计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(8):48−57 [doi: 10.15888/j.cnki.csa.007597]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

学者学术状态及竞争力可视化系统①

王　杨, 余敏楮, 单桂华, 陆忠华

(中国科学院 计算机网络信息中心, 北京 100190)
(中国科学院大学, 北京 100049)
通讯作者: 单桂华, E-mail: sgh@sccas.cn

摘　要: 随着每年出版的大量出版物, 学术数据迅速增长. 通过已发表的论文数据来准确、全面地呈现一个学者的科研水平和核心竞争力, 为大型科研机构的管理者、决策者或投资者提供辅助决策, 已成为文献大数据可视化的研究热点. 本文基于 Web Of Science (WOS) 论文数据, (1) 采用结合算法和交互式可视化的方法提升数据质量, 针对 WOS 论文数据特征, 设计实体分组算法和分组可视化校正工具, 实现了人名和单位名的消歧; (2) 根据常用的学术竞争力指标, 设计了学者画像可视化方法; (3) 研发了一套基于论文数据的学者画像可视化系统, 并通过具体的真实案例证明了该系统的实用性和有效性.

关键词: 学术状态; 学术竞争力; 可视化

引用格式: 王杨,余敏楮,单桂华,陆忠华.学者学术状态及竞争力可视化系统.计算机系统应用,2020,29(8):48–57. http://www.c-s-a.org.cn/1003-3254/7597.html

Visualization System of Academic Status and Competitiveness for Scholars

WANG Yang, YU Min-Zhu, SHAN Gui-Hua, LU Zhong-Hua

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
(University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: With a large number of publications published every year, academic data grows rapidly. Through published paper data to accurately and comprehensively present a scholar's scientific research level and core competitiveness so as to provide assistant decision-making for managers, decision-makers or investors of large-scale scientific research institutions, big data visualization has become a research hotspot of literature. This study based on Web Of Science (WOS) paper data, (1) in order to improve the quality of data, a combination of algorithms and interactive visualization is used to design entity grouping algorithm and grouping visualization correction tool for the data characteristics of WOS papers, which can eliminate the difference between person name and affiliation name; (2) according to the commonly used academic competitiveness index, the visualization method of scholar's profile is designed; (3) a set of visualization system of scholar's profile based on thesis data is developed, and the visualization system of scholar's profile based on publicated papers is developed. The real case of body proves the practicability and effectiveness of the system.

Key words: academic status; academic competitiveness; visualization

如今, 各种学术组织和学术机构遍布世界. 像中国科学院这种大型研究机构, 清华大学、浙江大学这样的大型高等学校, 中华人民共和国科学技术部、中国科协技术协会这样的大型科学组织, 人员众多, 内部结构复杂. 这些机构的管理者和决策者们希望全面掌握其研究人员的学术水平和质量. 投资者们也需要根据研

① 基金项目: 中国科学院“十三五”信息化专项课题 (XXH13504)
Foundation item: CAS Special Fund for Informatization Construction in 13th Five-Year Plan (XXH13504)
收稿时间: 2020-02-17; 修改时间: 2020-03-17; 采用时间: 2020-03-24; csa 在线出版时间: 2020-07-29

究人员的研究方向、学术竞争力、研究团队等多方面的信息来选择投资对象[1]. 一个研究人员所发表论文情况是描述其科研水平和核心竞争力的最重要的因素之一.

随着每年出版的大量出版物, 学术数据迅速增长. 到目前为止, 全球至少有出版物 2.3 亿份, 作者 2.3 亿人, 研究领域 71 万个, 会议 4.4 万个, 期刊 4.9 万种, 机构 2.6 万个. 这使得我们很难找出谁是某个研究领域中最有价值的专家, 或者谁是完成某项特定任务的最佳研究人员. 因此, 如何通过已发表的论文数据来准确、全面的呈现这些信息, 使用户快速定位合适的目标, 已成为文献可视化研究的热点.

常用的论文数据库有 Web Of Science (WOS)、Scopus、IEEE Xplore、谷歌学术、微软学术等. 数据库中的每条记录都有标题、作者姓名、作者单位、发表时间、发表的期刊/会议、摘要、参考文献、被引次数等. 然而, 记录中的作者姓名和作者单位的消歧面临着巨大的挑战. 一个作者姓名可能对应现实生活中多个作者; 一个现实生活中的作者在不同的论文中的署名方式可能不同. 一个单位的名称和地址在不同论文中的写法不同, 有使用全称的, 也有使用缩写甚至省略部分信息的, 并伴随一定比例的笔误. 这些数据问题会导致统计值不准确, 从而失去辅助决策的意义. 因此, 数据质量是数据驱动决策系统中极其关键的部分. 目前, 结合算法和交互式可视化的数据质量管理方法已成为可视分析领域的研究热点之一[2–4].

基于论文数据构建学者画像是近年来的研究热点问题. 学者画像的目标是提取研究人员各维度的属性信息进行信息挖掘和分析应用. 学者画像技术是大型智库实现专家发现、学术影响力评估等功能的关键. 如何选择反映研究人员学术竞争力的评估指标, 挖掘其科研团队, 并通过可视化为其全方位、高精度地构建画像, 是目前文献大数据分析需要解决的问题.

本文基于 WOS 数据, 包含了标题、作者姓名、作者单位、研究领域、参考文献、引文、会议或期刊、发表时间等丰富的信息, 设计了实体分组算法和分组可视化校正工具, 为分析提供尽可能准确的人名和机构名; 根据常用的学术竞争力指标, 设计了学者画像可视化方法; 基于合作关系挖掘学者的潜在研究团队. 最后, 研发了一套基于论文数据的学者学术状态及竞争力可视化系统, 并通过具体的真实案例证明了该系统的实用性和有效性.

