基于合作作者与隶属机构信息的同名排歧方法

尚玉玲[1] 曹建军[2] 李红梅[1] 郑奇斌[1]

(解放军理工大学指挥信息系统学院 南京 210007)[1] (国防科技大学第六十三研究所 南京 210007)[2]

摘 要 同名排歧是实体分辨领域的重要研究内容之一,其旨在分辨出相同姓名对应的不同人。针对传统同名排歧方法需要丰富的信息以及无法解决信息缺乏时的排歧问题,提出了一种基于合作作者和隶属机构信息的同名排歧方法。根据作者间的合作关系以及作者与机构间的隶属关系构造实体关系图,采用广度优先搜索策略搜索图中两两同名作者间的有效路径;根据有效路径长度、数目及路径上边的类型,计算两个同名作者间的连接强度,并将其与阈值进行比较,实现同名排歧。实验结果表明,所提方法比当前最好的方法具有更好的同名排歧效果,且能够实现单一作者的同名排歧。

关键词 数据质量,实体分辨,同名排歧,有效路径,连接强度

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.11.034

Co-author and Affiliate Based Name Disambiguation Approach

SHANG Yu-ling[1] CAO Jian-jun[2] LI Hong-mei[1] ZHENG Qi-bin[1]

(College of Command Information Systems,PLA University of Science and Technology,Nanjing 210007,China)[1]

(The 63rd Research Institute,National University of Defense Technology,Nanjing 210007,China)[2]

Abstract Name disambiguation is one of the most challenging issues in entity resolution domain,and it aims at solving the problem that the same name is shared by different people. However,most of the conventional approaches rely heavily on sufficient information of entities,and fail to realize the name identification with insufficient information. This paper proposesd a novel name disambiguation approach based on co-authors and authors' affiliates. Specifically,entity relationship diagram is constructed based on co-authorship and authors' affiliates,and the breadth-first search scheme is utilized to search the effective path between each pair of authors with the exactly same name in the constructed entity relationship diagram. A unique metric connection strength between authors is calculated according to the length of effective path,the number of effective path and the type of edge on path. And it is compared with the threshold to achieve name disambiguation. Experimental results show that the proposed approach is better than the state-of-the-art approaches,and it is able to disambiguate the authors sharing the same name without co-authorship.

Keywords Data quality,Entity resolution,Name disambiguation,Effective path,Connection strength

1 引言

一个或多个数据集中,同一实体可能存在不同的表示形式。不同实体也可能存在相同的表示形式,实体分辨(Entity Resolution)是提高数据质量的关键技术之一,目的是正确分辨出同一客观实体的全部不同数据对象表示[1]。名称分辨(Name Resolution)是实体分辨的一项重要研究内容,主要有人名、地名和机构名等实体的分辨。人名分辨有两种,相同姓名对应不同人的分辨(即同名排歧(Name Disambiguation,ND)),以及同一人对应多个不同姓名表示形式的分辨。本文主要针对同名排歧问题进行研究。

当前,人名分辨的方法很多,主要可分为有监督的方法、无监督的方法及其他方法。

有监督的方法一般是采用带有类标号的数据训练分类模型实现同名排歧。文献[2]提出了一种自训练的同名排歧方法,该方法首先选取稀少姓氏的共同合作作者记录和具有两个及以上较常见姓氏的合作作者的记录作为训练数据,以生成区分度函数,并用于对尚未分类的作者进行指派,实现同名排歧。文献[3]提出了一种结合用户反馈的有监督同名排歧方法,该方法首先用生成数据训练多个分类器,并人工组合这些分类器以达到较高的分类正确率;再根据类间的相似度进行类合并;最后根据用户反馈信息进行二次合并,实现同名排歧。

无监督的方法根据同名作者间的相似度,采用聚类的方

到稿日期:2017-10-25 返修日期:2017-12-07 本文受国家自然科学基金(61371196),中国博士后科学基金(2015M582832)资助。

尚玉玲(1990-),女,硕士生,主要研究方向为数据质量控制与数据治理,E-mail:1533765046@qq.com;**曹建军**(1975-),男,博士,副研究员,主要研究方向为数据质量控制与数据治理、数据智能分析与应用,E-mail:jianjuncao@yeah.net(通信作者);**李红梅**(1990-),女,博士生,主要研究方向为个性化推荐;**郑奇斌**(1990-),男,博士生,主要研究方向为数据质量控制与数据治理。

法实现同名排歧。文献[4]提出了一种无监督的启发式层次聚类的同名排歧方法，该方法将具有相同合作作者的记录划分到同一簇以提高效率和效果；再根据作者的研究领域、出版地名称的相似度合并簇，实现同名排歧。文献[5]提出了一种消除 DBLP 中同名作者的排歧方法，该方法将具有相同姓名的文献分到同一块，然后构造所有作者结点和文献结点的网络，将两篇文献间的最短路径作为两者间的距离，并与阈值进行比较，实现同名排歧。文献[6]提出了一种动态增长数据库的同名排歧方法，该方法根据作者姓名及其合作作者、研究领域和论文的出版地等信息，计算与库中同名作者记录的相似度，确定需要插入库中的类簇，实现同名排歧。文献[7]提出了一种只使用合作作者信息的同名排歧框架 GHOST。GHOST 根据作者间的合作关系构造图；根据图中待排歧作者间有效路径的数目和长度计算两两待排歧作者间的并联电阻相似度；最后对相似度矩阵聚类，实现同名排歧。文献[8]针对动态增加的数据，采用 DBSCAN 和随机森林相结合的方法，通过设置实例层约束条件和聚类层约束条件实现同名排歧。文献[9]提出了一种专利发明人的同名排歧方法，根据专利发明人、机构、合作发明人、代理人、团体、题目等属性进行同名排歧。该方法采用训练随机森林分类器进行两两同名者间的排歧，根据两两排歧结果进行分块并进行 DBSCAN 聚类，实现同名排歧。文献[10]根据作者间的合作关系构造合作关系图，通过图中同名作者间的有效路径数计算相似度；再根据论文摘要内容的 TF-IDF 模型计算相似度；最后分别对两相似度矩阵进行聚类及合并，实现同名排歧。文献[11]采用同名作者的合作作者、题目、期刊 3 个属性计算合作关系图中的“多路径游走”或作者-题目二部图的“P-SimRank”、题目间(TF-IDF 模型)的余弦相似度及期刊名称间的共用词相似度，并对 3 个相似度加权后的矩阵进行聚类，实现同名排歧。

