

融合多策略的中文科技文献机构名称规范化研究与实践^{*}

刘 燕 孙月萍 侯 丽

(中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

〔摘要〕 分析中文科技文献中机构著录项的组织特点和中文机构名称的命名特点, 详细阐述常见机构名称规范化方法、中文科技文献机构名称规范化处理流程, 提出利用字符串匹配词典和规则过滤等方法提取规范化的机构名称, 并基于机构-作者共现关系, 计算作者共现率, 结合绝对共现量和共现率阈值实现机构实体的消歧, 能够有效匹配同一机构的不同表现形式。

〔关键词〕 机构名称规范; 科技文献; 作者共现; 实体挖掘

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.12.007

Study and Practice on Institution Name Normalization of Chinese Scientific and Technical Literature Based on Multiple Strategies LIU Yan, SUN Yueping, HOU Li, Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China

〔Abstract〕 The paper analyzes the organization characteristics of institution description items in Chinese scientific and technical literature and the naming characteristics of Chinese institutions, expounds the common methods of institution name normalization and the process of institution name normalization for Chinese scientific and technical literature, and proposes that the methods of extracting the normalized institution names by using the methods of string matching, dictionary-based and rule-based filtering, calculating the co-occurrence rate of authors based on the co-occurrence relationship between institutions and authors, and disambiguating the institution entities through the number of absolute co-occurrence and the co-occurrence rate threshold, which can effectively match different forms of an institution.

〔Keywords〕 institution name normalization; scientific and technical literature; author co-occurrence; entity mining

〔修回日期〕 2022-06-17

〔作者简介〕 刘燕, 助理研究员, 发表论文 6 篇; 通信作者: 侯丽, 研究员。

〔基金项目〕 中国工程科技知识中心建设项目“医药卫生专业知识服务系统”(项目编号: CKCEST-2022-1-6); 国家社科青年基金项目“基于语义增强的医学学术出版创新融合研究”(项目编号: 18CTQ024)。

1 引言

面对日益激增的海量数字化文献资源, 如何利用规范化的机构体系对文献资源进行整合、挖掘、分析等一直是学界关注的重点^[1]。近年来, 学界加强了对机构规范文档^[2-3]、机构知识库^[4]等的构建与应用研究, 从机构名称统一标识^[5]、机构类别特征化^[6]、机构名称相似度计算^[7]等角度, 推进规范

化机构在各种服务场景中的应用。机构作为科技文献的重要组成元素之一,是开展科研评价、信息检索、学术资源组织与关联的基础。但现实中文献机构名称著录混乱、层级结构模糊、更名、重组、合并、拆分等现象频繁,加之名称存在缩写、简称、书写不规范等问题,导致机构名称识别度降低,各类数据库和搜索引擎很难准确统计机构对应的资源数量^[8],从而影响统计分析和评价结果的可靠性^[9]。因此为有效整合并利用机构实体不同名称下的信息资源^[10],进行机构名称规范化的研究与实践至关重要。

机构名称规范一般是指通过收集机构实体的所有表现形式,实现多个机构名称到一个机构实体的映射^[11]。对于科技文献中的机构名称规范研究而言,其核心问题是提取“作者单位”著录项中的机构名称,并进行机构名称的消歧,使同一机构实体的不同名称表现形式都指向一处。学者们据此开展诸多研究,取得较好效果,然而还无法有效解决表达形式差异较大的机构名称规范问题,如“北京安贞医院”与“首都医科大学第六临床医学院”。对此,有学者通过发文著者共现情况来判断机构名称的相似度^[12-13],取得了一定效果,但未考虑不同类型机构的差异。鉴于此,本文尝试从“机构-作者”共现和机构类型特征词的角度,进行机构名称的规范化研究,分析不同类型机构名称的命名特点,并结合机构共现作者和相似度计算方法进行中文机构名称的消歧,最后以医学领域机构为例进行实践。