1 相关工作

1.1 实体消歧

文献数据中作者姓名的歧义主要由两方面导致. 一是由于没有统一的署名标准, 同一作者在不同文献中署名不同; 二是不同作者的姓名可能相同. 人名消歧从 19 世界 60 年代开始就备受关注. 早期多采用人工消歧的方法. 但随着学者数量的快速增长, 人工消歧变得越来越不现实. 因此, 学者提出了大量先进的人名消歧算法模型来自动识别作者[5,6]. 目前大多数方法预先筛选出各种强特征, 计算它们的相似性, 以此来识别由同一个作者发表的论文. 例如, Milojevi´c[7] 验证了首字母在人名消歧中的有效性. 也有学者使用标题信息[8], 自引信息[9], 公共参考文献信息[10], 人名特征[7], 网页信息[11–13] 等增加人名消歧的准确性. 其中, 合作关系被证明是最易获取且最有效的特征[14]. 虽然, 随着特征的增加, 准确性可能会随之提高, 但只有很少一部分特征是普遍适用的. 有些挖掘算法模型需要的数据并不能轻松获取, 即使获取到了, 质量未必满足要求. 因此, Shen 等[15] 提出了一个新颖的可视分析系统, 用于交互式地对论文数据中的作者姓名进行消歧. 该系统量化了歧义姓名和确定姓名之间的相似度, 并将其可视化. 其相似度通过合作关系、发表论文期刊/会议、时间信息 3 个关键因素来计算. 该系统提供了可视化线索, 以帮助用户检查每一个有歧义的案例. 通过将用户引入消歧的过程中, 系统可以获得比采用挖掘算法模型更可靠的结果.

文献中单位实体的歧义主要是由于没有统一的书写标准以及笔误导致. 目前英文单位实体消歧相关的学术研究较少, 在工业界, 通常采用关键词规则匹配和上下文特征信息进行消歧.

本文涉及人名消歧和单位名消歧. 采用算法与交互式可视化相结合的方式来获得较高质量的数据. 人名消歧算法主要使用人名特征、合作关系、单位、研究领域等关键特征进行设计; 单位名消歧采用关键词规则匹配和基于莱文斯坦距离的相似度进行设计. 此外, 本文设计了分组可视化工具, 使专家可以实时查看经过算法分组后的人名和单位名, 并通过简单地拖拽实现对算法结果的校正.

1.2 学者画像的可视化

ACM Digital Library、DBLP、Google Scholar、

ResearchGate、Scopus、Semantic Scholar 等系统罗列了学者的基本信息,包括学者单位、论文列表、被引次数、合著者等. 但这些系统通常仅仅列出了上述信息, 或用少量基础图表展示上述信息, 用户需要通过链接访问不同的页面, 通过过滤器筛选感兴趣的数据, 才能获得需要的信息.

在学者画像的可视化方面, AMiner 从 h 指数、论文引用数、学者论文数等多个方面来评价学者的研究水平和研究状态, 为学者创建了包括学者简介、研究兴趣、论文列表、合著者网络、相关作者等信息的知识图谱. Latif[16] 分析了学者画像的可视化需求, 在对比了 ACM Digital Library、AMiner、DBLP、Google Scholar、ResearchGate、Scopus、Semantic Scholar 提供的列表式的学者画像的基础上, 提出了一种集成文本和视觉描述的信息呈现方法, 以突出论文数据中的模式. 他们利用基于模板的自然语言生成技术来总结显著的统计信息、研究主题的演变和合作关系; 无缝集成的可视化图表增强了文本描述的表现力, 并可以在图表间、文本间交互连接.

本文用直观的可视化图表从基本评价指标、研究兴趣、合著者和学术论文等方面呈现学者的研究状态和竞争力, 还通过交互式关系可视化方法呈现学者的论文、合著者、合作单位之间的关系, 并设计科研团队可视化方法, 帮助用户洞察学者的科研团队.

2 系统流程设计

学者学术状态及竞争力可视化系统 (图 1) 主要由 4 大模块组成: 关键词提取模块, 实体消歧模块, 团队挖掘模块和学者画像可视化模块. 关键词提取模块基于 *n*-gram 模型, 采用了自然语言处理技术, 从论文的标题、摘要、作者指定的关键词中提取名词性短语, 根据逆文档频率 (TF-IDF)、共现频次等特征对名词性短语进行排名. 最终取排名靠前的名词性短语作为论文的关键词. 实体消歧模块针对 WOS 数据质量问题, 对论文作者姓名和单位进行了特殊处理. 主要采用了结合基于算法的自动分组和基于交互式可视化的专家校验的方法. 团队挖掘模块主要负责从合著网络中发现专家学者稳定的合作团队. 学者画像可视化模块为用户提供了可交互的信息呈现和探索方法, 便于用户全方位地了解学者的学术状态和竞争力.

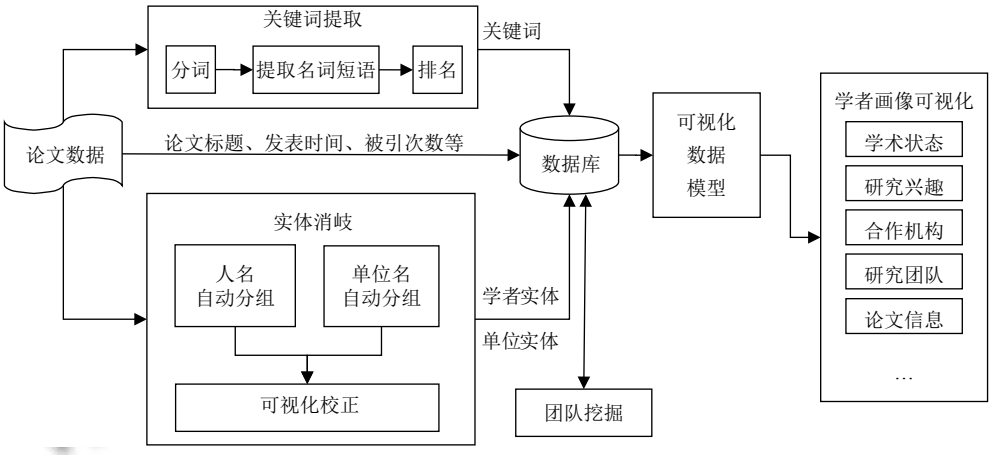

图 1 系统流程图

3 数据处理方法

3.1 单位名称处理

作者的单位名称从单位地址中提取得到. 通常, 同一个单位地址有不同的写法. 通过对数据集中单位地址的分析, 总结出单位地址的组成成分. 单位地址通常包含单位名、研究所名、学院名、部门名、实验室名、邮编、城市、省份、国家等元素. 通常有些单位地址会缺省其中部分元素, 如实验室名、邮编、省份、国家等信息 (本文简称可缺省元素), 但主要的单位名称、研究所名不会全部缺省 (本文简称主要元素). 比如表 1 中, 地址 0, 1, 2 实际上是同一单位, 但地址 0 省略了实验室名称, 地址 2 省略了邮编; 地址 8 和地址 6, 7, 9 单位名相同, 但所在城市不同, 不应该是同一个单位. 本文认为同一个单位应该具有相同的邮编, 位于相同的

