLP: Some lessons learned[J]. Proceedings of the VLDB Endowment, 2009, 2(2): 1493-1500.

[11] KIM J. Evaluating author name disambiguation for digital libraries: A case of DBLP[J]. Scientometrics, 2018, 116(3): 1867-1886.

[12] HAZIMEH H, YOUNESS I, MAKKI J, et al. Leveraging co-authorship and biographical information for author ambiguity resolution in DBLP[C]/Advanced Information Networking and Applications (AINA). [S.l.]: [s.n.], 2016: 1080-1084.

[13] 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. [S.l.]: [s.n.], 2004: 296-305.

[14] GILES C L, ZHA H, HAN H. 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). [S.l.]:[s.n.], 2005: 334-343.

[15] MALIN B. Unsupervised name disambiguation via social network similarity[C]//Workshop on Link Analysis, Counterterrorism, and Security[S.l.]: [s.n.], 2005:93-102.

[16] ZHANG Baichuan, AL-HASAN M. Name disambiguation in anonymized graphs using network embedding[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. [S.l.]:[s.n.], 2017: 1239-1248.

[17] PERZZI 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. [S.l.]:[s.n.], 2014: 701-710.

[18] 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. [S.l.]: International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[19] TANG Jian, QU Meng, MEI Qiaozhu. 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. [S.l.]:[s.n.], 2015: 1165-1174.

[20] GROVER A, LESKOVEC J. Node2vec: Scalable feature learning for networks[J]. Knowledge Discovery and Data Mining, 2016: 855-864.

[21] PHAM T H, PHAM X K, NGUYEN T A, et al. NNVLP: A neural network-based Vietnamese language processing toolkit[C]//International Joint Conference on Natural Language Processing. [S.l.]:[s.n.], 2017: 37-40.

[22] WU Fangzhao, LIU Junxin, WU Chuhan, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[J]. The World Wide Web Conference, 2019: 3342-3348.

[23] 甄然，于佳兴，赵国花，等.基于卷积神经网络的无人机识别方法仿真研究[J]. 河北科技大学学报，2019，40(5)：397-403.
ZHEN Ran, YU Jiaxing, ZHAO Guohua, et al. Simulation research on UAV recognition method based on convolutional neural network [J]. Journal of Hebei University of Science and Technology, 2019, 40(5): 397-403.

[24] 纪志强，魏明，吴启蒙，等.基于递归神经网络的 TVS 电磁脉冲响应建模[J]. 河北科技大学学报，2015，36(2)：157-162.
JI Zhiqiang, WEI Ming, WU Qimeng, et al. EMP response modeling of TVS based on the recurrent neural network[J]. Journal of Hebei University of Science and Technology, 2015,36(2): 157-162.