其他同名排歧方法包括有监督和无监督的混合方法。文献[12]是有监督和无监督的混合同名排歧方法，该方法用无监督方法选择训练数据，用混合监督方法学习分类模型，最后根据用户反馈信息迭代实现同名排歧。文献[13]提出了一种 DISTINCT 同名排歧方法。DISTINCT 根据同名作者记录间的合作作者、研究领域、期刊或会议名称、出版地等属性定义了邻接元组，并根据随机游走概率和邻接元组间的 Jaccard 系数得到两同名作者间的连接强度；选取稀少姓氏姓名的训练数据训练 SVM 分类器，得到不同路径的权重，根据连接路径及其权重得到两同名作者间的相似度；最后对相似度矩阵进行聚类，实现同名排歧。文献[14]提出了一种基于多核函数的实体链接同名排歧方法，并通过网络验证来提高排歧的准确性。文献[15]通过将实体链接到本体库或诸如 KIM、OpenCyc、维基百科等知识库，实现同名排歧。文献[16]主要针对微博中的同一人对应多个不同姓名表示形式的歧义性问题进行研究，提出了基于标签消歧算法和中文姓氏的二分类器姓名排歧方法，通过构造中文姓氏表，以及实体与百度百科、维基百科中实体的映射表，实现姓名排歧。文献[17]将待排歧同名作者表示成“编号、作者单位、关键词、合作作者、期刊及标题”的六元组，通过计算单个属性间的字符串模糊匹配相似度，并比较它们加权后的值与阈值的大小，来实现同名排歧。文献[18]针对英文作者姓名的歧义性，采用文献记录中作者的姓、名、隶属机构、合作作者、题目、期刊属性，计算两条记录的各个属性值间的 Jaccard 相似度或 Jaro-Winkler 字符串相似度，并将其与训练得到的阈值进行比较，实现姓名排歧。

上述方法尽管能够实现同名排歧，但所需信息较多，获取不易，缺乏普适性；有的方法虽然所需信息较少，如 GHOST 同名排歧框架，但无法解决单一作者时的同名排歧。事实上，通过对中国知网和万方数据库的随机抽样发现，单一作者的文献记录竟高达 30%。因此，本文同时使用作者间的合作关系信息及作者的隶属机构信息，提出一种基于合作作者和隶属机构信息的同名排歧(Co-Author and Affiliate based Name Disambiguation，CoAAND)方法，以解决传统方法所需信息较多或难以实现单一作者的排歧的问题。

2　基于合作作者和隶属机构的同名排歧

DISTINCT 方法在排歧过程中使用属性较多，没有充分利用记录间的关系信息，效果较差；GHOST 框架只使用了合作作者信息，不能解决没有合作作者时的同名排歧。相比获取作者的邮箱、研究方向、主页等信息，作者的隶属机构较易获取，因此本文基于合作作者和隶属机构信息，提出了一种新的同名排歧方法 CoAAND，以更好地解决单一作者文献记录中的同名排歧问题。

CoAAND 方法主要分 4 步：1)根据作者间的合作关系及作者与机构间的隶属关系构造图；2)搜索图中待排歧作者间的有效路径；3)根据每条有效路径对连接强度的重要程度(路径权重)及其边的类型、各个结点所连边的数目(路径概率)计算连接强度；4)比较连接强度与阈值的大小，实现同名排歧。

2.1　构造实体关系图

在构造图之前，先介绍多集的概念及其运算法则[19]。普通的集合要满足互异性，集合中不允许出现重复元素，而多集是允许出现重复元素的集合。在实际应用中，集合或多集的元素个数一般是有限的，因此，多集 A_i 可以表示为 $A_i=\{w_{i1}a_1,w_{i2}a_2,\cdots,w_{im}a_m\}$，其中 w_{ik} 为 $a_k(0\leqslant k\leqslant m,m\in\mathbf{N})$ 在 A_i 中的重复次数，$w_{ik}=0$ 表示 A_i 中不存在 a_k。多集的运算与普通集合类似，又略有不同。若 $A_1=\{a,5b,3c,6d,e\}$，$A_2=\{3a,2b,5c\}$，则 $A_1\cup A_2=\{3a,5b,5c,6d,e\}$，$A_1\cap A_2=\{a,2b,3c\}$，$|A_1|=16$，$|A_2|=10$。下面构造实体关系图。

