面向科学文献同名消歧的可视化分析方法

张鹏宇, 张勇*, 崔言杰, 尹宝才

(北京工业大学信息学部北京人工智能研究院 北京 100124)
(zhangyong2010@bjut.edu.cn)

摘 要: 直观地展示科研团队中的合作关系, 能够辅助用户在科学文献管理工作中更好地进行文献同名消歧. 由于科学文献作者合作关系复杂, 难以通过可视化进行直观展示, 因此设计并实现了面向科学文献的同名消歧可视化分析方法. 首先, 根据文献合著者存在的合作网络生成合作关系图, 揭示科研团队中作者的合作关系; 其次, 为了展示不同作者研究方向之间的相关性, 设计了合作关系图和发文期刊图之间的可视化联动; 最后, 通过结合深度学习模型分别对文献和作者进行分类, 实现了从作者和团队任意主体出发的交叉分析与连贯推理. 基于北京工业大学 4 000 篇论文构成的真实数据进行案例分析, 并邀请了科研管理人员和学生通过实验完成度和李克特量表进行评价, 验证了所提方法的有效性.

关键词: 可视化分析; 网络分析; 同名消歧; 多视图
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2022.19191

Visual Analysis for Name Disambiguation of Academic Papers

Zhang Pengyu, Zhang Yong*, Cui Yanjie, and Yin Baocai

(*Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology, Beijing* 100124)

Abstract: The visual presentation of academic social networks can help users better perform the name disambiguation work in academic paper management. The complex network among authors brings a great challenge to visual interaction. In order to help users better perform the name disambiguation work, a name disambiguation visualization analysis method for academic papers is designed and implemented. Proposed method first generates a cooperative relationship graph based on the cooperative network of co-authors, which is used to reveal the cooperative relationship of authors in the scientific research team. Then, to show the correlation between the research directions of different authors, the visual linkage between the collaboration graph and the published journal graph is designed. Finally, through the combination of the deep learning model, papers and authors are classified respectively to achieve the cross-analysis and coherent reasoning starting from the author or team. The system is based on 4 000 actual papers for case studies and professionals and students are invited to use and evaluate the system, proving the effectiveness in solving name disambiguation.

Key words: visual analytics; network analytics; name disambiguation; multi-view

收稿日期: 2021-06-21; 修回日期: 2021-07-20. **基金项目**: 国家自然科学基金(62072015, U19B2039); 中国高等教育学会专项课题(2020XXHYB16). **张鹏宇**(1992—), 男, 硕士研究生, 主要研究方向为大数据分析及可视化、人工智能; **张勇**(1979—), 男, 博士, 副教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为大数据分析及可视化、人工智能、图形学及虚拟现实仿真; **崔言杰**(1995—), 男, 硕士研究生, 主要研究方向为大数据分析及可视化、人工智能; **尹宝才**(1963—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为多媒体技术、跨媒体智能、视频编码.

当不同的人使用同一个名字时, 就会产生歧义, 这是对学术论文进行记录和统计时常会遇到的问题. 这种问题在中文环境中, 尤其是 2 个字的中文名字中极其常见[1]. 因为很多中文字虽然拥有不同的含义, 不同的写法, 但是却有同样的发音, 此时单一的姓名信息难以起到区分不同作者的作用. 例如, “王伟”和“王韦”的发音完全一样, 这 2 位作者在发表英文论文时, 他们名字的写法同为 Wang Wei.

确保可以正确、高效地区分有歧义的姓名, 是学术检索系统[2]以及各大高校科研管理部门面临的重要问题. 对于线上的学术检索系统, 如 ACM, DBLP 和 IEEE, 由于算法的局限性, 面临着论文作者姓名模糊以及论文归属错误的问题, 给科研人员在进行信息检索时带来了巨大困难. 对于各大高校的科研管理部门, 因为学术论文是反映高校整体学术水平和科研实力的重要指标, 所以要定期对本校教师发表的学术论文进行汇总统计. 而面对大量由校内教师发表论文而产生的同名问题, 科研管理部门需要在短时间内给出准确度较高的消歧结果, 这给相关人员带来了巨大困难. 如何将论文快速、准确地分配到对应作者名下, 是线上学术检索系统及线下科研管理部门的工作人员亟待解决的难题.

针对同名消歧问题, 已存在相关工作对同名作者进行区分[3]. 目前通常的做法是利用论文作者的姓名[4-5]、论文引用网络[6]等对网络上的公开数据集进行分类. 此类方法已经取得了一定的准确率, 但依然需要通过人工筛查才能最终完成同名消歧工作. 为了更好地进行人工筛查, 相关工作便将可视化方法引入人工消歧的工作中. 已有的可视化系统[1,7-8]将同名消歧算法与可视化方法结合, 对学术论文进行消歧. 但相关工作的可视化方法较为专业, 使用者在具备学术背景的同时需经过一定时间的培训, 这为论文消歧增加了成本.

为了解决上述问题, 本文面向科研管理人员对论文作者的同名消歧工作, 尝试将可视化分析方法与学术论文的消歧过程相结合, 提出了面向科学文献同名消歧的可视化分析方法, 以直观的可视化方法展示复杂的数据, 以帮助科研管理人员更高效、准确地完成基于学术论文的同名消歧工作.

1 相关工作

本文的工作涉及论文作者同名消歧方法、论文作者同名消歧可视化 2 个研究方向, 下面对国内外相关工作进展进行阐述.

(1) 论文作者同名消歧的方法

同名消歧的核心是将数据库中拥有相同名字的不同作者区分开来. 文献[9]在 1969 年便对这个问题进行了探讨. 根据文献[10], 在 20 世纪末, 姓名消歧常用方法为使用人工对数据进行消歧. 文献[11]提出利用论文属性自动识别作者身份. 在此基础上, 文献[12-13]将论文的名称与论文的其他属性进行融合, 以进一步提高识别的准确度. 这种做法启发了大量研究人员对论文属性融合进行更深层次的探索[14]. 在此基础上, 文献[15-17]利用作者简历与论文进行融合, 进一步提高识别的准确度.

已有的同名消歧工作已经取得了较好的效果, 但依然需要辅以大量烦琐的人工筛查工作进行论文消歧. 而本文的工作是在利用多种节点关系图的基础上, 将同名消歧算法与可视化方法结合起来, 以保证未受过专业训练的使用者可以直观地理解并利用算法分类的结果进行同名消歧.

(2) 论文作者同名消歧的可视化

将可视化方法与论文作者消歧结合的工作相对较少. 文献[8]提出通过利用论文期刊及会议信息进行聚类, 并通过聚类结果进行消歧, 同时结合可视化进行表示. 文献[18-19]将数据匹配算法与可视化方法结合, 并突出展示了作者在社交网络中的关系, 使作者之间的关系易于在界面中识别. 文献[1]提出了名为 NameClarifier 的新型视觉分析系统, 打破了传统姓名消歧的黑箱, 直观地解释了节点分类理由以及消除歧义的过程. 以上方法均取得了良好的效果, 但在验证用户操作的正确性上有所欠缺, 使用者若要验证消歧操作是否正确, 仍要付出大量努力.

在以上工作的基础上, 本文面向科研绩效管理评估工作, 以合作关系为主并辅以作者研究方向等信息设计可视化分析方法, 帮助使用者更好地发现论文及作者之间的联系, 同时使用者可以借助不同模块之间的联系直观地验证消歧操作是否正确.

2 任务分析与系统设计

为了更好地了解科研人员以及高校科研管理人员在论文匹配和同名消歧方面的需求, 本文分别对 2 位科研管理工作人员进行了多次调研. 截

至调研时, 2 位工作人员均在北京工业大学从事科研管理工作 10 年以上. 通过对调研内容的总结, 论文匹配与同名消歧工作普遍存在以下难题.

T1. 如何从海量的论文库中快速地定位到需要进行同名消歧的论文作者. 现有做法是聘请有经验的工作人员在论文库中搜索指定作者, 但搜索结果只有文字显示, 这种做法十分耗时.

T2. 如何区分同名作者. 当找出包含特定名字作者的所有论文后, 下一步要确定这部分论文是否存在被工作人员或算法错误分配的问题. 现有做法是首先根据论文名称判断论文方向, 其次分别查询同名作者每个人的研究方向, 最终确定待消歧论文属于哪位作者. 这种做法对查询人员有较高的学科分类背景知识要求.

T3. 如何利用多维度信息辅助使用者做出决策. 论文数据集中包含的数据维度较多, 包括论文发表刊物的级别、年份、论文作者的研究方向及合著者的研究方向等内容. 以上内容如果仅以文字形式表达, 很难发现同一团队中不同作者研究方向的共性.

针对上述任务, 本文提出了以下可视化功能, 用来指导完成论文匹配和同名消歧的过程.