2 常见机构名称规范化方法

机构名称的规范化建设经历了规范控制、访问控制、唯一标识符等阶段。其中,规范控制是为各机构设置一个规范名称并将其他名称都指向它,缺点是检索其他名称时只能获取包含该名称的资源;访问控制则不设置规范名称,而是将所有名称都加入一个可访问的白名单中,检索任意名称都能获取全部资源,但多次检索会加重系统负担^[14];国际标准名称识别码^[15](International Standard Name

entifier, ISNI)、Ringgold 标识数据库^[16]等希望通过唯一标识符来实现机构的唯一识别,但由于目前并未形成统一的全球化方案,在文献数据中的应用程度还较低,因此利用唯一标识符解决机构实体的歧义问题更多是愿景和辅助手段^[14]。常见机构名称规范化方法主要有基于字符串相似度的方法、基于规则的方法、基于统计关联的方法和混合策略的方法。

2.1 基于字符串相似度的方法

基本思路是利用字符串相似度计算的方法判定机构名称相似性程度。常用方法包括 Levenshtein 编辑距离^[17]、Cosine 相似度、Jaccard 相似度等。有学者^[18-19]基于字符串编辑距离的方法构建机构名称规范文档。Ferosh J^[20]利用 Levenshtein 编辑距离方法对求职简历中求职者机构名称进行规范。Jiang Y 等^[21]基于归一化的压缩聚类方法实现对同一机构不同名称的聚类。

2.2 基于规则的方法

主要思想是基于建立的规则库对错误匹配对进行过滤。有学者^[22-23]根据机构名称的特点,提出基于规则的机构名称消歧方法,并在 Web of Science 不同学科数据集中进行有效性测试。沈嘉懿等^[24]针对网络文本数据提出基于规则识别中文组织机构名称的方法,借助机构后缀词库、规则匹配和贝叶斯模型识别机构边界。

2.3 基于统计关联的方法

基本思路是利用 Web 大规模语料,通过计算不同机构名称字符串搜索结果中统一资源定位符(Universal Resource Locator, URL)的共现情况来判定机构名称相似度^[25]。Aum Mueller D 等^[26]基于谷歌和雅虎搜索返回的前 k 个 URL 共现重叠情况来计算两个机构名称匹配程度。

2.4 混合策略的方法

主要思想是通过整合两种或两种以上的方法,来实现更高的机构名称识别精准度。杨瑞仙等^[27]提出一种基于规则和向量空间模型的科研机构名称识

别方法。孙海霞等^[9]提出一种基于规则和编辑距离的机构名称匹配策略，并以中文生物医学文献数据库为例进行实践。张建勇等^[14]基于规则和相似度计算的方法对国家科技图书文献中心内的科研机构实体进行消歧，以便构建科研合作网络等。

3 中文科技文献机构名称规范化处理流程

本研究以中文科技文献中的机构为例开展名称规范化研究。设计中文科技文献机构名称规范实现流程，包括数据采集、机构名称提取和机构实体消歧 3 个步骤，见图 1。

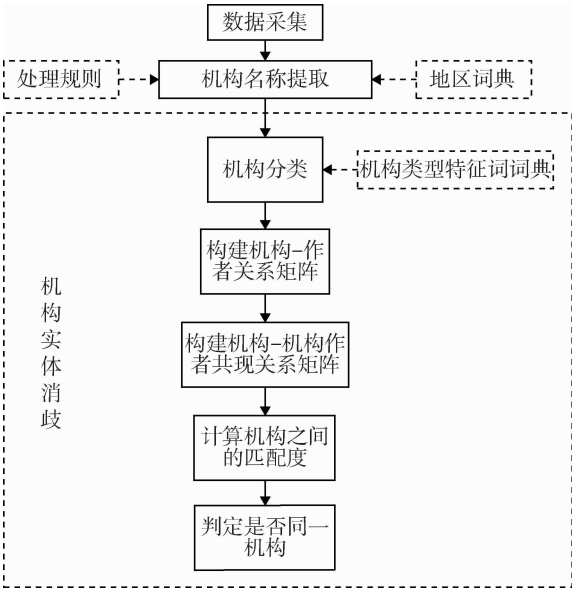


图 1 中文科技文献机构名称规范化处理流程

3.1 数据采集

科技文献来源包括数据库商、出版商、服务商等，不同来源的数据描述粒度不同，数据质量也有所差异。本研究制定数据采集方案如下：根据数据质量、权威性、采集要求，确定采集来源、时间范围、期刊等；确定需要采集的字段项，如题目、作者、机构著录项等；利用爬虫软件进行数据采集，完成格式转换与存储；制定规则对不完整数据和重复数据进行处理，将缺少文献题目、作者、机构等关键字段的数据直接剔除，删除重复数据中字段项较少的，判定重复数据的条件为两篇文献 DOI 是否

