外文数据库英译中文作者姓名消歧实践*

朱玉强[1] 江涛[2] 李翼飞[1]
(1. 山东师范大学图书馆，济南 250014；2. 海南医学院图书馆，海口 571199)

摘要：针对外文数据库英译中文作者姓名存在多记录指向同一人或同记录指向不同人等情况，模拟人工排检法，整合多源数据、学术社交网络、知识百科及在线翻译网站等语料库，利用网页文档对象自动操作、正则表达式、短文本相似度计算等技术编制程序开展英译中文作者姓名消歧实践。结果表明，算法架构稳定有效、扩展性强，成功率得到从业人员认可，为数据预处理和清洗工作提供了新思路和新方法。

关键词：姓名消歧；地址消歧；数据治理；外文数据库
中图分类号：G255.1　DOI：10.3772/j.issn.1673-2286.2022.02.005
引文格式：朱玉强，江涛，李翼飞. 外文数据库英译中文作者姓名消歧实践[J]. 数字图书馆论坛，2022 (2)：33-39.

Web of Science（WoS）、Scopus、Engineering Village（EI）等外文数据库收录英文学术论文，正文以外包括题名、作者、摘要、关键词等信息，其中作者信息包含作者姓名和所属机构名称（机构所在省市和邮编）。中国作者在外文期刊发文时按国际惯例须将中文姓名翻译为英文，因不同国家或地区期刊出版规范不同、不同历史时期数据加工标准有差异等原因[1-2]，有些作者姓名不是按汉语拼音方案翻译，如按威妥玛-翟理斯方案将“蒋介石”翻译为“Chiang Kai-shek”，将“张三丰”译为“Chang San-feng”[3]；即便使用汉语拼音方案，因期刊执行时格式有差异，同一作者有不同英译名或同一英译名对应不同作者的情况相当普遍，如“张三丰”有“Zhang（，）Sanfeng”[4]“Zhang（，）San-F（f）eng”“Sanf（F）eng（，）Zhang”“San-F（f）eng（，）Zhang”等译法，还有“Zhang（，）S.F.”“Zhang（，）SF”“S.F. Zhang”“Zhang（，）S.”“S. Zhang”等缩写版本，后两种译法甚至将“丰”字丢弃，可对应“章四凤”“张桑”“张思”等中文作者名。中文单姓单名英译因“姓前名后”或“名前姓后”原则不同造成的混乱尤甚，如将“姚明”翻译为“Yao Ming”[5]或“Ming Yao”[6]，后者亦可对应中文名“明瑶”“明尧”等。即便机构确切、人名拼音标记完全，还存在类似“明瑶”“明尧”音同字不同的情况。因此，仅依据外文数据库中作者英译姓名及机构名称确认其归属易错易漏，给文献计量工作带来诸多不便，进而使基于文献计量的情报分析、人才评价、参考咨询工作受到很大影响。有鉴于此，对外文数据库英译中文作者姓名进行消歧处理是进行数据清洗、提高数据质量的关键。英译中文作者姓名汉化消歧的难点在于英文缩写还原、同拼音汉字溯源及不同机构相同汉字人名身份的甄别，手工排检工作量繁杂巨大，如多人协作则数据质量难以统一，从业者对半自动或全自动数据处理工具的需求日益迫切。

1 相关研究与实践

英译中文作者姓名汉化消歧的解决方案按自动化程度可分人工、半自动和全自动3类[7]。人工排检实践方面，侯长来[8]对SCI论文中同一拼音著者，先将署名机构翻译为中文，再到《中国科技论文统计与引文分析数据库》中找到对应中文机构，查找该机构下有无同拼音著者进行辨识追踪。人工排检优点为结果准确，如原英文署名为“Hu G.C.”，找到对应中文机构后模糊匹配，