S1. 对数据库中的数据进行查询. 使用者可以对论文作者进行搜索或在由程序生成的待消歧论文作者中进行逐条检索, 目的是选出需要进行同名消歧的作者, 作为后续任务的输入.

S2. 对论文作者的合作关系网络进行探索. 将上一任务选出的作者及其合作作者关系通过网络结构的形式表示. 首先将所有与被选中作者有合作关系的作者以节点形式呈现在图中, 然后根据论文作者之间是否存在合作发表论文的情况确定节点之间是否有连线. 通过合作网络可以初步了解相同发文团队中各个作者的合作关系. 接下来通过关联程度图确定每个节点和团队中其他节点的关联程度, 关联程度低的作者可能为潜在的错误节点, 也是使用者需要重点调查的对象. 并且, 将合作关系图与发文期刊图进行组合设计, 通过不同图之间的联动交叉比对, 使用者可以更好地探索团队中作者的发文相关性.

S3. 引入基于多视图分类的方法对论文和论文作者进行分类, 以达到分析论文作者研究方向的目的. 通过作者发表论文的方向、发表论文的关键词和发表论文的期刊等信息对论文作者进行分类, 并将分类结果通过层次化突出显示, 从而确定个体与团队的研究方向.

S4. 通过使用者手动修改信息和添加强联系动态更新数据集. 由于团队中作者之间存在多种合作关系, 不同强度的合作关系在可视化界面中会有不同体现. 当使用者把握了某些联系之后, 可以手动添加作者之间的强联系或修改错误联系. 在修改错误联系后, 可视化方法提供基础信息模块供使用者进行验证, 确保修改信息正确.

明确要实现的可视化功能后, 本文考虑拥有相同姓名的不同作者通常不属于同一科研团队, 而不同科研团队之间在研究方向、发文习惯、发文期刊和发文时间上存在较大差异, 依据此特点, 使用以下准则作为使用者的判断指导.

G1. 同一团队中不同作者通常在较集中的几个期刊上发文, 相同团队中不同作者发表期刊重合率高.

G2. 同一团队中不同作者通常研究方向相似.

G3. 同一团队中不同作者合作关系通常比较固定, 不同作者之间根据合作次数可分为强合作关系和弱合作关系.

G4. 同一团队中不同作者之间通常会相互合作, 很少会出现团队中某位作者只与整个团队中一人合作过的情况.

G5. 作者的研究方向通常不会发生较大变化.

G6. 同一团队中不同作者的发文时间通常较为相似.

3　本文系统总体框架

针对上述任务, 本文设计的可视化系统分别由数据处理和数据可视化 2 部分组成. 可视化方法如图 1 所示.

数据处理部分中, 需要对论文原始数据进行数据清洗与降维, 并利用深度学习模型进行论文与论文作者的分类, 便于进一步结合可视化部分探索作者之间关系.

数据可视化部分包含 4 个可视化探索模块, 分别为查询模块、关联程度模块、合作关系网络模块和基础信息模块.

(1) 查询模块. 使用者在查询模块中搜索到某作者的姓名后, 所有包含被搜索作者的论文便会组成待消歧论文集, 待消歧论文集中出现的全部作者则组成了可能包含错误节点的合作团队, 此团队的相关信息会展示在可视化界面中.

(2) 关联程度模块. 使用者使用关联程度模块对搜索结果进行初步判断, 与被搜索作者团队关

联程度低同时发文数量高的节点为可疑节点.

(3) 合作关系网络模块. 使用者可以随时通过合作关系模块与基础信息模块中信息的联动进行辅助判断. 在此基础上, 使用者还可以增加节点之间的“强联系”. 对节点的操作或增加强联系的操作会影响待消歧论文集. 每次操作和交互, 系统便会动态地重新构建关联关系图. 这种交替更新的过程有助于更准确地显示节点关系, 从而帮助使用者更好地消歧.

(4) 基础信息模块. 当出现对算法分类结果怀疑或对消歧结果需要验证的情况时, 使用者可以通过基础信息模块进行确认, 基础信息模块包含了未经处理的原始论文数据, 通过原始可靠的信息辅助使用者进行消歧.

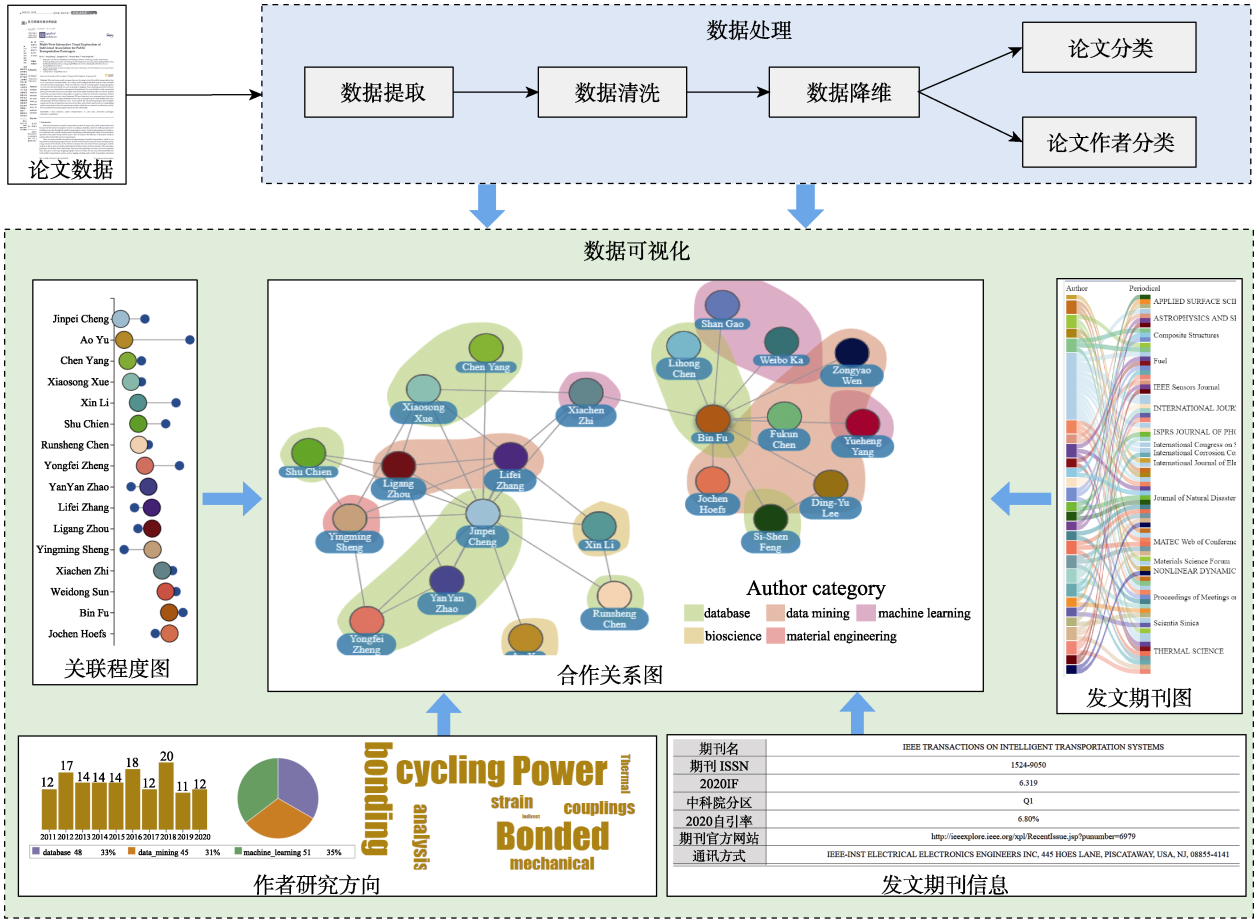


图 1 面向科研文献同名消歧的可视化分析方法

4 数据预处理

本文可视化方法使用的数据包括论文的发表时间、论文发表的期刊或会议、论文包含的关键词、论文中的合作作者, 这些特征均来自论文数据. 通过分析这些数据, 可以更好地对其进行分类.

本文使用的数据来源于 DBLP 公开论文数据集和北京工业大学非公开论文数据集. 为了提高数据质量, 以便更好地对论文进行分类, 将删除所用数据集中的某些噪声数据、重复记录的冗余数据以及记录不完整的缺失数据.

4.1 DBLP 公开论文数据集

DBLP 是计算机领域内对研究的成果以作者为核心的计算机类英文文献的集成数据集系统. 数据集中按年代列出了作者的科研成果, 包括国际期刊和会议等公开发表的论文. DBLP 中以论文作者为节点, 以 2 位作者之间的合作关系构建边. 如果 2 位作者曾经共同发表某篇论文, 那么他们之间便存在联系. 作者被分为 4 个领域: 数据库、数据挖掘、机器学习和信息检索. 使用公开数据集可以验证本文方法在其中同名消歧的有效性.