一致或题目、作者和期刊 3 项信息是否完全相同。

3.2 机构名称提取

3.2.1 概述 机构名称在科技文献中的表述形式多样，存在问题主要包括两点：机构合作客观存在，且 1 位作者可能会隶属于多个机构，故 1 篇文献可能会存在多个机构的现象（简称多机构）；机构著录项标注形式不统一，且不同期刊对机构著录项要求不同，如邮编位置、是否标注机构所在国家、机构是否为独立法人等。鉴于此，本研究将利用字符串匹配、词典和规则过滤的方法进行规范化机构名称提取。

3.2.2 多机构拆分 将包含多个机构的数据拆分为多条数据，确保 1 条数据只包含 1 个机构及其对应的作者，便于统计机构发表的文献及隶属于机构的作者。拆分方法是先利用字符串方法找到机构著录项之间的分隔符，并以分隔符为边界完成机构拆分。

3.2.3 机构著录项拆分与过滤 对单机构的机构著录项进行拆分并过滤邮编、行政区划地址等信息，以获取作者原始著录的机构名称信息。（1）机构著录项拆分。以逗号或空格为分隔符对机构著录项包含的字段进行拆分，考虑到机构名称长度至少为 4，可直接过滤掉长度小于 4 的字段。（2）邮编和行政区划地址过滤。判断剩余的字段是否为邮编和行政区划地址，若是则直接删除。其中，邮编可使用字符串编辑的方法处理，若该字段由 6 位连续的数字组成，则判定为邮编；行政区划地址可通过构建国内各省市地区字典来处理。

3.2.4 机构名称规范化提取方案 本研究的规范化机构名称是指法人级别的机构，因此要对部门、科室等二级机构名称进行识别并删除。通常，中文机构名称以“A+B”的形式表达，A 部分一般由方位词、序数词、动词等构成，B 部分一般为“大学”“研究所”“医院”等用来表示机构特征的中心语，故可以通过 B 部分来判定机构名称是否已规范至法人级别。本研究设计面向中文科技文献机构名称规范化提取方案包括：（1）机构名称分词。构建机构名称词库，利用中文分词工具 Jieba 对机构

名称进行分词，得到 A 和 B 两部分。(2) 构建机构特征词表。结合国家机构类型分类标准《组织机构类型 (GB/T 20091—2006)》，将机构分为科研机构、高等教育机构、医疗机构、事业单位、行政机构、公司企业、社会团体、其他 8 类，进而利用中文机构名称的命名特点，构建机构类型特征词表。(3) 识别机构名称著录深度。依次比较机构名称的 B 部分与机构类型特征词表有无匹配项，若有匹配项则不作处理，若无匹配项则表明该字段包含二级机构名，应从右至左依次遍历分词列表，直到匹配到正确的机构中心语，并将中心语右侧的二级机构名删除，得到规范的一级机构名称，见图 2。