* 本研究得到2021年度海南省哲学社会科学规划课题（编号：hnsz2021-19）资助。

关注“胡贵超”“胡国策”“胡桂朝”“胡国才”“呼革彩”等疑似作者，再根据二级机构、专业方向等进一步筛选确认。极端情况下两位作者机构、专业方向甚至所在教研室都完全相同，如作者发文时自行用性别、年龄等做了标识，人工排检时即可按已有标识记为“胡贵超（男）”或“胡贵超（大）”，如无标识又确需分清彼此则只能和原作者联系。人工排检的缺点为效率非常低且成功率受中文对照库丰富程度影响，如“胡贵超”只发表外文论文从未发表中文论文，仅从中文期刊库这一语料库查找就无解，只能再借助于搜索引擎或百科网站等其他语料库。何春建[9]、高营[10]开展了以正则表达式为主要技术的半自动排检实践，该技术可从字符串中灵活提取指定文本，但此类实践仅能筛选或微调检索结果，如取回网页源码中疑似作者姓名拼音的文本串“G.C. Hu”再转为中国人习惯的“Hu G C”，无法将姓名拼音补全，更无法汉字化，只能为后续人工介入提供比较干净的姓名拼音或机构名称，为补全拼音全称做准备。孙源[11]、何涛等[12]、霍朝光等[13]、盛晓光等[14]、邓启平等[15]开展了以词向量为主要技术的半自动排检实践，该技术思路为：将文本按一定规则数字化为空间坐标系中的点，各点连接构成大小、方向不同的向量，通过各向量差异（夹角、长度）表征文本相似程度，夹角越小、长度越相近表明两点越可能重合，则两点代表的原文本越相似。常用计算方法有编辑距离相似度[16]、余弦相似度等，如将“山东济南250014”通过一定规则数字化为平面直角坐标系中的点对（1，2），按同样规则将“山东济南250100”数字化为（2，3），两点各自与坐标原点连接成两条线段，使用余弦相似度计算两线段夹角余弦值为0.992 3，则夹角接近0°，表明两段文本非常相似。在半自动排检实践的语料库选择、应用方面，昌宁等[17]选用了中国知网、维普、万方和个人主页，刘玮辰等[18]选用引文网络，Waqas等[19]选用了作者个人网页、ResearchGate（RG）和Google Scholar，Zhang等[20]选用了Microsoft Academic Graph、Semantic Scholar和PubMed Knowledge Graph等，Rehs[21]选用了infomap社区。全自动排检理论研究方面，Kim等[22]指出随着ORCID继续推进，通过ORCID链接的标记数据可以使消歧数据总体得到改进，Author-ity2009的ORCID链接标记数据可公开用于验证，但全自动排检实践尚未见报道。

本文通过组合并改进正则表达式、词向量和多源数据等技术手段，将人工排检操作的共性部分如查找不同语料库并比对结果交由程序完成，减少人工介入并缩短操作时间。通过编制带用户界面的应用程序，为外文数据库英译中文作者姓名消歧工作提供更便捷有效的工具。程序对用户计算机操作能力几乎没有要求，工作组中一人导出数据源，清洗工作可由多人多台电脑分批合作完成，仍可保证数据处理质量统一、收割结果有序，甚至可以无人值守。

2 程序设计思路

系统总体目标是编制一个带用户界面的应用程序，以实现外文数据库英译中文作者姓名的汉化消歧。用户单击“开始工作”按钮即开展全自动清洗工作，先自动将英译作者姓名修正为作者本人或所在团队认可的、符合我国及国际标准的汉语拼音形式，然后自动处理英译作者姓名所属机构名称，包括翻译成中文、查询语料库、确定最可能的中文机构名称，再将作者汉语拼音姓名与中文机构名称同时提交语料库进行检索获取可能的作者中文姓名，结果以xls格式写出。程序根据前期调查问卷反馈结果预设可调节参数默认值，同时允许用户自由调整，如同时执行任务进程数量、网页就绪超时秒数、网页解析器失败时重试次数等，在工序顺畅和结果准确之间寻求平衡点，确保程序处理的全自动化，遇少量错误写出详细日志供后续人工处理或导入程序并用调整后的阈值再次自动处理，实现程序处理和人工介导剥离。