基于 DBLP 数据集, 本文构建了多视图的邻接矩阵和节点的特征信息. 其中, 包括 4 056 位论文作者节点, 节点关系边 500 万条, 包含了 3 种关系图, 即共同发表论文(2 位作者共同发表过某篇论文)、共同出席会议(2 位作者曾在同一个会议或期

刊中发表过论文)和共同使用关键词(2 位作者曾共同使用某个关键词发表过论文). 作者特征为作者发表过的论文关键字组成的词袋. 数据集以作者的研究领域为标签, 验证分类结果是否准确.

4.2　北京工业大学非公开论文数据集

北京工业大学非公开论文数据集来自北京工业大学图书馆, 包含了 2011—2020 年北京工业大学教师发表的论文. 数据集以论文名称为中心, 包含大量信息作为论文的属性. 此数据集的特征如下.

(1) 每位作者都有校内唯一工号与之对应, 每位教师在进入北京工业大学工作时会被分配唯一工号, 而每位教师只有唯一所属学院.

(2) 由于系统在研发过程中得到了北京工业大学各学院科研助理的配合, 所以在使用北京工业大学数据集验证时便于核查, 以验证高校在对科研绩效考核时本文方法的有效性.

(3) 本数据集在数据预处理时, 已使用大量人工利用传统消歧手段进行消歧, 解决了诸如将 Wang Wei 写为 Wang W、将 Beijing University of Technology 写为 BJUT 等写法不规范情况, 并得到各学院科研助理确认, 保证了数据集的质量.

本文基于北京工业大学非公开论文数据集, 构建了多视图邻接矩阵和节点特征信息, 包括 4 000 篇论文节点, 节点关系 400 万条; 并以教师所属学院作为节点的标签, 验证分类结果是否准确.

5　基于多视图卷积网络的论文与作者分类

考虑需要消歧的作者虽然拥有相同的姓名, 但通常有不同的研究方向, 因此, 本文利用拥有相同姓名的作者的研究方向不同作为可视化系统的消歧切入点, 分别对论文以及论文作者进行分类.

5.1　论文分类

利用论文的关键词和发表期刊等信息对论文进行分类, 并将论文分类结果展示在作者发文方向中(如图 1 中的作者研究方向图), 以此辅助使用者确定论文作者的研究方向.

5.2　论文作者分类

利用作者使用过的发文关键词作为特征对作者进行分类, 并将分类结果展示在合作关系图中(如图 1 中的合作关系图), 以此辅助使用者在可视化界面中快速直观地区分不同研究方向的作者以及作者之间的关系.

5.3　分类算法的具体实现过程

基于以上思想, 本文引入了基于图卷积网络的分类模型[20], 分别对论文和论文作者进行分类. 在分类过程中, 需要 2 部分数据作为算法的输入, 首先需要节点的特征向量, 由论文关键词映射到低维空间得出. 其次需要节点之间的关系图. 此前的做法大多用单视图进行分类, 然而, 在真实世界中, 不同节点之间的关系相对复杂, 很难用单视图表示节点之间的关系, 所以本文利用多视图进行论文的分类. 多视图与单视图相比, 提供了更多的节点信息. 本文通过 3 组关系图进行分类, 从而更全面地描述节点间的关系, 得出更准确的分类结果. 由于数据处理部分和数据可视化部分相对独立, 后续如有其他算法的分类效果好于现有算法, 本文系统支持将现有算法替换为效果更好的分类算法.

下面以对论文作者节点进行分类为例说明分类模型实现方法. 首先, 根据论文作者之间的关系以邻接矩阵的形式构造节点关系图. 本文以论文的合著关系构建第 1 个视图, 以论文作者共同发文期刊关系构建第 2 个视图, 以论文作者共现同一关键词关系构建第 3 个视图. 其中, 关系图为 $G=(V,E)$, 在对论文作者进行分类时, 将作者看作节点 V, 节点之间的边为作者之间的关系 E, 节点的特征为 $\boldsymbol{X}$. 在邻接矩阵中, $A_{ij}=1$ 表示节点 i 和节点 j 之间有一条边, $A_{ij}=0$ 表示节点 i 和节点 j 之间没有边. 特别地, 节点本身的连接为 0, 即 $A_{ii}=0$, 不考虑节点与自身的连边.

其次, 为了充分捕捉特征空间中的结构信息, 本文使用之前提到的节点特征生成的 k 近邻图作为特征结构图, 并对特征结构图和 3 组关系图分别作卷积. 对于输入图的第 l 层的输出 $\boldsymbol{Z}$ 可以表示为 $\boldsymbol{Z}^{(l)}=\mathrm{ReLU}\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{Z}^{(l-1)}\boldsymbol{W}^{(l)}\right)$. 其中, $\boldsymbol{W}^{(l)}$ 为 GCN 第 l 层的权重矩阵; 初始的 $\boldsymbol{Z}$ 即 $\boldsymbol{Z}^{(0)}$ 为节点特征矩阵; $\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$, $\boldsymbol{I}$ 为单位矩阵, $\boldsymbol{A}$ 为节点关系矩阵; $\tilde{\boldsymbol{D}}$ 为 $\tilde{\boldsymbol{A}}$ 的对角矩阵; ReLU 为激活函数.

4 个输入图的输出分别为 $\boldsymbol{Z}_1,\boldsymbol{Z}_2,\boldsymbol{Z}_3,\boldsymbol{Z}_4$. 接下来需要最大化 4 个输入图之间的不同, 以确保不同输入可以捕获不同的节点关系. 这里使用希尔伯特-施密特独立性准则(Hilbert-Schmidt independence criterion, HSIC)衡量 4 个输入图之间的分布差异, 从而尽可能地扩大不同输入之间的差距. 最终损失函数为

$$L=\mathrm{HSIC}(\boldsymbol{Z}_1,\boldsymbol{Z}_2)+\mathrm{HSIC}(\boldsymbol{Z}_1,\boldsymbol{Z}_3)+\mathrm{HSIC}(\boldsymbol{Z}_1,\boldsymbol{Z}_4)+\mathrm{HSIC}(\boldsymbol{Z}_2,\boldsymbol{Z}_3)+\mathrm{HSIC}(\boldsymbol{Z}_2,\boldsymbol{Z}_4)+\mathrm{HSIC}(\boldsymbol{Z}_3,\boldsymbol{Z}_4).$$

其中，$\mathrm{HSIC}(\boldsymbol{Z}_1,\boldsymbol{Z}_2)=(n-1)^{-2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{K}_1\boldsymbol{R}\boldsymbol{K}_2)$. n 为节点个数；$\boldsymbol{K}_1$ 和 $\boldsymbol{K}_2$ 为格拉姆矩阵，即由 2 个向量经过内积运算所组成的矩阵；$\boldsymbol{R}=\boldsymbol{I}-\frac{1}{n}\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}$，$\boldsymbol{I}$ 为单位矩阵，$\boldsymbol{e}$ 为一个全 1 的列向量.

5.4 多视图卷积网络效果评估

本文使用 ACM 和 DBLP 公开论文数据集对模型分类的准确性进行评估. ACM 数据集以论文为节点，以论文之间是否存在共同作者、共同参会或共现关键词确定节点之间是否有边. 节点分为 3 类: 数据库、无线通信和数据挖掘. 模型利用 2 025 个节点和 500 万条节点关系作为输入，1 000 个节点作为测试节点，预测的测试节点类别作为模型输出. DBLP 数据集以作者为节点，以作者之间是否存在合著、共同参会或共用关键词确定节点之间是否存在边. 节点分为 4 类: 数据库、数据挖掘、机器学习和信息检索. 利用 3 056 个节点和 700 万条节点关系进行训练，1 000 个节点作为测试节点，预测的测试节点类别作为模型输出. 评估方法为模型分类的准确率与 F_1 分数, 2 种评估指标均为越高越好. 具体结果如表 1 所示.

表 1 分类模型评估结果 %

数据集	准确率	F_1 分数
ACM	91.42	91.36
DBLP	91.72	91.78

由表 1 可知，分类模型的准确率较高，可以有效地帮助使用者快速确定合作关系图中节点的研究方向. 当分类算法将节点错误分类时，使用者依然可以通过关联程度模块确定每个单独节点与团队整体的关联程度寻找可疑节点. 在确定可疑节点后，使用者便可以通过基础信息模块中的“作者信息”标签页了解团队中每个节点未经任何算法处理过的原始信息，包括发文关键词、论文标题与摘要等，以此判断每个节点的研究方向. 这样可以保证在分类算法失效时，使用者依然可以通过本系统完成论文消歧.