将得到的包括论文题目(编号)、作者及其隶属机构信息的文献记录表示成图 $G=(V,E)$，其中 V 为顶点的集合，$V=A\cup O$，$A=\{a_1,a_2,\cdots,a_m,\cdots,a_n\}$ 为全部作者姓名的集合(排歧时，对待排歧作者的姓名进行编号以便区分)，$O=\{o_1,o_2,\cdots,o_s\}$ 为作者所隶属的机构名称集合；$E=\{S_{ij}\mid 0<i\leqslant n,0<j\leqslant n\}\cup ns$ 为边的多集，S_{ij} 是作者 a_i 和 a_j 合作过的所有论文的集合；$s=\{1,0\}$ 为隶属关系类型的边，$n=|A|$，ns 为隶属关系类型边的多集表示。

假设有表 1 所列的论文、作者及其隶属机构记录信息，则可构造实体关系图，如图 1 所示。

表 1 论文、作者及其隶属机构样例

Table 1 Sample of papers, authors and their affiliates

论文	作者及其隶属机构
p_1	李娜(o_1),魏永巨(o_1),秦身钧(o_1)
p_2	李娜(o_1),魏永巨(o_1),秦身钧(o_1),张玉平(o_1)
p_3	李艳廷(o_1),魏永巨(o_1),秦身钧(o_1),申金山(o_1)
p_4	李娜(o_1),李一凡(o_2)
p_5	李娜(o_1),李一凡(o_2),张银霞(o_2),王红(o_2)
p_6	李娜(o_1),李一凡(o_2),张银霞(o_2),王红(o_2)
p_7	李一凡(o_2),张银霞(o_2),王红(o_2)
p_8	李一凡(o_2),张银霞(o_2),王红(o_2)
p_9	李娜(o_3)
p_{10}	李娜(o_3)
p_{11}	李娜(o_4)
p_{12}	李娜(o_5)

注:o_1 为河北师范大学化学学院;o_2 为河北政法职业学院;o_3 为河北师范大学法政学院;o_4 为河北师范大学;o_5 为河北师范大学文学院

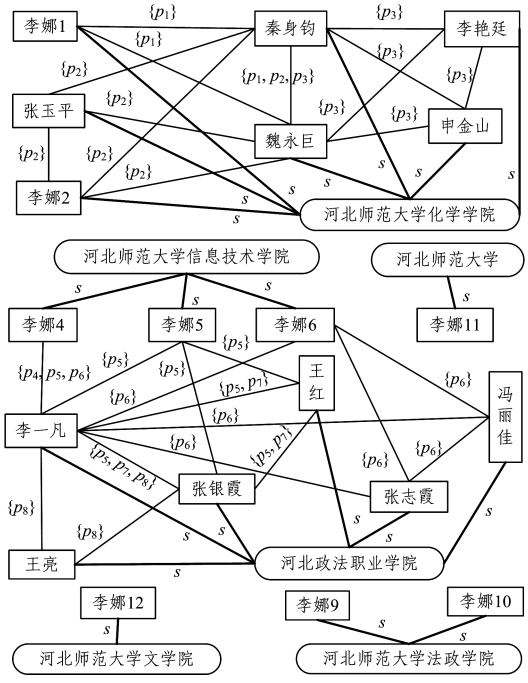

图 1 作者-机构实体关系图

Fig. 1 Author-affiliate entity relationship diagram

图 1 中,$V=A\cup O$ 为顶点集合,A 为作者姓名的集合,$A=${李娜 1,李娜 2,李娜 4,李娜 5,李娜 6,李娜 9,李娜 10,李娜 11,李娜 12,魏永巨,秦身钧,张玉平,申金山,李艳廷,李一凡,张银霞,王红,冯丽佳,张志霞},"李娜"为待排歧作者姓名,$O=\{o_1,o_2,o_3,o_4,o_5\}$ 为机构名称的集合;E 为边的多集,$E=\{2\{p_1\},5\{p_2\},5\{p_3\},\{p_1,p_2,p_3\},\{p_4,p_5,p_6\},3\{p_5\},6\{p_6\},2\{p_5,p_7\},\{p_5,p_7,p_8\},2\{p_8\},20\{1,0\}\}$。

假设作者的研究兴趣及其隶属机构不变,则他一定有稳定的合作团队和隶属机构。若两个同名作者不是同一人,他们的合作作者极有可能没有交叉,隶属机构也极可能不同(同一机构出现两个姓名完全相同的作者的概率是很小的,如对于院校等较大的机构,取较小的机构名称粒度可降低这一概率),因此图中这两个同名作者间几乎没有路径连接。反之,他们一般会有相对稳定的共同合作作者和隶属机构,在图中有路径连接,且路径数越多、长度越短,两者对应同一人的可能性越大。

2.2 选择有效路径

在实体关系图中,路径存在与否是两个作者关联性的直接反映,两个作者间的路径数目、长度及合作论文的篇数直接反映了作者的相关程度。本文根据作者间的路径实现同名排歧,下边给出有效路径的定义。

定义 3(有效路径[7]) 对于实体关系图 G 中的两顶点 a_s 和 a_e 间的一条无回路路径 P,$A=\{a_s,a_1,a_2,\cdots,a_n,a_e\}$ 为 P 上的顶点序列,记 P 以 a_s 或 a_i 为起点,a_{i+1} 为下一顶点,a_{i+2} 或 a_e 为 a_{i+1} 的下一个顶点,则有效路径为不满足 $|S_{i(i+1)}|=1$,$|S_{(i+1)(i+2)}|=1$ 且 $S_{i(i+1)}=S_{(i+1)(i+2)}$ 的路径。