系统由待处理数据集、工作层和结果数据集组成。待处理数据集由用户手工检索外文库后手工导出。程序软件会提醒用户根据自身需求在特定外文数据库手工检索并导出待分析记录，记录格式可为html、xls（x）、txt或csv格式中任意一种，允许用户通过单击按钮导入上述格式中任何一种或多种格式组合的一件或多件记录文档。软件自动根据源文件格式读取字段及对应数据、合并记录并写入数据库，并且允许用户通过单击按钮浏览、查找、增加、删除或修改数据库记录。工作层实现自动化操作，将待处理数据集中英译中文作者姓名补充完整并找出对应的中文姓名。这一过程借助多个语料库进行匹配，程序界面允许用户勾选一种或多种语料库，勾选越多则结果越精确。本文用到的语料库包括：多源数据，如中国知网、万方数据知识服务平台、维普网、读秀学术搜索；学术社交网络，如RG、Academia.edu、Mendeley、HumanitiesCommons

及科研之友（Scholar Mate）；网络知识库，如维基百科、百度百科；在线翻译网站，如金山词霸、海词词典。学术社交网络以RG为主。近5年，百度指数[23]和谷歌趋势[24]均表明RG在国内的影响力逐年攀升，谷歌提示中国对其搜索热度稳居全球第一，所以选用以RG为代表的学术社交网络作语料库开展署名作者姓名消歧实践在数据量上有一定保障。自动检索语料库使用多进程工作，使用一种或多种语料库在耗时方面没有显著差别。还要允许用户编辑、测试正则表达式，可根据应用场景分类管理，内置按汉语拼音方案编写的成熟的正则表达式并支持一键导入。主程序开放接口，针对不同语料库编写的网页文档对象自动操作脚本均使用独立插件（exe格式）方式提供，方便今后在不更新主程序的情况下更新旧插件或加入新插件，同时解决“大而全”程序的兼容性与准确性不可兼得问题，主程序根据用户勾选情况自动调用所需运行插件，使该工具兼容常见外文数据库如WoS、Scopus及EI等。工作层的处理结果自动写入结果数据集。

以WoS导出的一条文献为例，作者“Xiang，JW”“Hu，GC”“Zhang，XG”3人共同署名发表论文“*Equivalent linear damping model of nonlinear hydraulic damper for helicopter rotor*”，作者所属机构“Beijing Univ Aeronaut & Astronaut，Dept Aircraft Design & Appl Mech，Beijing 100083，Peoples R China”。首先，利用正则表达式提取文本，将题名、作者名、作者机构一一对应，得3个列表（每位作者1个列表），其中1个列表为“[Equivalent linear damping model of…，Xiang，JW，Beijing Univ Aeronaut & Astronaut…]”。取论文标题，自动在语料库如RG中检索此文，发现有同样题名的论文其3位作者姓名分别为Jinwu Xiang、Guocai Hu和Xiaogu Zhang。由此将作者完整的英文姓名自动替换3个列表中的作者英文名。有时入驻RG作者还会修改变更后单位名称（如工作变动或学校更名），程序不应该修改原文作者机构名称，但如果工作任务同时要求梳理发文作者工作单位变动情况，则可另立字段记录。如有作者未入驻RG，姓名仍未补全，可另寻语料库重试。下一步另设正则表达式，按汉语拼音方案将英译中文作者的姓与名位置调换，“Jinwu Xiang”自动转换为“Xiang Jinwu”。然后，取该作者所在列表第3个元素即机构名称，将分词或全部字符自动提交至翻译网站，得“北京”“大学”“航天”“航空”“航天和航空”“北京航空航天大学”等结果，利用文本相似度计算等方法，按得分最高者取“北京航空航天大学”。随后利用语料库如中国知网、百科网站、搜索引擎等，使用正则表达式、文本相似度检测等算法，使用此机构名称反复、多方位自动模糊检索“Xiang Jinwu”，已知语料库如中国知网支持模糊匹配汉语拼音，即在作者姓名检索入口允许输入汉语拼音并在检索结果输出可能的同音或相近汉字，最终挖掘出作者中文姓名为“向锦武”。如取不回任何结果，则记录详细日志，待后续人工介入，或使用作者合作网络等更多语料库，或使用该作者在学术社交网络标记的新单位等再使用程序自动检索。中英文机构对照表也可事先人工建立从而节省计算时间，程序自动映射时可追加或更新此表（如北京航空航天大学的官方英译已改为“Beihang University”）。实践中还发现有学者误领或冒领作者身份，导致取回错误的姓名全拼，故设计程序时应多方查找取概率最大者。