城市、省份或国家. 此外, 数据集中存在一定的笔误, 如地址 10 和地址 11 仅仅差了一个英文字母“h”. 这应该是书写时遗漏的. 本文认为这两个地址指代同一单位. 类似的书写错误在数据集中非常普遍, 尤其当单位名中包含英文字母“n”或“m”时, 将“n”写成“m”或将“m”写成“n”的情况大大增加.

表 1 原始单位地址

编号	地址
0	Chinese Acad Sci, Inst Geochem, Guiyang 550002, Peoples R China.
1	Chinese Acad Sci, Inst Geochem, State Key Lab Environm Geochem, Guiyang 550002, Peoples R China.
2	Chinese Acad Sci, Inst Geochem, Guiyang, Peoples R China.
3	Chinese Acad Sci, Shanghai Inst Tech Phys, Shanghai 200083, Peoples R China.
4	Chinese Acad Sci, Shanghai Inst Tech Phys, Key Lab Space Act Optoelect Technol, Shanghai 200083, Peoples R China.
5	Chinese Acad Sci, Anhui Inst Opt & Fine Mech, Hefei 230031, Peoples R China.
6	Chinese Acad Sci, Guangzhou Inst Geochem, State Key Lab Orgran Geochem, Guangzhou 510640, Guangdong, Peoples R China.
7	Chinese Acad Sci, Guangzhou Inst Geochem, Guangzhou, Guangdong, Peoples R China.
8	Chinese Acad Sci, Guangzhou Inst Geochem, Beijing, Peoples R China.
9	Chinese Acad Sci, Guangzhou Inst Geochem, Key Lab Marginal Sea Geol, Guangzhou 510640, Peoples R China.
10	Chinese Acad Geol Sci, Inst Geomech, Beijing 100081, Peoples R China.
11	Chinese Acad Geol Sci, Inst Geomec, Beijing 100081, Peoples R China.

本文首先通过“inst”, “univ”, “corp”等关键词提取了单位名, 通过正则表达式提取邮编, 通过词典提取城市、省份、国家. 接着, 为了降低因笔误导致单位名不能精确匹配而带来的影响, 通过基于莱文斯坦距离的相似度来对单位名进行分组, 同一组的单位名很可能指代同一个单位. 然后通过邮编、城市、省份、国家来自动校正分组结果. 此外, 由于不同的机构在不同的时期名称可能不同, 两个相似度很高的单位名可能指代不同的单位. 比如: NSF 指代美国国家科学基金会, NSFC 指代中国国家科学基金会, 两者的差异非常小. 而由笔误引起的单位名差异也非常小, 如前面所举的例子. 这种情况下, 尤其是其它可缺省元素缺省比较严重时, 算法难免会把一些名字差异非常小但实际上指代不同一单位的字符串分成一组, 因此有必要进行人工验证及校正. 为了便于对分组结果进行验证和校正, 本文设计了分组结果可视化工具. 通过该工具, 可以调整阈值并查看对应阈值下分组的结果以寻找最合适的阈值. 同时, 可以直接通过该工具对不合理却无法通过算法得到正确分组的结果进行校正. 下面以表 1 中的地址为例说明单位名称处理的具体步骤:

步骤 1. 地址预处理. 在该步骤中, 邮编、城市、省份、国家首先被单独提取出来, 作为后续步骤的校正信息. 然后将地址字符串中的邮编删除, 避免由数字组成的邮编对单位名提取造成干扰. 因为存在若干名称包含数字的单位, 如“Univ Paris 06”. 此外, 地址字符串中的无用字符也在该步骤中被删除.

步骤 2. 单位地址分组. 在该步骤中, 通过“inst”, “univ”, “corp”等关键词提取地址字符串中的单位名, 作为该地址指代的单位名, 如地址 0 中的“Inst Geochem”, 地址 5 中的“Anhui Inst Opt & Fine Mech”. 然后将提取出来的单位名中的“&”和空格删除, 并将其统一转成小写. 接着我们用式 (1) 两两计算单位名的相似度, 得到一个相似度矩阵, 如表 2 所示. 根据相似度矩阵, 可以对单位地址进行分组. 依次遍历未分组的单位地址, 筛选相似度大于阈值 T 的单位名对应的地址, 形成一个临时组. 如表 2 中, 当遍历到地址 0 时, 地址 1, 2, 11 和地址 0 的相似度很高 (用红色框高亮), 他们形成一个临时组. 然而, 不难发现地址 11 与地址 10 的相似度比其与地址 0 的相似度更高 (用绿色框高亮). 地址 11 与地址 10 应当分为一组. 因此, 需要从临时组中剔除并不是与当前地址最相似的地址, 从而得到最终的分组. 继续遍历, 直到所有单位地址均有分组. 最终, 表 1 中的地址在阈值$T = 0.7$时的分组如下: [0, 1, 2,], [3, 4], [5], [6, 7, 8, 9], [10, 11].

$$Similarity(x,y) = 1 - \frac{Levenshtein(x,y)}{avg(\text{len}(x),\text{len}(y))} \quad (1)$$

步骤 3. 分组结果自动校正. 在该步骤中, 通过步骤 1 中提取的邮编、城市、省份、国家对分组结果进行纠正. 如步骤 2 中的地址 6, 7, 8, 9 被分为一组, 但地址 8 指代的单位在北京, 地址 6, 7, 9 指代的单位在广州. 地址 8 与地址 6, 7, 9 应当是两个不同的分组. 因此, 通过邮编和城市名称对分组结果进行纠正非常必要. 通过纠正后, 表 1 中的地址的分组如下: [[0, 1, 2,], [3, 4], [5], [8], [6, 7, 9], [10, 11]].