由图的构造过程不难发现,对于任意两个作者 a_i 和 a_j,若他们与 a_k 合作完成了论文 p_c,且 S_{ik} 和 S_{jk} 中都只有 p_c,那么 S_{ij} 一定有 p_c[7]。只合作一篇论文不能保证两个作者具有稳定的合作关系,对连接强度的贡献较小,路径存在冗余。如图 2 所示,"秦身均"和"李娜 2"间的路径"秦身均-李娜 2"和"秦身均-张玉平-李娜 2"明显冗余,这就使得路径的连接强度被重复计算,影响排歧效果。反之,若一条路径上连续的两条边上多于一篇文献,说明两个作者具有稳定的合作关系,则认为该路径是有效路径。

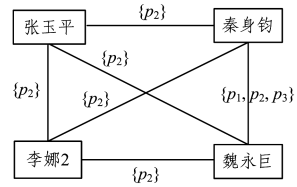

图 2 实体关系图例

Fig. 2 Sample of entity relationship diagram

采用广度优先搜索(BFS)策略搜索图中两个顶点间的有效路径。因 BFS 的时间复杂度为 $O(|V|+|C|)$,若同名作者的个数为 r,则共有 $r(r-1)/2$ 个两两同名作者对,时间复杂度为 $O((|V|+|C|)r^2)$。事实上,一篇论文的合作作者一般为 4 个左右,而路径越长,时间复杂度越高,且对排歧的贡献越小。已知"六度分割理论",即任何两个陌生人之间,最多通过 6 个人便能够互相认识对方[20],可见,路径过长不一定会提高排歧效果,因此限定路径长度可以提高搜索效率,且几乎不影响排歧效果。

2.3 计算连接强度

在实体关系图中,通常两实体间的路径越多、长度越短,两实体越相似[1]。同样,本文中两同名作者间的有效路径越短、数目越多,两者是同一人的可能性越大。基于 CoAAND 方法的图中同时有合作关系和隶属关系类型的边,本节区分边的类型以计算连接强度。

2.3.1 连接强度

连接强度是两顶点间全部有效路径的路径连接强度之和。记 R_L 为 a_i 和 a_j 间有效路径的集合,$P\in R_L$,cs_P 为 P 的连接强度,则 a_i 和 a_j 间的连接强度为:

$$cs(a_i,a_j)=\sum_{P\in R_L} cs_P \tag{1}$$

路径 P 的连接强度 cs_P 为:

$$cs_P=w_P\cdot p_P \tag{2}$$

其中,p_P 为路径概率,由 P 所含结点数及与顶点连接的边数决定;w_P 为路径权重,由 P 所含的边数决定,是该路径对两

顶点间连接强度重要程度的直接反映。

2.3.2　路径概率

路径概率是由该路径所含结点数及与其连接的边数决定的。路径越短，与其上顶点连接的边越少，对应的路径概率越大。

文献[13]提出了路径传播概率，但不区分边的类型，根据当前顶点的度计算选择某个邻接顶点的正向概率，并根据其下一顶点的度计算当前顶点的反向概率。如图 3 所示的路径 $P:a_i-a_j-a_e$，正向概率 $p(a_i-a_j)=1/5$，$p(a_j-a_e)=1/4$，则 $p(P)=p(a_i-a_j)p(a_j-a_e)=1/20$；反向概率 $p'(a_e-a_j)=1/2$，$p'(a_j-a_i)=1/4$，则 $p'(P)=p'(a_j-a_i)p'(a_e-a_j)=1/8$。

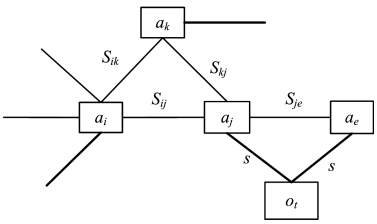

图 3　作者及其隶属机构关系示意图

Fig. 3　Relationship diagram of authors and their affiliates

图 3 中，若 a_i，a_k，a_j，a_e 分别为不同的作者顶点，o_t 为机构名称顶点，则该图中有两种不同类型的边，细线为合作关系类型边，粗线为隶属类型边。因反向概率计算复杂且贡献较小，本文只考虑正向概率，且计算时只考虑与从当前顶点到下一顶点类型相同的边数，即 $p(a_i-a_j)=1/4$，$p(a_j-a_e)=1/3$，$p(P)=p(a_i-a_j)p(a_j-a_e)=1/12$。

定义 4(路径概率)　对于实体关系图中的两个顶点 a_s 和 a_e，P 为 a_s 到 a_e 的一条路径，$A=\{a_s,a_1,a_2,\cdots,a_n,a_e\}$ 为 P 上的顶点序列，$S=\{s_c,s\}$ 为 P 上的两种关系类型集合，其中 s_c 为合作关系类型，s 为隶属关系类型；记 P 中以 a_s 或 a_i 为起点，以 a_{i+1} 或 a_e 为终点的边的关系类型为 s_{a_i}，$s_{a_i}\in S$，$N(a_i,s_{a_i})$ 为与顶点 a_i 连接的 s_{a_i} 类型的边数，则 P 的路径概率为：

$$p_P=\prod_{a_i\in A-a_e}\frac{1}{N(a_i,s_{a_i})} \tag{3}$$

2.3.3　路径权重

路径权重是路径对两顶点间连接强度重要程度的直接反映，大小由该路径上的边数决定。

定义 5(路径权重)　实体关系图中，$S=\{s_c,s\}$ 为关系类型集合，n_c，n_t 分别为路径 P 上合作关系类型 s_c 和隶属关系类型 s 对应的边数，则 P 的路径权重 w_P 为：