6 可视化系统设计

如图 2 所示，学术论文同名消歧可视化分析系统由 4 部分协同交互的模块组成，如图 2 所示. 可视化模块分别为查询模块、关联程度模块、合作关系模块和基础信息模块. 接下来给出每个模块的具体介绍并描述各模块的详细设计.

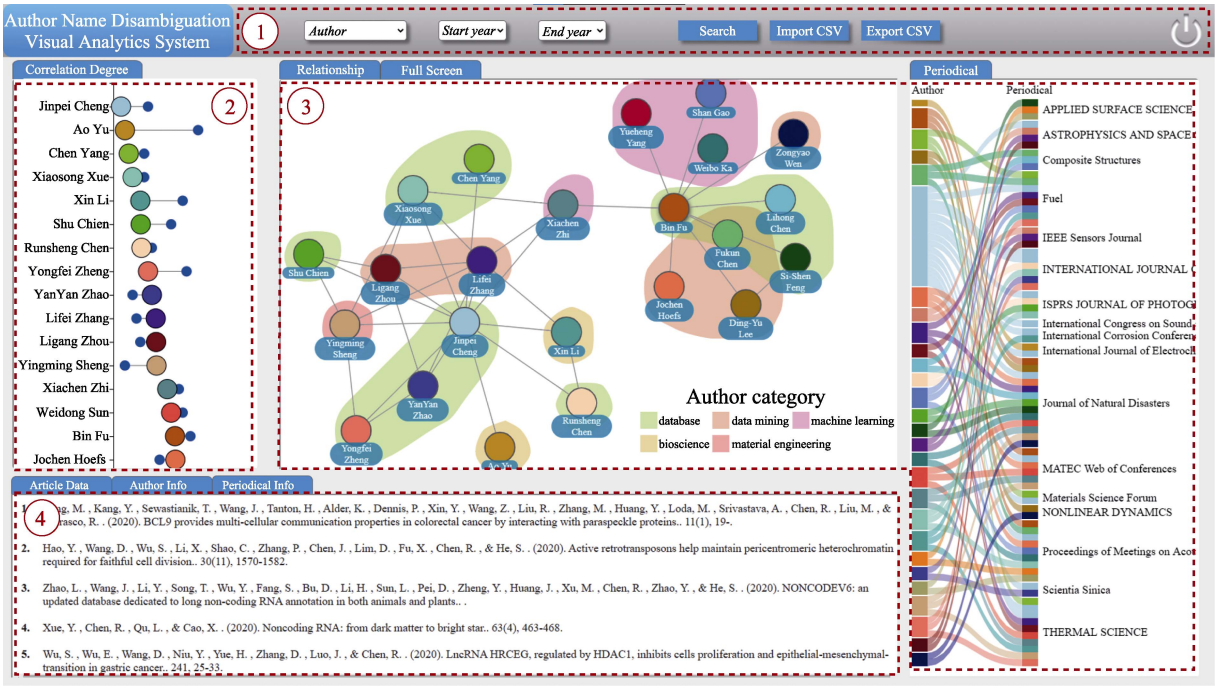


图 2 面向科研文献同名消歧的可视化系统

6.1 查询模块

查询模块是整个可视化程序的入口和出口. 从将数据以 CSV 格式的文件导入，查询某个特定姓名的论文作者，到最终将完成修改的数据导出

为 CSV 格式的文件, 都要使用此模块操作. 查询模块如图 2 中虚线框①所示. 具体操作过程如下.

Step1. 获取数据. 使用者可以在知网、DBLP 或任何公开学术检索工具中将论文数据导出为 CSV 格式的文件, 并作为本系统的输入.

Step2. 导入 CSV 文件. 使用者在进行搜索和消歧之前, 需要按照系统指定的格式导入 CSV 文件. 单击 Import CSV 按钮, 并选择准备好的 CSV 文件上传即可. 上传文件应包含论文标题、发文年份、发文期刊、发文关键词和论文作者等信息.

Step3. 搜索论文作者. 此部分包含了需要查询的 2 部分内容: 以论文作者作为节点的查询和时间范围的查询. 在 Author 下拉列表中选择论文作者的姓名后, 在 Start year 和 End year 下拉列表中选择希望查询的时间. 例如, 在 Author 下拉列表中选择查询 Wang Wei 的论文, 选择查询论文的时间为 2011—2020 年, 选择完成后单击 Search 按钮即可. 在确定查询作者和查询时间后, 系统会在已上传数据集中搜索被查询的作者姓名, 包含被查询姓名的所有论文都会被从上传的总数据集中检索出来, 被检索的论文形成一个小型的待消歧数据集. 对于论文作者进行同名消歧和其他模块中的可视化渲染都依据这个待消歧数据集.

Step4. 导出 CSV 文件. 在使用者消歧结束后, 系统会自动保存消歧结果. 此时单击 Export CSV 按钮, 并选择导出文件的存储位置, 即可完成 CSV 格式文件的导出.

6.2　关联程度模块

可视化系统将待消歧数据集中所有论文作者看作同一团队, 此团队中的人数是不确定的, 规模可能从数人到数十人. 因为使用者对团队中的每位成员均详细调查会耗费大量时间和精力, 所以系统引入哑铃图帮助使用者快速确定团队中的“可疑节点”, 即最有可能被算法错误分配到团队中的论文作者, 使用者仅需要对可疑节点或与可疑节点相连的节点重点关注即可.

在关联程度模块中, 使用哑铃图可以直观地看到每位作者与团队中其他作者之间的关联程度和此作者的发文数量, 分别用大、小 2 个圆表示. 哑铃图中纵坐标代表不同作者, 每位作者都被赋予了一种独有的颜色, 同一位作者在不同模块中均使用同一种颜色, 以提高相同作者在不同模块之间的识别度, 保证使用者可以更快速、准确地区分不同作者.

大圆的颜色为作者独有的颜色, 大圆横坐标代表此作者与整个团队的关联程度得分. 关联程度得分情况由每位作者和团队中其他作者的合作发文数、发文方向重合度、连接中心性等数值综合得出. 如图 2 中虚线框②所示, 关联程度得分低, 则大圆位置偏左, 代表此作者与团队中其他作者关联程度更低, 更“不合群”; 关联程度得分高, 则大圆位置偏右, 代表此作者与团队中其他作者关联程度更高, 更“合群”.

然而, 仅使用关联程度得分这项指标会产生将刚加入科研团队的年轻教师误判为可疑节点的问题, 因为刚加入某个科研团队的年轻教师, 由于与团队中其他成员合作发文数量较少, 故如果仅考虑关联程度得分这项指标, 年轻教师并不属于此科研团队. 因此, 本文在关联程度模块中引人入了每位作者的发文数量得分作为对比, 在图中以小圆表示. 为了便于使用者对比, 小圆的颜色统一指定为蓝色, 小圆的横坐标代表了此作者发文数量得分. 作者发文数量少, 则得分少, 小圆偏左; 作者发文数量多, 则小圆偏右. 例如, 图 2 中虚线框②最上面的教师 Jinpei Cheng, 从大圆位置可看出此节点与团队关联程度得分最低, 但是从小圆位置可以看出此人发文数量得分较低, 可能是新加入团队的年轻教师. 而排名第 2 位的教师 Ao Yu, 与团队关联程度得分较低, 同时发文数量得分较高, 由此可以判断 Ao Yu 节点更有可能是被错误分配到团队中的错误节点, 需要使用者重点关注.

6.3　合作关系模块

与已有的工作[1]不同, 本文系统通过展示团队成员之间的合作关系以及团队成员的研究方向辅助使用者进行论文消歧. 网络关系图将要观察的作者抽象为点, 将 2 个作者之间的合作关系抽象为边, 这就构成了基本的网络关系图. 如果仅使用基础的网络关系图, 会显得杂乱, 当节点数量较多时很难看出节点之间的不同特征. 因此, 本文系统在网络关系图的基础上, 使用力导向布局. 力导向布局是网络关系图中的一种, 它除了点与点之间的联系外, 还使用了空间中节点的聚集程度这一指标, 可以反映不同团队、不同节点分类之间的关系. 力导向布局模拟了节点之间的引力和弹力, 即当 2 个节点距离过近时会彼此弹开, 而 2 个节点距离过远时会彼此吸引, 避免了当图中包含较多节点时导致的不同节点之间相互覆盖而产生的视觉混乱问题.