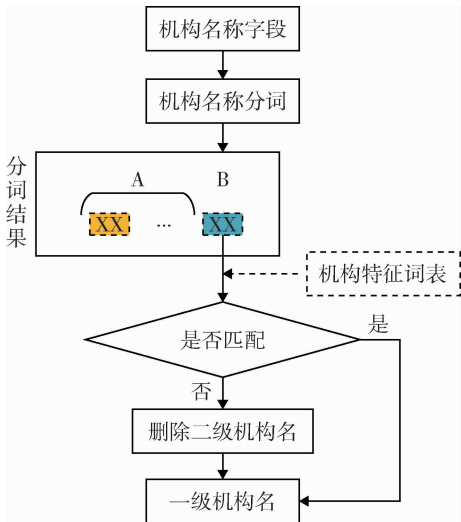


图 2 中文科技文献中的机构名称规范化提取方案

3.3 机构实体消歧

3.3.1 概述 可用于机构实体匹配的文献特征有机构名称、行政区划地址、邮编等，但很多机构著录项中的行政区划地址和邮编信息并不完整。因此本研究考虑从机构名称出发，构建“机构－作者”关系表，并基于机构类型特征词典对机构进行分类，进而面向不同机构类别分别构建“机构－机构”作者共现矩阵、计算作者共现率，以实现机构实体消歧。假设不同类别中的机构名不可能指向同一机构实体，即无需匹配不同类别之间的机构名，这样一方面可以减少机构之间两两匹配的次数，提高计算效率；另一方面能够降低错误匹配的机率，提升匹

配准确率。

3.3.2 构建“机构－作者”关系表 通常，机构发表的文献都不止 1 篇，故本研究先以机构为中心对文献进行聚类，聚类个数即为待消歧机构名称的数量，从而得到各机构发表的文献集合，整合对应集合中的作者，完成“机构－作者”关系表构建。作者消歧是实体消歧的另一关键问题，非本研究重点，故暂不考虑作者同名的情况。

3.3.3 机构分类 利用分词工具对上述规范至法人级别的机构名称进行分词处理，选取能够代表机构类型的中心语，即分词列表中的最后一个词，依次与机构类型特征词表中的特征词进行比较，据此得到各机构名称的分类。以“中国人民大学”为例，首先分词得到“中国/ns 人民/n 大学/n”，然后选择分词列表中的最后一个词“大学”与机构特征词表进行匹配，发现该机构名称属于“高等教育机构”。

3.3.4 机构消歧 本研究假设，在一段时期内机构成员会保持相对稳定^[28]，因此可通过机构之间的作者共现率来推断不同机构名是否指向同一实体。此外，考虑到本研究涉及的机构类型多样，如公司企业、社会团体等机构发文量难以保证，无法避免由于发文量低而导致的重名风险，即若某机构发文量极低（如小于 5），则可能因个别作者重名而导致作者共现率超过阈值^[29]，影响消歧准确率。因此综合考虑机构对之间的作者共现率和作者绝对共现量指标，即针对不同类别机构的数据，循环遍历“机构－作者”关系表中的 n 个机构，依次比较机构 m ($1 \leq m \leq n$) 和剩下的 $n - 1$ 个机构，统计两机构各自的作者数、机构间的共同作者数和全部作者数，构建“机构－机构”共现矩阵，计算机构对之间的作者共现率，此处共现率是指机构的共同作者占全部作者的比值，见公式 (1)，进而确定共现率阈值（如 0.3），并据此筛选出具有同一关系的候选机构对；利用作者绝对共现量（机构间的共同作者数）指标控制重名风险：若作者绝对共现量大于等于 2，判定两个机构名称指向同一机构实体，否则即使机构对的作者共现率大于等于阈值，仍将其判定为非同一实体。

作者共现率 = $\frac{A \cap B}{A \cup B} \times 100\%$

(1)

其中, A 和 B 分别为两个机构对应的作者集合, $A \cap B$ 为两个机构的共同作者数, $A \cup B$ 为两个机构的全部作者数。

3.4 评价指标

主要采用准确率 P 来评价本文提出的中文科技文献机构名称规范化方案有效性, 见公式 (2)。

$$P = \frac{n}{N} \times 100\%$$
 (2)

其中, n 为人工审核的正确机构对数量, N 为识别出的机构共现对数量。

4 机构名称规范化实践

医药卫生知识服务系统 (<https://med.ckcest.cn>) 整合大量医学领域的科技文献、专家、机构、专利等学术资源, 但科研成果中的机构名称存在著录混乱、