$$w_P=\frac{1}{n_c+n_t} \tag{4}$$

由式(4)可知，路径权重与路径上所有类型的边数之和成反比，即路径越长，路径权重越小，对连接强度的贡献越小。

2.4　算法描述

根据 2.3 节得到待排歧作者间的连接强度，通过比较其与阈值的大小实现同名排歧。CoAAND 算法的伪代码如算法 1 所示。

算法 1　CoAAND 算法

输入：文献记录数据 Data、阈值 δ 和待排歧作者数 len

输出：对应同一人的作者集 C

```
1. 根据文献记录数据 Data 构造图 G;
2. For (i=1:len)
3.    For ( j=(i+1):len)
4.       搜索 a_i,a_j 间的全部有效路径,并存入有效路径集合 R_L;
5.       依据 R_L 计算 a_i,a_j 间的连接强度 sim(a_i,a_j);//可得矩阵 sim
6.    End;
7.    按行归一化得到的连接强度矩阵 sim,即 Sim_Min 为第 i 行的最
      小值,Sim_Max 为第 i 行的最大值,则第 i 行的第 j 个元素 SIM
      (a_i,a_j)=Sim_Min x sim(a_i,a_j)/( Sim_Max−Sim_Min);
8.    For ( j=(i+1):len)
9.       If SIM(a_i,a_j)>δ
10.         将(a_i,a_j)添加到对应同一人的作者对集合 Pairs 中;
11.      End;
12.   End;
13. End;
14. 对 Pairs 进行闭包运算,得到对应同一人的集合 C;
15. Return C.
```

3　实验验证

3.1　数据准备

为验证所提方法的有效性，使用待排歧作者的类标号进行比较。从万方数据库获取“李丹、李强、李伟、王鹏、王伟、王艳、杨静、张慧、张静、张伟”的文献记录，并标注类标号，具体步骤为：

1)从万方数据库中导出这些人名对应的作者文献记录，根据原始论文中作者的合作作者、邮箱、机构、性别、出生年份及研究方向等，逐一标注类标号，并丢弃无法确定类标号的记录。

2)因路径长度为 3，需导出步骤 1)中同名作者的合作作者文献记录，如相同的作者姓名 a_1，a_2 对应同一人，a_1 仅与 b 合作发表了论文，a_2 仅与 c 合作发表了论文，b 与 c 又合作发表了论文，那么 a_1 和 a_2 间便有路径“$a_1-b-c-a_2$”，长度为 3，即中间有两个作者顶点 b 和 c。

3)将从万方数据库中导出的文献记录预处理为“题目、作者、隶属机构”的结构化形式，并将作者的隶属机构名称保留到二级子机构，如大学机构名称保留到院系；对于只署名一级机构名称的情况，全部保留；对于作者署名多个机构名称的情况，则分别保留。

经过以上处理，得到表 2 所列的详细信息，其中，第 1－3 列分别为同名作者的姓名、包含的论文数(每篇论文对应一个待排歧同名作者)以及对应的不同作者数，第 4 列为合作作者数，第 5 列为作者的隶属机构数。

表 2　含单一作者的待排歧作者信息表

Table 2　Ambiguous authors' information with single author

姓名	论文数	作者数	合作作者数	机构数
李丹	74	20	881	227
李强	44	11	506	173
李伟	71	16	1723	537
王鹏	23	8	317	61
王伟	146	25	1666	373
王艳	25	8	255	68
杨静	150	10	1026	212
张慧	50	14	574	160
张静	48	12	440	137
张伟	40	5	261	60

因为 GHOST 同名排歧方法不适用于单一作者的情况，所以删除表 2 中单一作者的记录，得到表 3，该数据集只用于验证 GHOST 方法的有效性。

表 3 不含单一作者的待排歧作者信息表

Table 3 Ambiguous authors' information without single author

姓名	论文数	作者数	合作作者数	机构数
李丹	74	20	881	227
李强	34	10	506	172
李伟	60	15	1723	537
王鹏	20	7	317	60
王伟	135	25	1666	373
王艳	21	7	255	60
杨静	145	9	1026	211
张慧	36	13	574	159
张静	36	10	440	135
张伟	39	5	261	60

3.2 实验方法

为验证 CoAAND 方法的有效性和优越性，将其分别与 DISTINCT，GHOST 方法进行比较。实验设置如下。

方法 1：文献[6]的 DISTINCT 方法。将数据库表示成图；根据邻接元组间的相似性及图中两条记录间的随机游走概率度量两条记录间的相似度；用 SVM 训练得到邻接元组间的不同连接类型边的权重，得到待排歧的两条作者记录间的相似度；最后用近邻传播聚类实现同名排歧。

方法 2：文献[8]的 GHOST 方法。将数据库表示成图；广度优先搜索待排歧作者间的有效路径；根据并联电阻相似度公式计算待排歧同名作者间的相似度，得到相似度矩阵；用近邻传播聚类实现同名排歧。

方法 3：本文提出的 CoAAND 方法。

3.3 评价指标

用实体分辨领域常用的查准率(Precision)、查全率(Recall)和 $F1$ 值衡量同名排歧的效果[21]。设 N_1 是数据集中正确排歧的同名作者数，N_2 是排歧方法得到的同名作者数，N_c 是 N_2 中正确排歧的同名作者数，则 P，R 和 $F1$ 的公式为：

$$P=\frac{N_c}{N_2}\times 100\% \tag{5}$$

$$R=\frac{N_c}{N_1}\times 100\% \tag{6}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{7}$$

$F1$ 越大，对应的查准率和查全率越大，排歧效果越好。

3.4 实验结果

根据 3.1 节中的数据，GHOST 和 CoAAND 方法构造的图中包含的顶点数和边数如表 4 所列。由表 4 可知，由于 CoAAND 方法增加了作者的隶属机构信息，图的规模较 GHOST 方法大，因此搜索有效路径的时间较长。由于 CoAAND 方法区分合作关系类型和隶属关系类型的边，而 GHOST 方法只有合作关系类型的边，因此 CoAAND 的时间复杂度比 GHOST 方法的高，但实验发现，CoAAND 与 DISTINCT 方法的时间复杂度相当。