充分了解团队中论文作者的研究方向对论文

消歧有很大帮助, 如果要确定论文作者的研究方向, 重要的参考标准之一便是作者投稿的期刊与会议. 某位作者的投稿期刊和会议在某种程度上代表了此人的研究方向, 也代表了团队的研究方向. 例如, 某团队中大部分作者都在化学类期刊发表过论文, 而团队中只有一人在历史学相关的期刊发表过论文, 那么就可以大致确定此人被错误划分到这个团队中, 也为接下来的论文消歧提供了指引. 于是, 此处需要一种可以清晰明确展示不同作者发文期刊的图. 桑基图由边、流量和支点组成, 其中边代表了流动的数据, 流量代表了流动的具体数值, 边的宽度与流量成比例显示, 边越宽, 数值越大. 在发文期刊图中, 系统使用桑基图展示团队中作者的发文期刊情况.

当使用者使用系统时, 会默认进入合作关系模块中的普通模式, 此时可以看到模块中包含了左侧的网络关系图和右侧的发文期刊图, 如图 2 中虚线框③所示.

(1) 网络关系图与发文期刊图

在图 2 虚线框③中, 左侧的网络关系图中的每个节点代表一位作者, 作者节点的颜色与之前提到过的关联程度模块相同. 如果 2 名作者曾经合作发表过同一篇论文, 2 个节点之间便会产生连线. 此处需要说明, 因为本文系统使用颜色作为区分不同节点的重要手段, 故为了保证团队中每个节点都有独一无二的颜色, 同时当团队中节点较多时, 也能通过颜色对不同节点产生较好的区分度, 本文首先将颜色利用 RGB 色彩模式(RGB color mode, RGB)分为 3 个数值, 然后利用欧几里得距离度量 2 个 RGB 颜色值的区分度, 最后通过随机或贪心算法依次获得指定数量的颜色值.

如果分类算法将多位作者判断为同一研究方向, 系统会用同一颜色的色块将同一研究方向的节点包裹起来, 达到更直观的效果. 在图 2 虚线框③中, 29 位作者被算法分为 5 个不同的研究方向, 系统用 5 种不同的颜色将研究方向相同的作者包裹起来, 直观地描绘出基于多视图分类的结果与不同作者的合作关系.

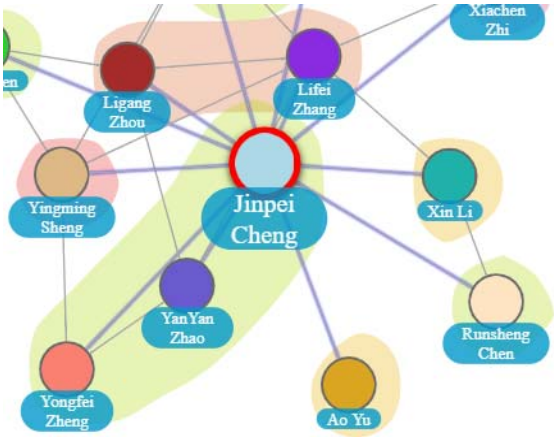
右侧的发文期刊图, 采用桑基图的呈现方式展示论文作者的发文期刊. 展示的信息分为 2 列, 左列为团队中包含的所有论文作者的姓名, 且作者姓名颜色与关系图中同一作者颜色一致, 便于使用者直观地了解作者发文情况; 右列为论文作者的发文期刊. 如果作者曾在某个期刊发表论文, 那么左列作者名和右列期刊名之间就会产生连线. 为了避免使用者混淆, 每名论文作者都使用之前提到的方法被赋予独一无二的颜色, 而每个期刊并没有被赋予颜色, 期刊颜色由在此期刊发表过论文的作者所决定. 例如, 红、蓝 2 名作者总计在期刊 A 中发表论文 10 篇, 其中红色作者发表论文 9 篇, 蓝色作者发表论文一篇, 那么期刊 A 在发文期刊图中的面积便由 9/10 的红色和 1/10 的蓝色组成.

(2) 关系图与期刊图的交叉分析

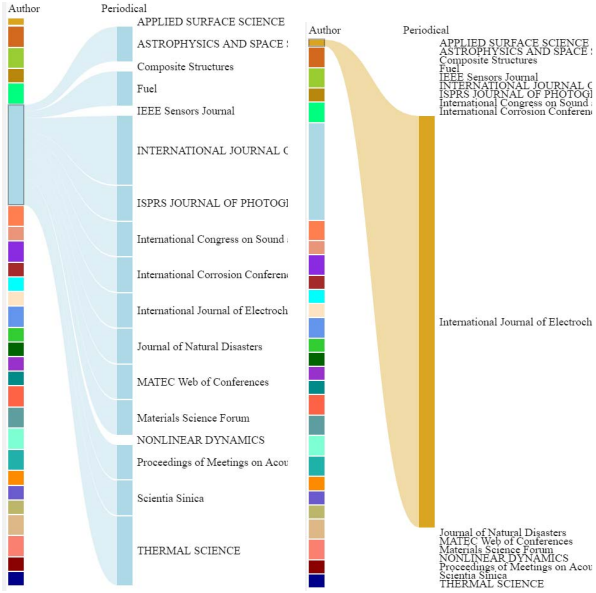
当使用者点击合作关系图中的 Jinpei Cheng 节点, 节点会被高亮显示, 如图 3a 所示. 与此同时, 发文期刊图中会自动隐去其他作者所发表的期刊, 只显示被点击作者发表期刊情况, 如图 3b 所示. 此时图中显示的是正确情况, 即被点击作者属于此团队的情况. 而当使用者点击被错误划分为此团队的作者时, 会看到可疑节点与其他节点不同, 仅在一个期刊上发表过论文, 如图 3c 所示. 此作者可能并不属于该团队, 只是由于人工或算法对论文分配错误导致. 如要验证猜想, 可将鼠标移动到可疑作者发表期刊上. 由于同一团队中作者通常都会在一个或几个期刊中发表论文, 如果看到同一期刊中同时有多位团队中作者发表过论文, 则代表此时数据分配正确, 如图 3d 中的 Fuel 期刊. 错误情况如图 3e 所示, 此时可以看到发表在此期刊的作者只有 2 人, 不符合 G1.

(3) 关联关系的修改

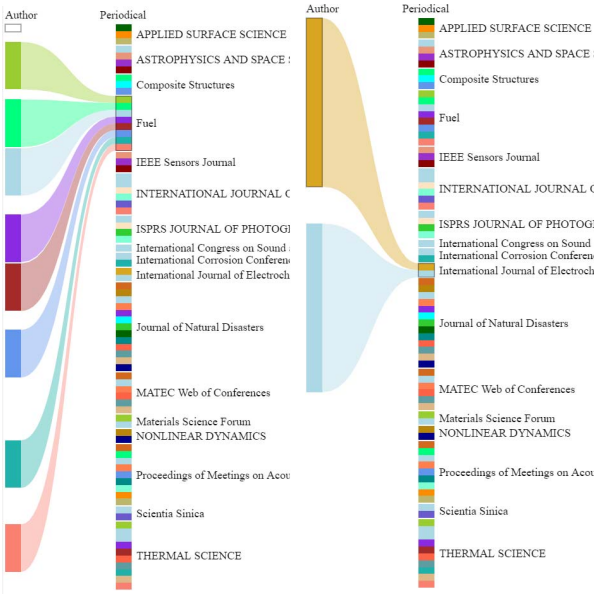
如果使用者需要进一步了解团队中作者之间的关系, 可以使用合作关系图的全屏模式. 全屏模式分为左侧的关联论文、中间的关系图和右侧的强联系 3 部分, 整体情况如图 4a 所示. 当使用者确定 2 个节点之间一定有合作关系时, 便可点击 2 个节点之间的线段, 被点击的线段会高亮强调(图 4b), 同时右侧强联系框中会出现 2 人已添加强联系的显示(图 4c). 被添加强联系的作者会被认定为一定有合作关系, 此结果会被反馈到分类算法中, 用来提高算法准确度. 使用者还可以同时点击 2 个节点(图 4d), 在 2 个节点高亮显示的同时, 左侧也会显示出被点击的作者因为哪些论文而产生的联系(图 4e). 使用者可以通过此功能判断节点之间的联系是否正确. 关联论文和强联系使用示例如图 4f 所示. 当使用者确定可疑节点为错误节点时, 通过在错误节点上右击, 在弹出的快捷菜单中选择添加或删除节点, 便可对错误数据进行修改.