3 技术方案

整体技术方案如图1所示。以文献标题为抓手，综合利用网页机器人、网络爬虫、正则表达式和短文本相似度检测技术，抓取特定文献标题对应的不同版本作者英文姓名和机构名称，去粗取精、去伪存真，计算对应中文姓名和机构名称。“去粗取精”指将使用翻译网站自动翻译的机构名称如“北京航空的和航天的大学”精简为“北京航空航天大学”；“去伪存真”指将外文文献中作者提供的非官方机构名称（如将“浙江大学”按方言自行翻译为“Zheijing Univ.”[25]）通过程序自动检索中外文语料库或规则表予以纠正，如将“Zheijing Univ，Coll Med，Affiliated Hosp 1”对应为“浙江大学医学院附属第一医院”。考虑到各编程语言主要适用方向及书写便利性，使用易语言设计界面引擎提供人机交互、Python实现主体算法、AutoHotKey承担全局热键脚本任务、JavaScript设计各语言产品联络中间件（如配置文件、日志文件等）。

编制程序关键技术与方法包括网页文档对象操作、短文本相似度检测、正则表达式技术、使用多进程代替多线程作业。

3.1 网页文档对象操作

该技术应用于程序中网页相关操作，如语料库检

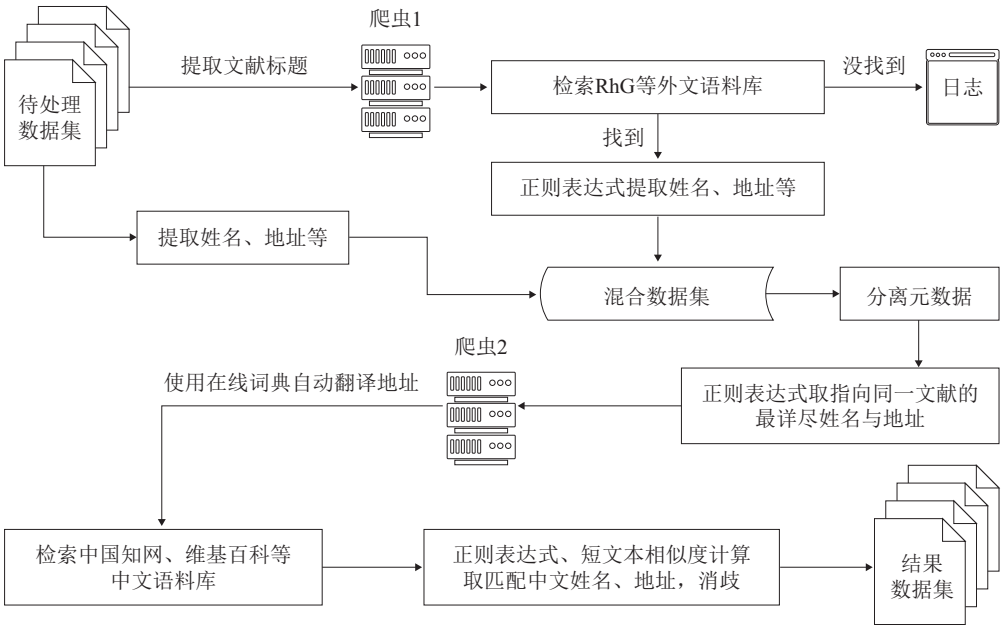

图1 技术方案

索并提取结果、解析元素后配合正则表达式提取格式化文本串等。程序自动操作网页的抓手是元素，故首先从网页源码中分离、识别、定位网页各元素，再通过读写其InnerText属性或Value属性获得或改写对应文本，或通过“click方法”自动点击，实现自动在语料库网页选取检索入口、输入文本、单击按钮检索、等待网页就绪、抓取结果页文本等。常用获取元素方法有“通过元素ID获取”（getElementById）、“通过元素名称获取”（getElementsByName）和“通过元素标签名获取”（getElementsByTagName）等。程序涉及跨域网页文档对象操作对象，即主页面使用IFrame技术嵌套不同域名的独立网页，可使用document.getElementsByTagName取回所有IFrame再按需筛选。程序框架使用一家网站对应一个独立插件思路，遇网站微改版，主程序和插件程序可读取更新后配置文件调整解析语法，甚至无须调整源码并另行编译。

3.2 短文本相似度检测