表 2 单位地址相似矩阵

编号	0	1	2	3	4	5	6	7	8	9	10	11
0	1.000	1.000	1.000	0.097	0.097	0.097	0.419	0.419	0.419	0.419	0.636	0.714
1	1.000	1.000	1.000	0.097	0.097	0.097	0.419	0.419	0.419	0.419	0.636	0.714
2	1.000	1.000	1.000	0.097	0.097	0.097	0.419	0.419	0.419	0.419	0.636	0.714
3	0.097	0.097	0.097	1.000	1.000	0.300	0.450	0.450	0.450	0.450	0.032	0.000
4	0.097	0.097	0.097	1.000	1.000	0.300	0.450	0.450	0.450	0.450	0.032	0.000
5	0.097	0.097	0.097	0.300	0.300	1.000	0.250	0.250	0.250	0.250	0.226	0.133
6	0.419	0.419	0.419	0.450	0.450	0.250	1.000	1.000	1.000	1.000	0.161	0.200
7	0.419	0.419	0.419	0.450	0.450	0.250	1.000	1.000	1.000	1.000	0.161	0.200
8	0.419	0.419	0.419	0.450	0.450	0.250	1.000	1.000	1.000	1.000	0.161	0.200
9	0.419	0.419	0.419	0.450	0.450	0.250	1.000	1.000	1.000	1.000	0.161	0.200
10	0.636	0.636	0.636	0.032	0.032	0.225	0.161	0.161	0.161	0.161	1.000	0.905
11	0.714	0.714	0.714	0.000	0.000	0.133	0.200	0.200	0.200	0.200	0.905	1.000

步骤 4. 单位地址标准化. 在该步骤中, 本文使用每一个分组中使用最频繁的单位地址作为该分组的标准单位地址.

步骤 5. 分组结果可视化校正. 如图 2(a) 所示. 通过对单位地址的拖拽, 专家可以将单位地址拖至更合适的分组或者新建一组. 如图 2(b) 所示, 把单位地址拖到 Affiliation address 节点, 自动连线成功即表示成功新建分组. 也可以把单位地址拖到别的圆圈处, 连线成功即表示成功调整分组. 确认无误后进行单位名称的提取, 更新数据库.

(a) 可视化校正工具图1

(b) 可视化校正工具图2

图 2 分组结果可视化校正工具

3.2 姓名处理

文献数据中作者姓名主要存在两大问题. 一是同一作者在不同论文中署名不同, 二是不同作者拥有相同的署名. 在描述本文解决方案之前, 需要先分析中文名和英文名的结构, 以帮助理解后面的算法.

英文姓名主要有名, 中间名, 姓组成. 在文献中, 姓不会简写; 名会写成全名或首字母, 但不会省略; 中间名会写成全名, 首字母, 或直接省略. 因此, 名, 中间名和姓的不同形式的组合导致了作者英文署名的多样化. 此外, 有的期刊要求先名后姓, 有的要求先姓后名, 这也增加了作者识别的难度. 比如:“Craig Brian, Agnor”, “Agnor, Craig Brian”, “Agnor, Craig B.”, “Agnor, C. Brian”, “Agnor, C.B.”, “Agnor, C. –B.” and “Agnor, C.” 等都可以指代同一个作者.

中文名主要由姓和名组成, 有时先名后姓, 有时先姓后名. 文献中, 作者的名有时全写, 有时只写名的拼音的首字母. 如表 3 所示, 当作者的名由两个及以上的字构成时, 字之间可能会用“-”, “,”或空格以及它们的组合来分隔, 且字也会采用拼音首字母, 有时甚至只保留名中第一个字的拼音首字母. 此外, 作者署名中还存在复姓, 如“Pu Yang”, “Ou Yang”, “Ai Xing Jue Luo”等

表 3 一个中文名的不同写法

类型	示例
名姓	Xiao-Ming Wang; Xiao-ming Wang; XiaoMing Wang; Xiaoming Wang; X.M. Wang; X. M. Wang; X. Wang;
名, 姓	Xiao-Ming, Wang; Xiao-ming, Wang; XiaoMing, Wang; Xiaoming, Wang; X.M., Wang; X. M., Wang; X., Wang;
姓名	Wang Xiao Ming; Wang XiaoMing; Wang Xiaoming; Wang Xiao-Ming; Wang Xiao -Ming; Wang Xiao -ming; Wang Xiao-ming; Wang X.M.; Wang X. M.; Wang XM; Wang X.
姓, 名	Wang, Xiao Ming; Wang, XiaoMing; Wang, Xiaoming; Wang, Xiao-Ming; Wang, Xiao-ming; Wang, Xiao -Ming; Wang, Xiao -ming; Wang, X.M.; Wang, X. M.; Wang, XM; Wang, X

在对作者姓名结构进行分析的基础上, 本文设计了基于规则的作者姓名识别算法, 并用 3.1 节中提到的分组可视化工具辅助专家校正算法结果. 姓名识别算法步骤如图 3 所示. 其中, 涉及的主要的方法如下.