a. 关系图中作者高亮显示

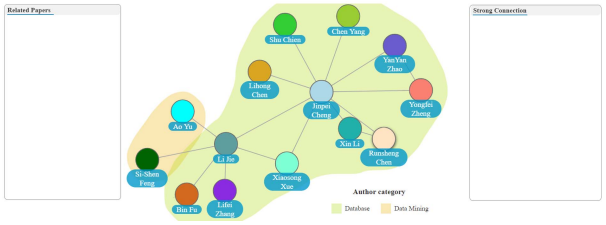


b. 正确作者发文期刊　　c. 错误作者发文期刊

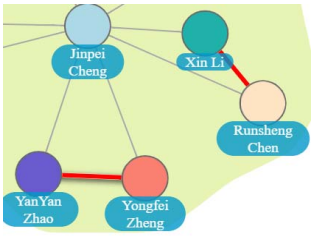


d. 正确期刊对应作者　　e. 错误期刊对应作者

图 3　关系图与期刊图的交叉分析

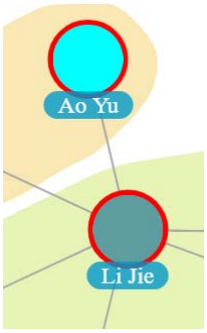


a. 合作关系模块全屏模式

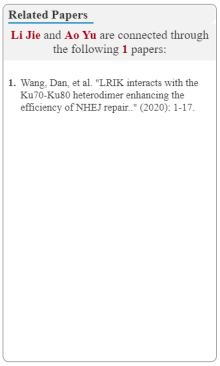


b. 节点间线段高亮

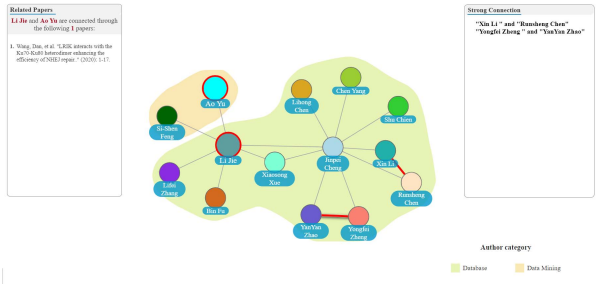
c. 添加强联系



d. 同时选中 2 名作者



e. 查看作者间关联论文



f. 合作关系模块全屏模式

图 4　关联关系的修改

6.4　基础信息模块

基础信息模块可以帮助使用者在使用关联程度和合作关系模块时，了解作者或期刊的详细信息. 此模块包含了论文原始数据标签页、作者信息标签页和期刊信息标签页.

(1) 论文数据标签页

系统会默认进入论文数据标签页，此标签页中包含了待消歧数据集中所有的论文数据，如图 5

所示. 可以看到论文名、发文年份、发文期刊、发文作者和发文关键词的信息. 论文数据标签页为使用者提供了最原始的论文数据供参考.

(2) 作者信息标签页

当使用者在合作关系图中点击某个节点时, 基础信息模块会自动显示被点击节点的相关信息, 包括作者的发文数量、发文年份、发文关键词和发文方向. 其中发文方向是通过分类算法对作者发表的论文进行分类得到的结果, 如图 6 所示. 作者信息标签页可以帮助使用者了解不同作者的研究方向, 以便更好地进行同名消歧.

(3) 期刊信息标签页

期刊信息标签页中包含了待消歧数据集中发文期刊的详细信息, 包含期刊名、影响因子和期刊方向等信息, 如图 7 所示. 使用者可以通过点击发文期刊图中的右侧期刊名进行切换. 期刊信息标签页可以帮助缺乏经验的使用者快速了解不同期刊的研究方向, 从而更准确地定位可疑节点.

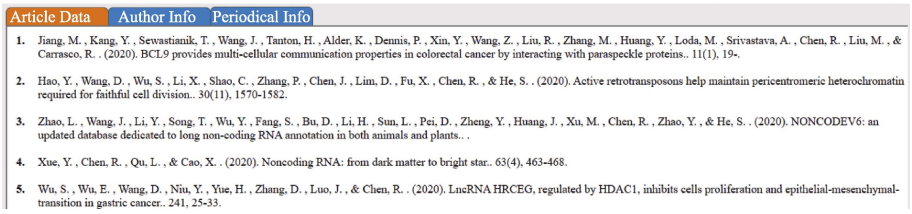


图 5 论文数据标签页

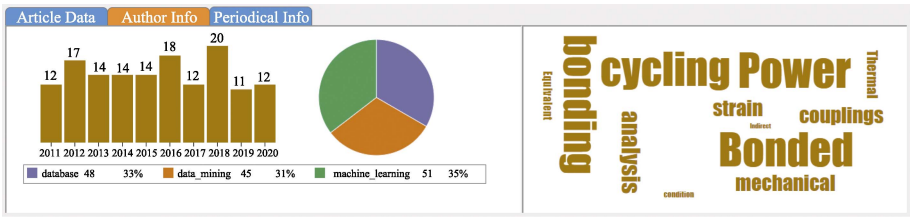
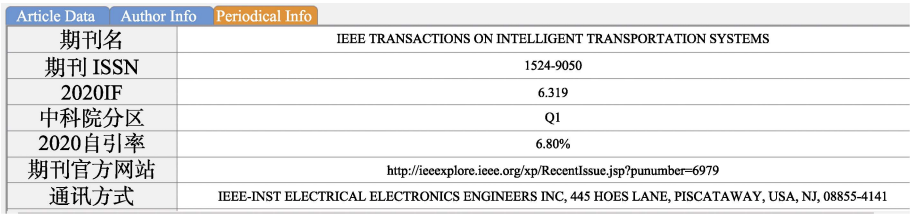


图 6 作者信息标签页



期刊名	IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS
期刊 ISSN	1524-9050
2020IF	6.319
中科院分区	Q1
2020自引率	6.80%
期刊官方网站	http://ieeexplore.ieee.org/xp/RecentIssue.jsp?punumber=6979
通讯方式	IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC, 445 HOES LANE, PISCATAWAY, USA, NJ, 08855-4141

图 7 期刊信息标签页

7 案例分析

下面通过 2 个案例说明本文方法在分析团队合作关系方面的优势. 与已有的偏向展现数据集中宏观的同名情况的可视化方法相比, 本文方法更着重展现微观的合作团队内的作者关系. 通过合作关系模块, 使用者可以直观地了解团队内的发文情况以及作者间彼此的合作情况. 2 个案例分别代表 2 种常见错误. 第 1 个案例用 Wang Wei 演示当多个发文团队中出现同名作者而导致需要消歧的情况; 第 2 个案例用 Li Jie 展示不同团队中出现学生与教师同名而需要进行同名消歧的情况. 为了便于验证同名消歧结果是否正确, 案例均来自北京工业大学非公开论文数据集.

7.1 案例 1——不同研究方向教师同名

经过对北京工业大学科研管理人员的采访, 发现北京工业大学中有多名教师姓名为 Wang Wei, 这给科研管理人员在论文与作者匹配时带来极大困难.

根据表述, 首先在查询模块中搜索作者姓名 Wang Wei, 时间范围为 2011—2020 年. 单击 Search 按钮后, 如图 8 所示, 可以通过合作关系图看到此团队分为 3 部分, 且每部分的研究方向均不相关. 这是同名消歧中较为典型的情况, 即多个不同研究方向的团队由于团队中均存在某个同名作者而被混在一起. 此时可以依据算法对论文作者的分类结果, 以及合作关系模块右下角图例初步推断待消歧论文集中包含了 3 个研究方向不相关的发

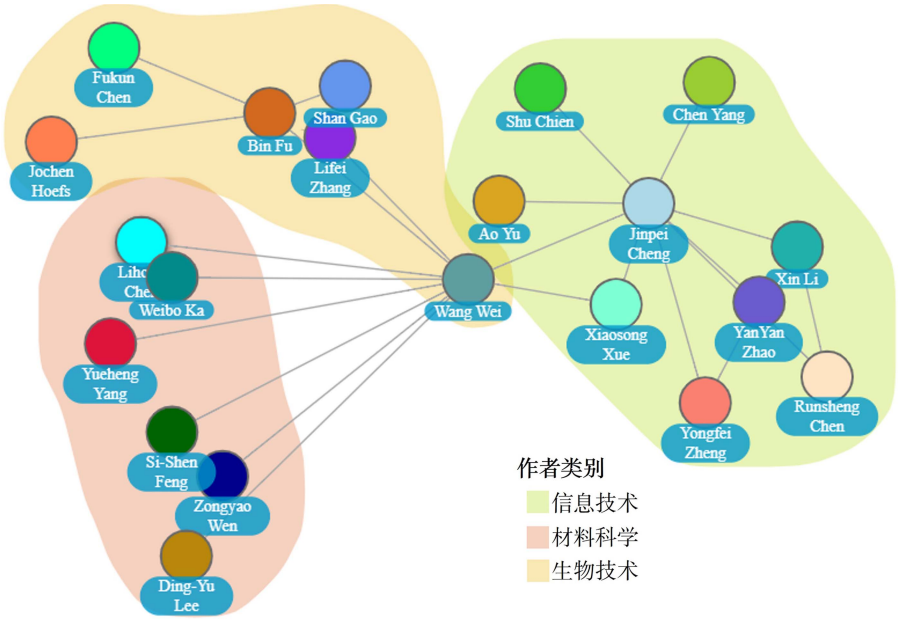


图 8　合作关系图

文团队，而每个团队中均有一位作者的姓名为 Wang Wei. 此时，目的为厘清 3 位 Wang Wei 的发文团队以及在不同团队中分别添加 Wang Wei 01, Wang Wei 02, Wang Wei 03 不同节点，以便日后科研人员统计或检索.