该技术应用于程序中计算两段文本是否相似及相似程度，用于计算最可能的英译汉机构名称、判断语料库机构名称与原文机构名称是否存在本质变化，并由此推测作者机构变更、语料库被自动补全拼音的作者是否存在误领等。本文采用TF-IDF模型计算短文本相似度[26]。TF-IDF模型计算相似度技术方案为：将两段待计算文本各自分词写入列表；合并两列表，去重，写入集合；转换集合为词典，为各分词建立索引；按分词出现位置分别编码两列表，文本首次实现简单数字化；对应词典，将两列表进行独热编码（OneHot），文本正式数字化、向量化，且将含不同成员数的列表编码统一为相等成员数；计算空间向量余弦夹角。

由于该模型没有考虑特征词位置因素对文本区分度影响[27]，本研究在计算时扩展待检索词提高计算可信度，如计算“机构名称”时使用“省市+邮编+机构名称”组织待检词。为进一步提高计算可信度，可先构造映射规则，如建立机构常用简称与全称对照表，自动将“中科院”先映射为“中国科学院”再参与计算。初期可手工建立映射规则，后期可将程序计算并达到一定阈值的数据写入数据库自动建立。

以计算“中科院水生所”和“中国科学院水生生物研究所”相似度为例，结果为15.81%；将“中科院”按规则映射为“中国科学院”后，计算“中国科学院水生所”与“中国科学院水生生物研究所”相似度为71.71%；另加入省市邮编，“湖北武汉430072中国科学院水生所”与“湖北武汉430072中国科学院水生生物研究所”相似度达78.94%；“湖北武汉430072中国科学院水生生物研究所”与自身相似度为100%。同样代码计算“北京100081中国农业科学院农业经济与发展研究所”与“湖北武汉430072中国科学院水生生物研究所”相似度为40.36%。由此认为，“中国科学院水生

生物研究所”可能是“中国科学院水生所”更详尽地址。当语料库返回样本足够多时，取字符更多、相似度最大者，正确率随之提高。

3.3 正则表达式技术

该技术应用于程序中从杂乱无章的网页源码提取非格式化文本，或用于验证指定文本是否具有特定属性等。非格式化文本相对格式化文本而言，后者明显的特征为使用标记语言书写，提取标记间文本只需使用常规方法如“取文本中间”等。非格式化文本往往无规律可循，如提取网页源码中随机出现的疑似中国邮政编码的文本（前后均无特殊、固定标记），只需将计算式“[0-9]\d{5}(?!\d)”应用于正则算法，表示提取仅6位、开头可为0、连续数字型且其后不可紧跟数字的文本。验证属性时，如提取到“Chiang Kai-shek”，人工排检时可非常方便地判断该文本不符合汉语拼音方案，至少“Ch”后不可跟“iang”，据此原理可编制“声母后可跟韵母”“声母后不可跟韵母”两种算法的正则表达式，用来判断某文本是否为汉语拼音（标准拼音、威妥玛-翟理斯方案拼音或邮政拼音）或非汉语拼音。程序反复使用正则表达式技术，将任务细化，在不同场合编辑不同正则表达式，遇多种需求则组合不同可执行文件实现，提升各表达式功能确切性，尽最大可能保证工序顺畅，提高自动化程度。