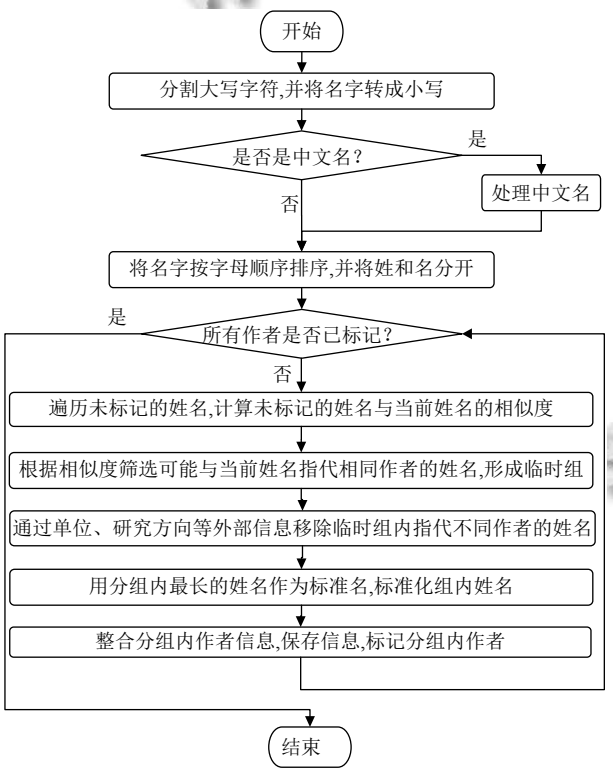

图 3 姓名处理流程图

方法 1. 分割大写字母. 连续大写字母的数量通常情况下与组成名或姓的字的数量一致, 每个大写字母是对应字的拼音的首字母. 因此, 这个重要特征可以用来识别作者名字的结构, 从而进一步计算两个名字的相似度. 尤其当作者姓名中存在多个缩写时, 分割大写字母的重要性更显凸显. 如“White, Simon D. M.”经常会写成“White, SDM”. 如果没有对大写字母进行分割, “SDM”会被识别成一个名.

方法 2. 中文名的识别和预处理. 由于拼音数量和中文名结构有限, 可以通过有限的词典和规则来识别中文名. 优先处理中文名不仅可以缩小后续匹配的数据范围, 还可以降低识别规则的复杂度. 在该步骤中, 首先识别中文名, 并根据结构特点区分姓和名. 接着将多个字组成的名中间的分隔符去掉并转成小写. 如“Wang, Xiao-ming”会处理成姓“wang”名“xiaoming”, 这样可以时后续字符串匹配和相似度计算提供帮助.

方法 3. 姓名归一. 文献中, 作者的姓一定排在最前面或最后面, 第一个名一定排在中间名前面. 因此, 当识别两个作者姓名 $A=\{A_1,A_2,\cdots,A_n\}$ 和 $B=\{B_1,B_2,\cdots,B_m\}$ 是否指代同一个人时, 可以通过如下规则判断, 其中 $m\le n$:

(1) 若 $\{A_1,A_n\}\cap\{B_1,B_m\}=\varnothing$, 那么 A 和 B 的姓不同, A 和 B 指代的不是同一个人;

(2) 若 $\{A_1,A_n\}\cap\{B_1,B_m\}\neq\varnothing$, 删除 A 和 B 中完全相同的部分, 剩余部分标记为 A' 和 B', 表示为 $A'=\{A'_1,\cdots,A'_{n-1}\}$, $B'=\{B'_1,\cdots,B'_{m-1}\}, m\le n$. 如果 A' 和 B' 都是空集, 那么这两个名字被认为指代了同一个人. 否则, A 和 B 指代同一人的必要条件是 A' 和 B' 中留下的是名, 即删除的完全相同部分是姓或姓和部分名.

(3) 依次匹配 A'_i 和 B'_i, $i\in[1,m-1]$. 只要有一个不匹配, A 和 B 指代的不是同一个人. 否则, A 和 B 被临时认为是同一个人, 分到同一个临时组. 需要注意的是, 此处所说的匹配并不是指相同. 在 A'_i 和 B'_i 都不是简写

的情况下, 只有当A_i'和B_i'完全相同时, 才算是匹配. 在A_i'和B_i'存在简写的情况下, 只要首字母相同, 就认为是匹配.

方法 4. 额外信息自动校正. 由于不同的作者的姓名可能相同, 所以不能认为相同的姓名指代了同一个人. 为了区分出同名作者, 本文用单位、研究领域、合著者等额外信息来验证同名作者是否是同一个人.

3.3 团队挖掘

研究团队挖掘可以转化为在以作者为节点, 以合著论文数为边的合作关系图上发现社团的问题. Louvain 算法[17] 是性能最好的社团发现算法之一, 是一种层次模块度优化算法, 具有快速、准确的特点. 模块度最初用于衡量社团发现算法结果的质量, 它的本质是刻画社团的紧密程度, 即社团内部紧密度越高, 社团之间紧密度越低, 社团划分的质量越高. 本文直接采用 Louvain 算法进行社团发现, 其它社团发现算法也可以用于挖掘本文所指的研究团队.

4 学者画像可视化

4.1 可视化指标选择

学者的竞争力评价是一项相当复杂的工作, 应当从尽可能多的维度开展评价. Keathley-Herring 等[18] 通过研究 1983~2016 年期间发表的 123 篇领域成熟度评估相关文章, 提出了一个通用的领域成熟度评估指标体系, 包括研究主题、论文质量、论文数量、合作交流、任职情况、学历、专利、项目、成果转化、学术影响力、社会影响力等多个维度的指标. 该指标体系不仅可以用于评估一个领域的成熟度, 对学者的综合实力评估也有很大的参考价值. 本文从该指标体系中抽取基于论文信息的指标, 并综合考虑数据的可获取性, 从科研产出、科研合作、科研影响 3 个方面筛选指标. 科研产出主要考虑学者发表的论文数量及其论文质量. 论文数量不仅包括论文的总数, 还可以是高质量论文的数量、不同研究方向的论文数量、近几年的论文数量等等. 论文质量通常可以在一定程度上通过论文发表的期刊影响因子来反映. 而影响因子与期刊的被引次数和文章数量有关, 因此也可以从学者的被引次数和文章数量上体现学者的论文质量. 科研合作可以反映学者的研究团队及其研究模式. 比如有的学者合作的学者和机构非常广泛, 有的学者有稳定的合作团队, 前者可能更适合做交叉领域的研究, 后者可能更适合在专业领域内进行科研攻关. 科研影响主要包括学术影响和社会影响. 学术影响通常使用被引次数来展现. 一篇论文的被引次数越高, 在一定程度上说明该论文对越多的科学研究产生了影响. 社会影响是对学术影响的重要补充, 可以通过在社交网络、在线科研论坛等网络平台上的热度, 对政策、法律等的影响来体现. 但分析社会影响所需数据的获取难度较大, 本文基于论文数据, 仅考虑学者的学术影响.