接下来，分别对合作关系图中每个发文团队进行研究. 可以看到，浅黄色色块包裹的节点的研究方向均为生物，浅粉色色块包裹的节点的研究方向均为材料，绿色色块包裹的节点的研究方向为信息技术. 逐个点击每个团队中的节点，并观察发文期刊图和基础信息模块中的内容. 可以发现，不同团队的发文期刊相互独立，相交节点均为 Wang Wei. 同时，不同团队中的节点的发文关键词均有很大的独立性. 点击发文期刊图右列期刊名，逐个了解每个期刊的研究领域后，可以确定 3 个团队的研究方向分别为生物、材料和信息技术. 这验证了之前的猜想.

最后，需要添加节点，即把合作关系图中心的 Wang Wei 节点分为 3 个节点，分别对应每个发文团队，如图 9 所示. 右击 Wang Wei 节点，选择添加节点，并分别标记 Wang Wei 01 为生物方向, Wang Wei 02 为材料方向, Wang Wei 03 为信息技术方向. 当成功区分节点后，如果次年有信息技术方向的论文加入数据库中，系统则会自动分配到 Wang Wei 01 节点下，避免每年重复对相同的姓名进行消歧. 在与相关学院科研助理确认后，得知消歧结果正确.

7.2　案例 2——相似研究方向教师与学生同名

经过对北京工业大学科研管理人员的采访，获悉北京工业大学信息学部 Li Jie 老师反映，自己团队中教师的论文并没有正确分配.

首先在查询模块中搜索作者姓名 Li Jie，时间范围为 2011—2020 年. 单击 Search 按钮后，可以通过合作关系图(图 4a)看到所有节点均属于同一团队，并没有明显的边界. 且根据作者研究方向进行分类结果同样较为相似，并不能通过色块代表的研究方向区分不同作者.

其次，观察关联程度模块(图 2 中虚线框②)，发现教师 Ao Yu 与团队中其他作者关联程度较低，同时发文数量较多，可以猜测教师 Ao Yu 被工作人员或算法分配到了错误的团队. 随后逐个点击节点，并观察作者信息标签页以及发文期刊图. 在作者信息标签页中发现团队中其他教师的发文关键词均与计算机相关，而 Ao Yu 的发文关键词则属于自动化类. 与此同时，在发文期刊图中，观察到除 Ao Yu 以外的教师的发文期刊均与团队中其他教师的发文期刊有重合且发文期刊丰富(图 3b)；只有教师 Ao Yu 的发文期刊单一(图 3c)，并且 Ao Yu 所发期刊只与教师 Li Jie 有交集(图 3e)，这不符合 G4. 同时注意到，如图 10 所示, Ao Yu 发文数量很多，却集中在 2015 年之前，不像团队中其他教师是逐年递增趋势，这不符合 G6.

接下来，进入全屏模式，并先后点击 Ao Yu 与 Li Jie 节点(图 4d)，发现虽然 2 人发文数量很多，但 Ao Yu 与 Li Jie 却只合作过一篇论文(图 4e). 由此可以推断，此论文是 Ao Yu 与学生合作，而学生名叫 Li Jie，与教师 Li Jie 重名.

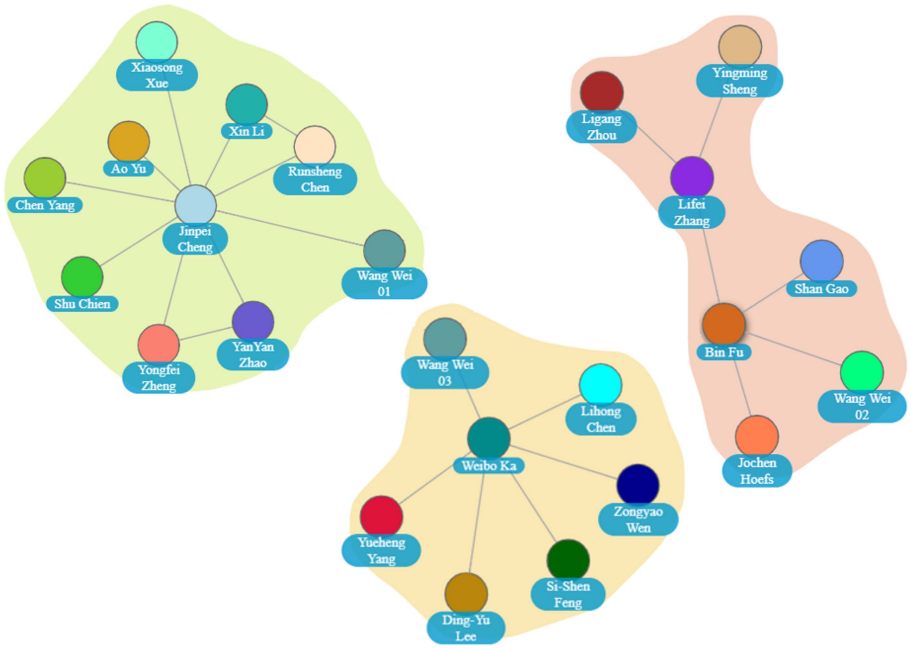


图 9 案例 1 消歧后合作关系图

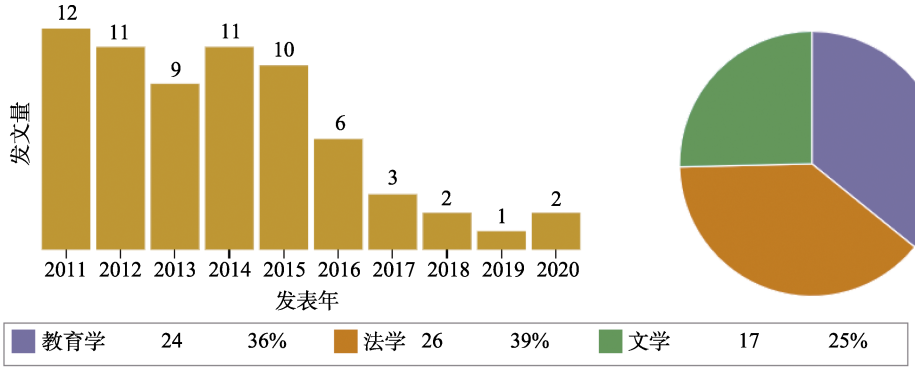


图 10 Ao Yu 发文分析

最后，在 Li Jie 节点上右击，标记计算机学院教师为 Li Jie 01，自动化学院学生为 Li Jie 02，如图 11 所示. 在询问相关学院科研助理后，得知消歧结果正确.

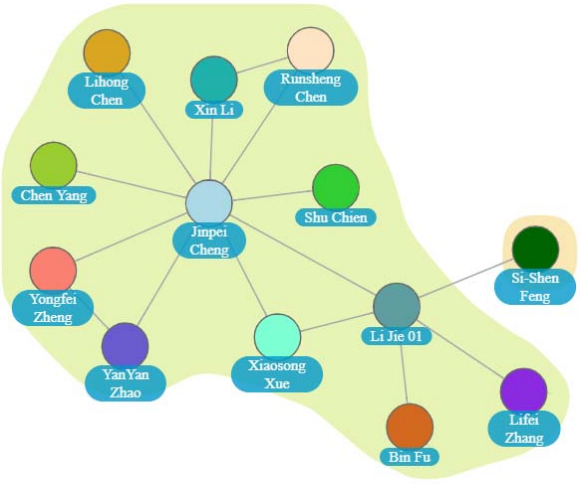


图 11 案例 2 消歧后合作关系图

8 系统评估

本文采用北京工业大学实际论文数据集验证上述可视化探索工具的可用性. 所用数据均来自北京工业大学 2011—2020 年实际发表的论文数据，分析系统完成下列任务的能力：区分多个不同研究领域中出现教师同名的情况；区分在论文分配错误导致相似领域中教师与学生同名的情况.

8.1 志愿者组成

本文招募 9 名北京工业大学的研究生和 2 名来自北京工业大学科研管理部门工作 10 年以上的工作人员作为志愿者，评估可视化系统的可行性. 在 9 名研究生中，包括女性 4 人，男性 5 人. 参与实验的人员均从未使用过本系统.