层级结构模糊、更名频繁等问题, 导致机构名称识别困难, 难以开展文献、专家、机构等科研实体之间的进一步关联分析与深入挖掘。为进一步提高机构名称识别效率, 打通不同类型学术资源之间的壁垒, 提高用户信息检索效率, 需要对机构名称进行规范化处理。本研究以医药卫生领域的中文科技文献为例, 开展机构名称规范化实践, 验证提出的机构名称规范化处理方案是否可行。

4.1 数据采集

选取医药卫生知识服务系统作为数据来源, 筛选医药卫生领域相关的期刊进行采集, 采集内容包括文献题目、作者、机构著录项等, 共采集 1999—2020 年发表的文献数据 10 万条, 完成数据格式转换与存储, 并对不完整数据和重复数据进行预处理, 剔除文献题目、作者、机构等关键字段不完整的数据, 得到相对规范、完整的数据, 见表 1。

表 1 部分采集样例数据

题目	作者	机构著录项
1 例 HIV 抗体阳性待确定产妇的个案调查	李虹璇 ¹ ; 冷冰 ¹ ; 王德伟 ¹ ; 许学江 ¹ ; 陈丹芍 ¹ ; 罗英 ¹	成都市成华区疾病预防控制中心, 成都 610051 ¹
复发性流产绒毛、蜕膜组织中 TIPE2、TNF-α、IL-10 的表达及意义	王丽丽 ¹ ; 赵敏 ²	山西医科大学研究生学院, 太原 030001 ¹ ; 山西省妇幼保健院妇科, 太原 030001 ²

4.2 机构名称提取

4.2.1 多机构拆分 从采集的中文科技文献数据可知, 其机构著录项之间都是通过分号进行分割。因此以分号为分隔符, 利用字符串方法对机构进行拆分, 拆分后共得到包含单机机构记录的数据 350 587 条。

4.2.2 机构著录项拆分与过滤 对于拆分后的单机机构记录, 其机构名称、行政区划地址和邮编之间均以空格或逗号作为分隔符, 据此可先对机构著录项进行初步拆分, 并直接剔除长度小于 4 的字段。然后, 基于字符串编辑方法过滤掉剩余字段中的邮编。最后, 基于构建的国内各省市地区字典识别并删除行政区划地址, 只保留作者原始著录的机构名称。

4.2.3 机构名称规范化处理 系统分析并构建医药卫生领域机构类型特征词表, 该词表共覆盖 8 种

类型机构, 包含特征词 103 个, 其中医疗机构最多 (41 个), 其次为事业单位 (22 个), 见表 2。

表 2 医药卫生领域机构类型特征词

机构类型	特征词
科研机构	研究所、研究院、科学院等
高等教育机构	大学、高等专科学校、职业技术学院、职业学院、职业技术学校等
医疗机构	医院、保健院、卫生院、疗养院、诊所、卫生所、医务室、卫生室、防治院等
事业单位	血站、服务站、控制中心、监测中心、血液中心、疾控中心等
行政机构	厅、局、委员会等
公司企业	公司、集团等
社会团体	医学会、协会、学会等
其他	总队、部队、师、支队等

对作者原始著录的机构名称数据进行分词、识别机构著录深度并删除相应的二级机构名称，完成机构名称规范化处理，见表 3。

表 3 规范化机构名称部分示例

题目	作者	机构
1 例 HIV 抗体阳性待确定产妇的个案调查	李虹璇; 冷冰; 王德伟; 许学江; 陈丹苒; 罗英	成都市成华区疾病预防控制中心
复发性流产绒毛、蜕膜组织中 TIPE2、TNF-α、IL-10 的表达及意义	王丽丽	山西医科大学
复发性流产绒毛、蜕膜组织中 TIPE2、TNF-α、IL-10 的表达及意义	赵敏	山西省妇幼保健院

4.3 机构实体消歧

- 4.3.1 构建“机构-作者”关系表 以机构为中心对文献进行聚类，共得到 15 088 个聚类集合，分别整合各集合中的作者，构建“机构-作者”对应关系表。
- 4.3.2 机构分类 基于医药卫生领域机构类型特征词表，对上述机构名称进行分类，其中，医疗机构占比最高，其次为事业单位，社会团体最低，见表 4。