综上所述, 本文采用论文数量、被引次数和篇均被引次数作为基本指标. 同时增加了 H-index 指标供参考. H-index 定义为: 一个指数为 h 的学者发表了 h 篇论文, 并且每篇论文至少被其他论文引用 h 次. 它被认为能比较准确地反映学者的学术成就. 除了基本指标外, 本文从论文产出、研究兴趣、科研团队、合作情况多个维度来展现学者的科研状态.

4.2 可视化设计

图 4 展示了某个学者的学术状态和竞争力. 该学者的学术状态可视化主要由 6 个主要部分组成: 图 4(a) 学者姓名和单位 (此处为了保护隐私, 去掉了真实姓名和单位); 图 4(b) 学术状态基本指标; 图 4(c) 研究兴趣; 图 4(d) 论文合作情况浏览器; 图 4(e) 合作矩阵图和图 4(f) 论文发表和被引情况.

学术状态基本指标用雷达图来展示. 4 个坐标轴对应 4 个指数: 论文总数、H-index、总被引次数、篇均被引次数. 每个轴的最大值是该学者所属单位中所有学者对应指标的最大值. 蓝色实线表示该学者的各项指标的数值, 黑色实线表示该单位所有学者各项指标的平均值.

研究兴趣采用关键词云来展示. 关键词云显示了该学者所有论文中出现频次最高的前 10 个关键词. 字体大小代表该词在该学者所有论文中出现的次数. 这些词可以通过鼠标点击进行交互. 当选择一个词时, 其它图表将同步更新, 以展示该研究方向相关的学术信息.

论文合作情况浏览器是对该学者的论文、合著者和合作单位的交互式总体可视化. 图表中间是合著者列表. 左边带灰色圆点的是合作单位, 右边带白色圆圈的是合著论文. 这些列表之间的连线表示合作关系. 当鼠标悬停到某个条目 (合著者、论文或合作单位) 上时, 与其有关系的相应条目将高亮显示. 点击某个合著者的

名字将显示该合著者的所属单位和该合著者与该学者合作发表的论文信息, 如图 5 所示. 在图 5 中, 中间较大的黑色圆点代表选中的合著者. 左边的小白点是与该学者合著的论文. 与之相连的小灰点是合著论文的其他作者. 右边的小灰点是合著者所属的单位, 与之相连的小灰点是与该合著者属于同一单位且与该学者合著过论文的学者.

合作矩阵图用于展示该学者与其合著者之间的合作信息, 并可以发现研究团队之间的关系. 如图 6 所示, X 轴和 Y 轴均代表该学者所有合著者. 如果两学者之间有合作, 那么我们用颜色填充相应坐标的网格. 合作的次数越多, 颜色就越深. 合作的详细次数可选择地显示在网格中. 属于同一科研团队的学者会组成一个内部填充较为密集的区域, 如图 6 中红线所围部分.

论文发表和被引情况显示该学者论文发表的时间、论文的影响力. X 轴为发表年份, Y 轴为月份, 一个圆点代表一篇论文. 圆点的半径表示该论文的被引用次数; 圆点的颜色表示该论文的合作单位. 折线图显示了该作者历年发表论文的数量.