分别用 DISTINCT，GHOST 和 CoAAND 方法对 3.1 节中的数据进行排歧，运行 10 次计算 P，R 和 $F1$ 的均值和标准差，实验结果如表 5 所列，GHOST* 和 GHOST 分别为表 2 和表 3 数据集上的排歧结果。

表 4 待排歧同名作者对应的实体关系图信息

Table 4 E-R information of same name authors

方法	姓名	顶点数	边数
GHOST	李丹	881	2678
CoAAND		1108	4220
GHOST	李强	506	2188
CoAAND		679	3083
GHOST	李伟	1723	5274
CoAAND		2260	8142
GHOST	王鹏	317	798
CoAAND		378	1257
GHOST	王伟	1666	4155
CoAAND		2039	6916
GHOST	王艳	255	323
CoAAND		581	1046
GHOST	杨静	1026	1238
CoAAND		2677	4992
GHOST	张慧	547	734
CoAAND		2193	3480
GHOST	张静	440	577
CoAAND		1166	2110
GHOST	张伟	261	321
CoAAND		551	1035

表 5 不种方法的同名排歧效果

Table 5 Disambiguation results of different approaches

姓名	方法	P/%	R/%	$F1$ 值/%
李丹	DISTINCT	98.78±3.39	28.08±1.97	43.70±2.49
	GHOST*	45.21±4.44	48.66±1.25	46.77±2.59
	GHOST	46.62±4.20	48.93±0.87	47.66±2.35
	CoAAND	**100.00±0**	**92.41±0**	**96.06±0**
李强	DISTINCT	89.12±11.52	39.11±8.22	54.12±9.26
	GHOST*	42.40±4.62	81.16±5.33	55.59±4.67
	GHOST	59.66±3.99	95.66±1.41	73.42±3.21
	CoAAND	**100.00±0**	**100.00±0**	**100.00±0**
李伟	DISTINCT	**100.00±0**	34.30±2.82	51.02±3.14
	GHOST*	39.99±3.61	46.09±2.24	42.76±2.65
	GHOST	52.13±2.42	60.24±1.57	55.86±1.58
	CoAAND	**100.00±0**	**100.00±0**	**100.00±0**
王鹏	DISTINCT	94.55±9.54	49.00±9.22	64.13±9.19
	GHOST*	20.22±8.66	23.50±5.68	21.43±7.00
	GHOST	28.84±10.82	27.81±7.43	28.08±8.23
	CoAAND	**100.00±0**	**97.50±0**	**98.73±0**
王伟	DISTINCT	**92.98±2.78**	22.66±0.69	36.43±1.00
	GHOST*	42.57±2.13	53.88±1.43	47.53±1.60
	GHOST	47.38±2.45	58.94±1.23	52.50±1.78
	CoAAND	80.10±0	**84.93±0**	**82.45±0**
王艳	DISTINCT	96.80±10.12	54.77±13.57	69.25±12.30
	GHOST*	24.73±3.61	79.77±2.50	37.61±4.04
	GHOST	31.31±4.87	**84.75±3.62**	45.54±5.44
	CoAAND	**100.00±0**	79.55±0	**88.61±0**
杨静	DISTINCT	96.94±2.45	6.46±0.35	12.11±0.61
	GHOST*	93.03±0.25	97.77±0.10	95.34±0.14
	GHOST	93.52±0.20	98.35±0.05	95.88±0.10
	CoAAND	**100.00±0**	**100.00±0**	**100.00±0**
张慧	DISTINCT	95.43±9.75	46.11±7.26	62.04±8.28
	GHOST*	36.84±3.84	64.74±3.82	46.80±2.89
	GHOST	63.30±8.30	85.63±2.63	72.49±5.59
	CoAAND	**100.00±0**	**98.95±0**	**99.47±0**
张静	DISTINCT	**99.30±1.57**	39.48±4.16	56.40±4.24
	GHOST*	41.56±5.07	73.10±3.17	52.83±4.30
	GHOST	77.44±5.76	90.69±5.15	83.50±5.22
	CoAAND	96.43±0	**93.10±0**	**94.74±0**
张伟	DISTINCT	**100.00±0**	37.64±2.10	54.66±2.20
	GHOST*	73.49±3.70	48.63±1.18	58.49±1.37
	GHOST	80.00±4.50	51.34±0.36	62.50±1.41
	CoAAND	**100.00±0**	**100.00±0**	**100.00±0**

由表 5 可以看出，相比于 DISTINCT 和 GHOST，CoAAND 方法的 $F1$ 明显较大，查全率最高，查准率也较高。为了进一步说明 CoAAND 方法的有效性和优越性，表 6 列出了 10 个同名作者的平均查准率、查全率以及 $F1$ 值。

表 6　10 个数据集上 3 种方法的平均指标

Table 6　Three approaches' mean value on ten data sets

（单位：%）

方法	查准率 P	查全率 R	$F1$ 值
DISTINCT	99.13	39.06	53.78
GHOST*	44.78	60.44	49.55
GHOST	58.18	69.18	61.58
CoAAND	97.65	94.64	96.01