表 4 医药卫生领域机构名称分类情况

机构类型	数量（个）	百分比（%）
医疗机构	9 917	65. 73
事业单位	1 365	9. 05
高等教育机构	1 343	8. 90
行政机构	913	6. 05
科研机构	734	4. 86
其他	727	4. 82
公司企业	68	0. 45
社会团体	21	0. 14

4.3.3 构建“机构-机构”共现矩阵 按照分类，依次计算各类别中机构对之间的作者共现率。经统计

共 14 592 个机构对间存在作者共现情况，考虑到共现率小于 0. 1 时误判率过高，分析意义不大，本研究只针对共现率大于等于 0. 1 的 2 088 个机构对进行比较分析，并将根据不同类型机构在数据集中所占的比例，按照同等比例从中随机遴选 300 个机构共现对，进行准确率的分析。需要说明的是由于“其他”类型中共现率大于等于 0. 1 的机构对共 2 个、“社会团体”共 0 个，故实际遴选出来的相较按比例的数量少（若按比例应遴选“其他”14 个、“社会团体”1 个），因此最终子集共包含机构共现对 287 个。由专业人员进行结果准确性测评，经分析，将共现率阈值设置为 0. 1 时准确率可达 89. 2%，具有较高的机构实体消歧能力，尽管随着阈值的提升，准确率也呈上升趋势，但提升幅度较小，同时也会过滤掉很多虽然共现率低但实际为同一实体的机构对，故本研究暂将共现率阈值设置为 0. 1。

4.4 结果

通过统计，随机遴选的阈值大于等于 0. 1 的 287 个机构共现对中，人工认为其中 256 个机构对是同一机构，整体准确率为 89. 2%，具有较好的可参考性。此外，为进一步比较该方法对于不同类型机构的消歧效果，针对各类机构分别进行了误判率统计。其中，“其他”类型误判率最高，究其原因是该类型数据太少，少量误判就会造成大的结果偏差；“高等教育机构”和“行政机构”类型误判率也显著高于其他类别，其原因可能是这两类机构存在更为频繁的更名、重组、拆分等现象，依据较低的共现率难以实现机构实体的有效识别。后续可通过进一步扩大数据集或提升共现率阈值来提高其准确率。

5 结语

规范化的机构名称是开展面向机构的科技评价、异构学术资源整合、学术图谱构建等工作的基础与关键。本研究从“机构-作者”共现和机构类型特征词的角度，开展面向中文科技文献数据的机构名称规范化研究，通过分析科技文献中不同类型机构名称的著录特点，并结合作者共现情况进行机

构名称的消歧,最后在医学领域进行验证。经测试评估,该策略能够有效匹配同一机构的不同表现形式。后续将进一步优化消歧策略,扩大实验数据集并尽快推进其在医药卫生知识服务系统中的应用。通过机构间的作者共现率可以有效规范机构名称,实现机构实体不同名称形式的全面聚类与挖掘。但从长远发展来看,建议积极落实对机构唯一识别码的使用,特别是发表论文、专利等成果时,准确标识不同机构实体,从而更好地开展机构评价、构建机构知识库、构建学术知识图谱、规范存储机构知识资源等工作。