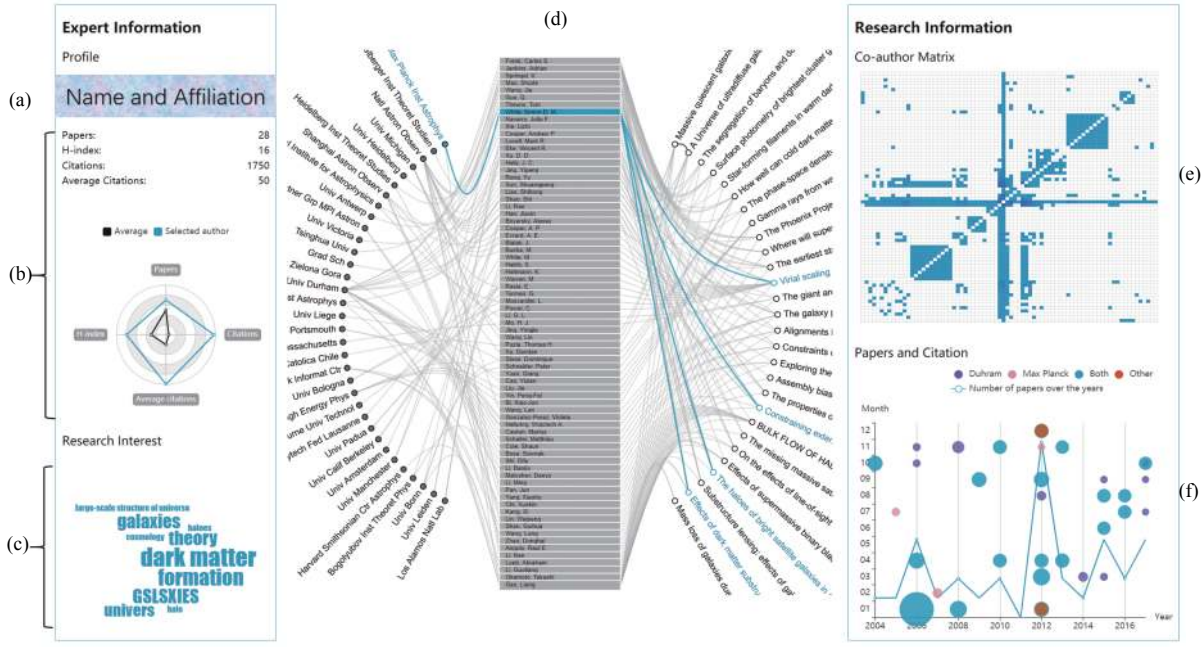

图 4 某学者画像的可视化

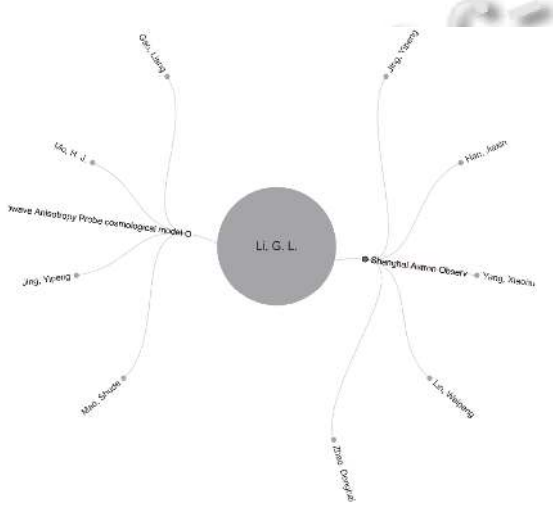

图 5 某个合著者的信息

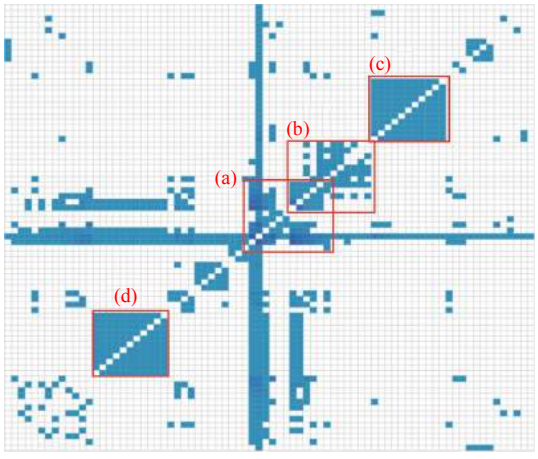

图 6 学者 G 的合作矩阵图

5 案例分析

我们研究了中国科学院发表的论文, 并选择某研究所 H 及其中的一个学者 G 作为研究案例, 因为 H 是中国科学院发表论文数量排名前 10 的研究所, G 在 H 研究所的所有学者中各项学术指标 (论文总数、H-index、总被引次数、篇均被引次数) 中都排名前五. G 的学术状态如图所示.

从图 4(b) 的基本指标可以看到学者 G 在总被引次数和篇均被引次数方面相当突出, 是该单位所有学者中这两个指标上表现最好的. 他在其他两个指标上的表现都比平均水平好, 但和这两个指标的最大值之间还存在明显的差距.

图 4(c) 中的研究兴趣关键词云在一定程度上体现了学者 G 的研究兴趣点. 从图 4(c) 中可以看出, 学者 G 是天文学专业的学者, 主要从事宇宙形成的理论研究. 他的研究集中于宇宙学、暗物质、星系研究等方向.

从图 4(d) 所示的论文合作情况浏览器中, 我们可以方便地探索学者 G 的论文、合著者和合作单位. 学者 G 与大约 72 个不同的学者合著过论文. 这些合著者来自约 34 个不同的组织. 学者 G 本人主要与 3 个机构有联系: 国家天文台、达勒姆大学和马克斯普朗克天体物理研究所.

学者 G 的合作矩阵图如图 6 所示: (a) 区域表示学者 G 的主要科研团队, 这种团队具有高度填充的网格和相对密集的深色色块等特征. 高度填充的网格和相对密集的深色色块意味着团队中几乎每个人相互之间存在多次合作. (b)~(d) 区域中团队成员之间合作次数很少, 绝大部分只合作一次, 颜色是浅的, 我们可以看出他们只是暂时合作写了一篇论文. (a) 和 (b) 有重叠部分, 说明这两个团队之间有共同的成员.

学者 G 的论文发表和被引情况如图 4(f) 所示. 我们可以看到学者 G 是一个高产的研究人员. 一篇发表于 2006 年的论文被引用率相当高. 2012 年是 G 学者最有成效的一年, 论文的数量和质量都是历年来最好的一年. 除了 2011 年外, 每年都有论文发表. 学者 G 主要与英国达勒姆大学和德国马克斯普朗克天体物理研究所合作. 紫色圆点代表学者 G 只与达勒姆大学合作的论文. 粉色的圆点代表学者 G 只与马克斯·普朗克天体物理研究所合作的论文. 蓝色的圆点代表学者 G 与两机构一起合作的论文. 从图表上可以明显看出, 学者 G 在 2004~2017 年的学术生涯有几个发展段: 2004~2005 年、2007~2011 年、2012~2014 年, 这 3 个阶段为学术累积期, 学者 G 发表论文的数量和质量都相对较低. 2006 年、2012 年、2015 年为学术爆发年, 经过累积期的积累, 迎来了高质量高产的年份.

综上所述, 我们可以将学者 G 的学术状态描述为: 学者 G 是天文学专业的学者. 他主要从事宇宙形成的理论研究. 他的研究集中于宇宙学热点、暗物质、星系等方向. 学者 G 在他所属的科研单位的所有学者中科研实力出众, 尤其是在论文被引用次数方面首屈一指, 是一位高被引作者. 他几乎每年都发表论文, 学术研究很有活力. 学者 G 的学术合作范围很广, 合著者约有 72 人, 来自约 34 个不同的科研机构, 主要合作机构有 3 个: 国家天文台、达勒姆大学和马克斯普朗克天体物理研究所. 学者 G 拥有一个由 5~7 名研究人员组成的主要研究团队, 并且与其他 3 个研究团队有过短暂合作. 2006 年、2012 年、2015 年是 G 学者的学术爆发年, 论文质量和数量都很好. 其中 2012 年是最突出、最有成效的一年.