8.2 实验内容

考虑上文的案例总体操作顺序相似，但是每个案例侧重点又有不同，所以实验部分以案例 1 作

为实验 1, 它提供了完整的系统操作流程, 志愿者要求在没有帮助与提示的情况下, 完成实验 1 中的任务.

实验 2 为简化并分解之后的案例 2, 通过文字描述指导参与者完成实验. 具体步骤如下(Q1~Q3). 实验 2 在实验 1 完成一周之后才进行, 这样做保证了志愿者忘记了实验 1 中大部分细节.

Q1. 搜索教师 Li Jie, 并在关联程度模块中通过每个节点与其他节点的关联程度确定可疑节点.

Q2. 通过 Q1 中确定的可疑节点, 在合作关系模块中, 完成以下子任务.

Q2.1. 通过点击节点并观察作者信息标签页的方法, 了解教师 Li Jie 及团队中其他教师的研究方向.

Q2.2. 通过点击节点并观察作者信息标签页的方法, 找到可疑节点与团队中其他节点发文年份的不同.

Q2.3. 通过进入全屏模式并点击 Li Jie 和可疑节点的方法, 确定不同教师因为哪篇论文而产生的联系.

Q3. 判断出错误节点, 并通过在 Li Jie 节点上添加节点完成同名消歧的过程.

8.3　实验步骤

在实验开始前, 首先由一名开发人员向参与实验的志愿者介绍系统的各模块, 并且介绍各模块之间的联系. 其次, 开发人员为每名志愿者均分配一台计算机, 志愿者可以随意操作可视化系统 20 min, 以便熟悉整个操作流程. 接下来, 志愿者根据上文提到的实验步骤进行操作. 如果一个任务完成, 且经过开发人员确认, 那么就在实验结果后标记为 1, 否则标记为 0.

每次实验结束后, 志愿者立即对系统的直觉性、操控性和可用性进行评估, 评估标准利用李克特量表进行评定. 量表范围从非常不同意(+1)到非常同意(+5)进行统计. 量表中包含 20 个题目, 从“我认为我愿意经常使用这个系统”, 到“我需要学习很多东西才能开始使用这个系统”. 每位志愿者的整体可用性量表得分为 0~100.

8.4　结果分析

志愿者均以 95%以上的准确度完成实验, 因此在有工作经验的科研管理人员和学生之间对系统的可用性进行评分时没有产生明显的差异. 这些结果表明, 无论志愿者有什么样的工作经验, 或被给予什么样的实验指导, 可视化系统都可以提供良好的可用性和有效性, 以帮助完成现实世界中的同名消歧任务. 然而, 更明确的实验指导会使志愿者的可用性得分更高. 例如, 志愿者在实验 2 中的平均可用性得分为 87.67±3.95, 高于实验 1 中的 83.67±4.16.

从实验结果还可以看出, 实验指导的详细程度和志愿者的经验共同影响着志愿者对培训的需求. 例如, 在实验指导较少的实验 1 中, 出现了明显较高的指导需求. 这一点可能与志愿者多数为学生有关, 因为学生拥有的先验经验较少, 无法快速从实际的案例中提炼出清晰的提示.

此外, 不同性别的志愿者在不同实验中的不同评价项目上都没有出现统计学上的显著差异, 因为本实验预先为每位志愿者明确了目标以及提供了热身练习, 保证志愿者能够全神贯注于对系统的探索, 不会产生性别上的认知差异.

8.5　专家评估

开发人员向专家介绍了已有的相关工作[1]是如何进行论文消歧工作的, 并记录了专家对可视化系统的点评. 专家们表示, 本文可视化系统具有良好的可用性和有效性, 可以辅助解决现实世界中存在的同名消歧任务. 专家对系统的便利性、良好的整合性和交互性都给予了较高的评价, 并表示与前人的工作对比, 本文系统在面临高度模糊的消歧任务中表现虽不及前人的相关工作, 但能更直观地体现团队中作者之间的合作关系, 方法更易懂, 使用者需要较少的培训即可熟练使用. 同时, 由于大多数科研管理人员并非科研领域内专业人士, 而他们在进行论文消歧工作时经常会面临高校内不同领域的专业论文, 使缺少背景知识的人员很难进行消歧工作. 本文系统由于作者信息标签页和发文期刊图的加入, 使用者不需要对消歧领域有所了解, 也可以顺利地完成论文同名消歧任务.

9　结　语

本文以提高科研管理人员绩效考核时的效率和优化科研人员检索科研论文的体验为出发点, 首先总结了高校在进行绩效考核时考核人员的若干需求, 其次提出了同名消歧可视化分析方法, 并研发了同名消歧可视化分析系统. 该系统可以帮助科研管理人员高效地完成论文同名消歧工作.

当然, 本文可视化方法也存在一些不足之处. 一方面, 现有的可视化系统对于数据的利用不够全面, 仅包含论文的合作网络, 导致系统面对高度

模糊的消歧案例表现欠佳. 未来可以将可视化技术与更多数据挖掘和数据分析方法结合, 如论文参考文献构成的共被引网络等. 另一方面, 在使用关联程度模块时, 使用者虽然能够以半自动的方式锁定与团队相关程度较低的潜在错误节点, 但此模块占用的面积较大, 同时传达的信息密度较低. 未来可以考虑使用此模块表现更多可以帮助使用者作出决策的信息, 最终做到帮助使用者更快速地探索不同作者研究方向之间的相关性.

参考文献(References):

[1] Shen Q M, Wu T S, Yang H Y, *et al*. NameClarifier: a visual analytics system for author name disambiguation[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 141-150

[2] Ferreira A A, Gonçalves M A, Laender A H F. A brief survey of automatic methods for author name disambiguation[J]. ACM SIGMOD Record, 2012, 41(2): 15-26

[3] 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

[4] Wang X, Ji H Y, Shi C, *et al*. Heterogeneous graph attention network[C] //Proceedings of the World Wide Web Conference. New York: ACM Press, 2019: 2022-2032

[5] Wang X, Lu Y F, Shi C, *et al*. Dynamic heterogeneous information network embedding with meta-path based proximity[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1117-1132

[6] Bo D Y, Wang X, Shi C, *et al*. Structural deep clustering network[C] //Proceedings of the Web Conference 2020. New York: ACM Press, 2020: 1400-1410

[7] Zhu J, Zhou X F, Fung G P C. A term-based driven clustering approach for name disambiguation[M] //Advances in Data and Web Management. Heidelberg: Springer, 2009: 320-331

[8] Strotmann A, Zhao D Z, Bubela T. Author name disambiguation for collaboration network analysis and visualization[J]. Proceedings of the American Society for Information Science and Technology, 2009, 46(1): 1-20

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

[10] Myer S. Maxwell's guide to authority work[J]. Serials Review, 2003, 29(2): 162-163

[11] Smalheiser N R, Torvik V I. Author name disambiguation[J]. Annual Review of Information Science and Technology, 2009, 43: 1-43

[12] Tang L, Walsh J P. Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps[J]. Scientometrics, 2010, 84(3): 763-784

[13] Milojević S. Accuracy of simple, initials-based methods for author name disambiguation[J]. Journal of Informetrics, 2013, 7(4): 767-773

[14] Kang I S, Na S H, Lee S, *et al*. On co-authorship for author disambiguation[J]. Information Processing & Management, 2009, 45(1): 84-97

[15] Kanani P H, McCallum A, Pal C. Improving author coreference by resource-bounded information gathering from the web[C] //Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 2007: 429-434

[16] Pereira D A, Ribeiro-Neto B, Ziviani N, *et al*. Using web information for author name disambiguation[C] //Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. New York: ACM Press, 2009: 49-58

[17] Yang K H, Peng H T, Jiang J Y, *et al*. Author name disambiguation for citations using topic and web correlation[C] //Proceedings of International Conference on Theory and Practice of Digital Libraries. Heidelberg: Springer, 2008: 185-196

[18] Bilgic M, Licamele L, Getoor L, *et al*. D-dupe: an interactive tool for entity resolution in social networks[C] //Proceedings of the IEEE Symposium on Visual Analytics Science and Technology. Los Alamitos: IEEE Computer Society Press, 2006: 43-50

[19] Wang Yang, Yu Minzhu, Shan Guihua, *et al*. Visual comparison analysis of the competitiveness of scientific research groups[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(4): 542-550(in Chinese)
(王杨, 余敏楮, 单桂华, 等. 科研团队竞争力可视对比分析方法[J]. 计算机辅助设计与图形学学报, 2020, 32(4): 542-550)

[20] Wang X, Zhu M Q, Bo D Y, *et al*. AM-GCN: adaptive multi-channel graph convolutional networks[C] //Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2020: 1243-1253