参考文献

- 1 杨昭. 基于元路径的机构名称归一化研究 [J]. 情报学报, 2020, 39 (10): 1069-1080.
- 2 孙海霞, 李军莲. 学术论文作者机构规范文档构建 [J]. 医学信息学杂志, 2015, 36 (11): 42-47.
- 3 高星, 戴玮, 黄利辉, 等. 中文生物医学文献机构名称规范化研究 [J]. 医学信息学杂志, 2010, 31 (12): 56-60.
- 4 李慧佳, 马建玲, 张秀秀, 等. 中文机构名称规范库建设的实践与分析——以“中科院机构名称规范库”建设为例 [J]. 图书与情报, 2016 (1): 133-139.
- 5 贤信, 曾建勋. 科研实体唯一标识系统研究 [J]. 图书情报工作, 2015, 59 (12): 113-119.
- 6 李家琪, 赵捷, 杨代庆. 机构规范文档构建中机构筛选方法研究 [J]. 中华医学图书情报杂志, 2018, 27 (5): 35-41.
- 7 Wan H, Zhang Y, Zhang J, et al. AMiner: Search and Mining of Academic Social Networks [J]. Data Intelligence, 2019, 1 (1): 58-76.
- 8 唐金玲. 国际三大检索系统论文作者机构名称问题研究——以高校机构名称为例 [J]. 情报探索, 2014 (9): 80-84.
- 9 孙海霞, 王蕾, 吴英杰, 等. 科技文献数据库中机构名称匹配策略研究 [J]. 数据分析与知识发现, 2018, 2 (8): 88-97.
- 10 曾建勋, 郭红梅. 基于知识组织的机构规范文档构建方法研究 [J]. 中国图书馆学报, 2021, 47 (1): 61-75.
- 11 周毅, 张建勇, 刘峥, 等. 科研实体名称规范的关联数据模型构建 [J]. 图书情报工作, 2020, 64 (10): 109-117.
- 12 吴英杰, 李军莲, 孙海霞, 等. 基于著者共现的 CBM 机构名称规范研究 [J]. 医学信息学杂志, 2019, 40 (12): 48-53.
- 13 贾君枝, 曾建勋, 李捷佳, 等. 科研机构名称归一化实现 [J]. 图书情报工作, 2018, 62 (13): 103-110.
- 14 张建勇, 钱力, 于倩倩, 等. 科研实体名称规范的研究与实践 [J]. 数据分析与知识发现, 2019, 3 (1): 27-37.
- 15 ISNI. International Standard Name Identifier [EB/OL]. [2022-06-14]. <https://isni.org/>.
- 16 Ringgold. Identity Database [EB/OL]. [2021-11-22]. <https://www.ringgold.com/identify>.
- 17 Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals [J]. Soviet Physics Doklady, 1966, 10 (1): 707-710.
- 18 French J C, Powell A L, Schulman E, et al. Automating the Construction of Authority Files in Digital Libraries: a Case Study [M]. Berlin: Springer, 1997.
- 19 French J C. Using Clustering Strategies for Creating Authority Files [J]. Journal of the American Society for Information Science, 2000, 51 (8): 774-786.
- 20 Ferosh J. SCool: a System for Academic Institution Name Normalization [C]. USA: 2014 International Conference on Collaboration Technologies and Systems (CTS), 2014.
- 21 Jiang Y, Zheng H, Wang X, et al. Affiliation Disambiguation for Constructing Semantic Digital Libraries [J]. Journal of the Association for Information Science and Technology, 2011, 62 (6): 1029-1041.
- 22 Huang S, Yang B, Yan S, et al. Institution Name Disambiguation for Research Assessment [J]. Scientometrics, 2014, 99 (3): 823-838.
- 23 杨波, 杨军威, 阎素兰. 基于规则的机构名规范化研究 [J]. 现代图书情报技术, 2015 (6): 57-63.
- 24 沈嘉懿, 李芳, 徐飞玉, 等. 中文组织机构名称与简称的识别 [J]. 中文信息学报, 2007 (6): 17-21.
- 25 Jonnalagadda S R, Topham P. NEMO: Extraction and Normalization of Organization Names from PubMed Affiliations [J]. Journal of Biomedical Discovery and Collaboration, 2010, 5 (1): 50-75.
- 26 Aumuller D, Rahm E. Web-based Affiliation Matching [C]. Potsdam: Proceeding of International Conference on Information Quality, 2009.
- 27 杨瑞仙, 毛一雷. 面向知识评价的我国科研机构命名识别方法研究 [J]. 情报杂志, 2015, 34 (7): 179-183.
- 28 刘进, 沈红. 中国研究型大学教师流动: 频率、路径与类型 [J]. 复旦教育论坛, 2014, 12 (1): 42-48, 92.
- 29 吕冬晴, 陆红如, 成颖, 等. 基于机构-作者向量的科研机构名称演化识别方法研究 [J]. 情报学报, 2020, 39 (6): 565-578.