从表 6 可知，GHOST 方法的实验效果略好于 DISTINCT 方法，而两者都远差于 CoAAND 方法，其原因为：1）路径长度强制设定为 3，不能完全发挥合作作者之间的关系信息作用；2）所选数据只涵盖了待排歧同名作者文献记录及其合作作者的合作作者文献记录，信息量有限；3）实验数据信息只选择了待排歧作者、其直接合作作者以及作者的隶属机构信息。在前两点的限制下，GHOST 方法不能发挥其最优性能；因 DISTINCT 方法若要发挥其最优性能，需要期刊、出版地等更多的信息，但其只使用了合作作者及隶属机构信息，信息量较少，因此不能达到最优性能；但即便是在信息受限的情况下，本文所提的 CoAAND 方法依然具有较好的性能，这进一步说明了 CoAAND 方法的优越性。

根据实验可得 CoAAND 方法的最优 $F1$ 值以及 $F1>70\%$ 的阈值区间，如表 7 所列。

表 7　CoAAND 方法的最优阈值及阈值区间

Table 7　CoAAND's best thresholds and their intervals

作者	最优 $F1$ 值/%	最优 $F1$ 降 5% 后的阈值区间	$F1>70\%$ 的阈值区间
李丹	96.06	[0.60,0.85]	[0.15,0.95]
李强	100.00	[0.00,0.65]	[0.00,0.95]
李伟	100.00	[0.20,0.85]	[0.05,0.95]
王鹏	98.73	[0.00,0.90]	[0.00,0.90]
王伟	82.45	[0.20,0.75]	[0.15,0.95]
王艳	88.61	[0.15,0.55]	[0.15,0.95]
杨静	100.00	[0.05,0.95]	[0.00,0.95]
张慧	99.47	[0.05,0.80]	[0.00,0.95]
张静	94.74	[0.05,0.20]	[0.05,0.40]
张伟	100.00	[0.00,0.35]	[0.00,0.75]

从表 7 可以看出，大部分同名排歧作者所对应的较优阈值区间都包含[0.05,0.55]区间，因此，对未知的同名作者进行排歧，可以将阈值设置为[0.05,0.50]的区间内的某个值，其大小由用户根据其对查准率和查全率的要求而定。

结束语　本文提出的同名排歧方法 CoAAND 较好地解决了包括单一作者信息的同名排歧问题。该方法的具体优点如下：根据作者间的合作关系及作者与机构间的隶属关系构造的实体关系图，涵盖了全部关系信息；根据图中两顶点间的路径长度、数目、边的类型定义连接强度，较好地反映了两顶点间的相似程度；通过设定阈值，得到对应同一人的姓名对，进行闭包运算，高效地实现了同名排歧。

CoAAND 方法仍存在一定的局限性，如阈值的设定对于不同的同名作者进行排歧时，很难保证取得最优排歧效果；对单一作者的同名作者排歧时，若其署名机构名称粒度过粗，则不能有效实现其排歧。

参 考 文 献

[1] TAN M C，DIAO X C，CAO J J，et al. Relationship Type Based Connection Strength Model for Relationship-based Entity Resolution[J]. Journal of Computational Information Systems，2015，11(16)：5947-5957.

[2] ANDERSON A F，VELOSO A，MARCOS A G，et al. Self-training Author Name Disambiguation for Information Scarce Scenarios[J]. Journal of the American Society for Information Science & Technology，2014，65(6)：1257-1278.

[3] EMILIA A S，ANDERSON A F，MARCOS A G. Combining Classifiers and User Feedback for Disambiguating Author Names[C]// Proceedings of JCDL'16. Knoxville，Tennessee，USA，2015：259-260.

[4] COTA R G，ANDERSON A F，MARCOS A G，et al. An Unsupervised Heuristic-based Hierarchical Method for Name Disambiguation in Bibliographic Citations[J]. Journal of the American Society for Information Science & Technology，2010，61(9)：1853-1870.

[5] FAKHRI M，PHILIPP M. Using Co-authorship Networks for Author Name Disambiguation[C]// 2016 IEEE/ACM Joint Conference on Digital Libraries(JCDL). 2016：261-262.

[6] CARVALHO A P，ANDERSON A F，ALBERTO H F，et al. Incremental Unsupervised Name Disambiguation in Cleaned Digital Libraries[J]. Journal of Information and Data Management，2011，2(3)：289-304.

[7] FAN X M，WANG J Y，PU X，et al. On Graph-based Name Disambiguation[J]. ACM Journal of Data and Information Quality，2011，2(2)：1-23.

[8] MADIAN K，PUCKTADA T，LEE C G. Online Person Name Disambiguation with Constraints[C]// ACM/IEEE-CS Joint Conference on Digital Libraries. 2015：37-46.

[9] KIM K，KHABSA M，GILES C L. Inventor Name Disambiguation for a Patent Database Using a Random Forest and DBSCAN[C]// Proceedings of the 16th ACM/IEE-CS on Joint Conference on Digital Libraries. 2016：269-270.

[10] ZHENG C S，JI D，CAI D F. The Method of Expert Name Disambiguation Based on System Combination[J]. Journal of Shenyang Aerospace University，2014，31(2)：74-78.（in Chinese）
郑才松，季铎，蔡东风. 基于系统融合的专家同名区分方法[J]. 沈阳航空航天大学学报，2014，31(2)：74-78.

[11] CHEN W L. Name Disambiguation Based on the Coauthorship Association Graph of Scholar Papers[D]. Hangzhou：Hangzhou Dianzi University，2017.（in Chinese）
陈未路. 基于科研论文合作者关系图的同名排歧方法研究[D]. 杭州：杭州电子科技大学，2017.

[12] THIAGO A G，RICARD S T，ARIADNE M B，et al. A Relevancd Feedback Approach for the Author Name Disambiguation Problem[C]// Proceedings of ACM/IEEE Joint Conference on Digital Libraries'13 Indianapolis. Indiana，USA，2013：209-218.