6 结论与展望

本文基于 WOS 论文数据, 采用了可视化和算法相结合的实体消歧方法, 针对数据特征设计相应的人名、单位名自动分组算法, 并设计了分组可视化工具来帮助用户对算法结果进行校正, 以获得能满足分析要求的高质量数据. 接着, 选取了若干主要的学术评价指标, 设计了学者学术状态和竞争力的可视化方法. 在此基础上, 研发了一套学者状态和竞争力可视化系统. 此外, 本文在合著网络的基础上进行团队挖掘, 并设计可视化方法帮助用户更好地洞察研究人员的科研团队. 最后通过对中国科学院某研究所某学者的学术状态进行了分析, 证明了本文方法在分析学者的学术状态和竞争力方面的有效性和实用性.

然而本文存在一定的不足. 首先, 使用的数据存在一定局限性. 该局限性体现在两方面, 一方面本文使用的数据可能并不包括需要分析的机构和该机构的学者的所有论文数据, 另一方面仅仅基于论文数据不能全方位体现学者的综合实力. 其次, 评价指标不够丰富全面. 导致该不足的因素主要包括两方面, 一方面是因为本文的分析建立在论文数据上, 无法使用更全面的评价指标, 另一方面是因为学术竞争力评估本身是一个需要深入研究的复杂问题. 但本文为学者的学术状态和竞争力评估提供了新的思路, 通过可视化来全方位展现学者的综合实力能避免因使用单一指标引起的一

刀切问题. 随着可用数据维度的增加以及学术评价研究的日益深入, 可以用更多更权威的评价指标来替换并扩展本文中采用的指标. 此外, 本文将人引入数据清洗环节, 让人可以检查并校正算法结果, 可能会带来一定风险, 比如恶意破坏数据质量等. 但在正常使用情况下, 人对数据的校正是建立在经过验证或核实的基础上的, 对结果的客观性影响有限.

今后的工作将集中在以下 3 个任务. (1) 改进实体消歧模块. 目前实体消歧是将原始数据作为算法输入, 算法输出作为分组可视化工具的输入, 分组可视化工具的输出即为实体消歧结果. 未来需要考虑将算法改进成能交互式学习的模型, 并将分组可视化工具的输出反馈至算法, 从而形式闭环, 使系统更加高效智能. (2) 改进团队挖掘算法. 目前, 我们的在团队挖掘方面的工作仅限于寻找团队, 未来可以进一步挖掘团队内部关系、预测合作趋势. (3) 融合多种数据源, 建立更加全面的指标体系. 本文由于论文数据的局限性, 指标体系不能够全方位展示学者的竞争力. 如果能融合更多数据, 如专利、项目、经费、任职等, 可以从更多维度筛选指标以更加全面地评估学者竞争力. 此外, 可视化方法也需要进一步优化以适应更多的指标展现.

参考文献

1 Wang Y, Yu MZ, Shan GH, *et al*. VISPubComPAS: A comparative analytical system for visualization publication data. Journal of Visualization, 2019, 22(5): 941–953. [doi: 10.1007/s12650-019-00585-2]

2 Liu SX, Andrienko G, Wu YC, *et al*. Steering data quality with visual analytics: The complexity challenge. Visual Informatics, 2018, 2(4): 191–197. [doi: 10.1016/j.visinf.2018.12.001]

3 Liu SX, Chen CJ, Lu YF, *et al*. An interactive method to improve crowdsourced annotations. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 235–245. [doi: 10.1109/TVCG.2018.2864843]

4 Xiang SX, Ye X, Xia JZ, *et al*. Interactive correction of mislabeled training data. Proceedings of 2019 IEEE Conference on Visual Analytics Science and Technology. Vancouver, BC, Canada. 2019. 57–68.

5 Ferreira AA, Gonçalves MA, Laender AHF. A brief survey of automatic methods for author name disambiguation. ACM SIGMOD Record, 2012, 41(2): 15–26. [doi: 10.1145/2350036.2350040]

6 Torvik VI, Smalheiser NR. Author name disambiguation in MEDLINE. ACM Transactions on Knowledge Discovery from Data, 2009, 3(3): 11.

7 Milojević S. Accuracy of simple, initials-based methods for author name disambiguation. Journal of Informetrics, 2013, 7(4): 767–773. [doi: 10.1016/j.joi.2013.06.006]

8 Zhu J, Zhou XF, Fung GPC. A term-based driven clustering approach for name disambiguation. Proceedings of the Joint International Conferences on Advances in Data and Web Management. Berlin, Germany. 2009. 320–331.

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

10 Tang L, Walsh JP. Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. Scientometrics, 2010, 84(3): 763–784. [doi: 10.1007/s11192-010-0196-6]

11 Kanani P, McCallum A, Pal C. Improving author coreference by resource-bounded information gathering from the web. Proceedings of the 20th International Joint Conference on Artifical Intelligence. San Francisco, CA, USA. 2007. 429–434.

12 Pereira DA, Ribeiro-Neto B, Ziviani N, *et al*. Using web information for author name disambiguation. Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. Austin, TX, USA. 2009. 49–58.

13 Yang KH, Peng HT, Jiang JY, *et al*. Author name disambiguation for citations using topic and web correlation. Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries. Berlin, Germany. 2008. 185–196.

14 Kang IS, Na SH, Lee S, *et al*. On co-authorship for author disambiguation. Information Processing & Management, 2009, 45(1): 84–97.

15 Shen QM, Wu TS, Yang HY, *et al*. NameClarifier: A visual analytics system for author name disambiguation. IEEE Transactions on Visualization and Computer Graphics, 2016, 23(1): 141–150.

16 Latif S, Beck F. VIS author profiles: Interactive descriptions of publication records combining text and visualization. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 152–161. [doi: 10.1109/TVCG.2018.2865022]

17 Blondel VD, Guillaume JL, Lambiotte R, *et al*. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008. [doi: 10.1088/1742-5468/2008/10/P10008]

18 Keathley-Herring H, Van Aken E, Gonzalez-Aleu F, *et al*. Assessing the maturity of a research area: Bibliometric review and proposed framework. Scientometrics, 2016, 109(2): 927–951. [doi: 10.1007/s11192-016-2096-x]