（下转第 260 页）

[3] ZHANG Z,XU Y,YANG J,et al. A Survey of Sparse Representation:Algorithms and Applications[J]. IEEE Access,2017,3:490-530.

[4] LIU L,TRAN T D,SANG P C. Partial face recognition:A sparse representation-based approach[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2016:2389-2393.

[5] QIU D,LIU Y. Improved image super-resolution via sparse representation[J]. Video Engineering,2016,12(8):100-104.

[6] BOLÓN-CANEDO V,SÁNCHEZ-MAROÑO N,ALONSO-BETANZOS A. Feature selection for high-dimensional data[J]. Progress in Artificial Intelligence,2016,5(2):65-75.

[7] LIU J H,LIN M L,ZHANG J,et al. A kind of heuristic local random feature selection algorithm[J]. Computer Engineering and Applications,2016,52(2):170-174. (in Chinese)
刘景华,林梦雷,张佳,等. 一种启发式的局部随机特征选择算法[J]. 计算机工程与应用,2016,52(2):170-174.

[8] ZHANG Z,HANCOCK E R. A Graph-Based Approach to Feature Selection[C]// International Conference on Graph-Based Representations in Pattern Recognition. Springer-Verlag,2017:205-214.

[9] RIVEROMORENO C J,BRES S. Texture Feature Extraction and Indexing by Hermite Filters[C]// International Conference on Pattern Recognition. IEEE,2017:684-687.

[10] JIANG F,LI G H,YUE X. Semantic-based Feature Extraction Method for Document[J]. Computer Science,2016,43(2):254-2589. (in Chinese)
姜芳,李国和,岳翔. 基于语义的文档特征提取研究方法[J]. 计算机科学,2016,43(2):254-258.

[11] ZHOU G,CICHOCKI A,ZHANG Y,et al. Group Component Analysis for Multiblock Data:Common and Individual Feature Extraction[J]. IEEE Transations Neural Networks Learning Systems,2016,27(11):2426-2439.

[12] XIAO L Y,CHEN X H,LIN X L. Feature Weighted and Improved Partition Fuzzy C-Means Cluster Algorithm[J]. Microelectronics &Computer,2016,33(10):143-146. (in Chinese)
肖林云,陈秀宏,林喜兰. 特征加权和优化划分的模糊 C 均值聚类算法[J]. 微电子学与计算机,2016,33(10):143-146.

[13] ZHANG L,JIANG L,LI C,et al. Two feature weighting approaches for naive Bayes text classifiers[J]. Knowledge-Based Systems,2016,100(C):137-144.

[14] LUO Y,ZHAO S L,LI X C,et al. Text keyword extraction method based on word frequency statistics[J]. Journal of Computer Applications,2016,36(3):718-725. (in Chinese)
罗燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用,2016,36(3):718-725.

[15] CHEN Z,XIA J B,BAI J,et al. Feature Extraction Algorithm Based on Evolutionary Deep Learning[J]. Computer Science,2015,42(11):288-292. (in Chinese)
陈珍,夏靖波,柏骏,等. 基于进化深度学习的特征提取算法[J]. 计算机科学,2015,42(11):288-292.

[16] ZENG Q S,HUANG X Y. Fast Data Mining Algorithm Based on FP-tree[J]. Journal of Chongqing University of Technology (Natural Science),2009,23(10):72-76. (in Chinese)
曾庆森,黄贤英. 基于 FP-tree 的快速数据挖掘算法[J]. 重庆理工大学学报(自然科学),2009,23(10):72-76.

(上接第 225 页)

[13] YIN X X,HAN J W,PHILIP S Y. Object Distinction:Distinguishing Objects with Identical Names[C]// Proceedings of International Conference on Data Engineering(ICDE). 2007:1242-1246.

[14] XU R F,GUI L,LU Q,et al. Incorporating Multi-kernel function and Internet Verification for Chinese Person Name Disambiguation[J]. Frontiers of Computer Science,2016,10(6):1-13.

[15] HIEN T N,TRU H C. Named Entity Disambiguation:A Hybird Statistical and Rule-Based Incremental Approach[C]// Proceedings of the Semantic Web:the 3th Asian Semantic Web Conference(ASWC). 2008:420-433.

[16] FU J L,QIU J,GUO Y L,et al. Entity Linking and Name Disambiguation Using SVM in CHINESE Micro-blogs[C]// Proceedings of International Conference on Natural Computation. IEEE,2016:468-472.

[17] LI Y P. Bibliometric Analysis and Name Disambiguation Research Based on Knowledge Clustering[D]. Nanjing:Nanjing University of Posts and Telecommunications,2016. (in Chinese)
李永萍. 基于知识聚类的文献统计与重名消歧机制的研究[D]. 南京:南京邮电大学,2016.

[18] MIN S,ERIN H K,HA J K. Exploring author name disambiguation on PubMed-scale[J]. Journal of Informetrics,2015,9(4):924-941.

[19] MU L M. Research of the Nature & Operation of Finite Multi-Set[J]. Journal of Neijiang Normal University,2009,24(4):5-8. (in Chinese)
牟廉明. 有限多重集的运算及性质[J]. 内江师范学院学报,2009,24(4):5-8.

[20] TRAVERS J,MILGRAM S. An Experimental Study of the Small World Problem[J]. Sociometry,1969,32(4):425-443.

[21] MONGLI L,TOK W L,WAI L L. Intelliclean:A Knowledge-Based Intelligent Data Cleaner[C]// ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. 2000:290-294.