3.4 使用多进程代替多线程作业

程序中调用多语料库检索时，如按用户勾选语料库顺序依次操作则耗时较长，无法充分发挥电脑计算潜力。如在可执行文件内部开启多线程任务，首先因为网页文档对象操作时容易混淆元素，其次线程池操作不稳定，在不同电脑表现不同，为获得更好效果，程序使用多进程代替多线程作业。思路为：将实现某相对完整功能的核心算法封装在插件文件中，插件运行后首先查看主程序有无为其分配任务，如有，首先领取任务ID，执行任务，将结果写出带任务ID的xls文档等，待主程序发起合并结果指令，按任务ID顺序合并为整体结果，合并算法由主程序提供，确保数据处理质量统一、收割结果有序。

4 应用效果

程序可在32位和64位Windows 7与Windows 10操作系统下平稳运行，在下载网速平均60Mbps、上传网速平均50Mbps、使用Ping命令访问www.a.shifen.com平均耗时10ms、网络抖动平均2.67ms、丢包平均0.5%网络环境下7×24小时运行未见崩溃且功能确切。

笔者陆续向大连理工大学、河南农业大学、东北师范大学、赣南医学院、青岛农业大学、曲阜师范大学等图书馆情报分析或参考咨询岗位同人分发软件测试版，通过软件内置模块回收用户有效反馈表139份，共处理文献3 685批（共计1 842 841篇），批均处理约500篇，统计结果见表1，满分值均为100%。

表1 软件评分

%

项　目	平均得分
汉化人名自动化率	63.24
汉化地址自动化率	98.17
人工复核人名汉化消歧正确率	71.39
人工复核地址汉化正确率	93.18

汉化人名自动化率为63.24%，指100位作者姓名中，约63位可通过中外文语料库匹配为中文姓名，约37位因学术社交网络无人认领、无文章被中文数据库收录等原因取不回汉化结果，但程序算法依然适用。姓名汉化消歧总体成功率为63.24%×71.39%，即45.15%，尚有很大提升空间，但对于长期从事情报分析、数据治理的从业人员来讲，自动成功处理45%的工作量依然颇具应用价值，96.46%的受访问者有继续使用意愿并希望软件持续更新。

以前文提到的文献“*Equivalent linear damping model of nonlinear hydraulic damper for helicopter rotor*”为例，消歧前后数据对照表如表2所示，人工复核汉化消歧单笔成功率100%。

5 结语

数据预处理是数据治理、情报分析工作中必不可少且非常重要的一环。本文通过编制程序，近乎全自动地实现了以往工作中需要人工投入大量精力的英译中文作者姓名汉化消歧，其优势在于将情报分析等相关从业人员从烦琐的数据清洗工作中部分地解放出来，使

表2 某文献作者姓名、机构名称消歧前后对照表

消歧前		消歧后	
机构名称	Beijing Univ Aeronaut & Astronaut， Dept Aircraft Design & Appl Mech， Beijing 100083，Peoples R China	自动汉化机构名称	北京航空航天大学飞行器设计与应用力学系，北京 100083
作者姓名	Xiang，JW；Hu，GC；Zhang，XG	自动补全姓名全拼且转姓前名后	Xiang Jinwu；Hu Guocai；Zhang Xiaogu
		自动匹配汉字姓名	向锦武；胡国才；张晓谷

其可以将精力更多地用于探索数据背后的逻辑。139份程序试用反馈表显示，有92%用户认为功能确切，81%用户认为执行速度快，97%用户表示工具运行不受第三方软件影响，96%用户表示有继续使用意愿并希望软件持续更新，用户对程序正向认可程度为91%。程序的不足之处在于自动化程度偏低，成功率和精确率尚有待进一步提高。但该工具框架下的语料库具有可扩展性，用户无须更新主程序，只需在程序运行目录添加独立动态链接库（dll）文件即可扩展语料库。工具算法适用于对信息爬取、数据清洗有需求的应用场景，包括但不限于情报分析、关联挖掘、查收查引及自引识别与排除等领域。程序在操作便利性、爬虫稳定性与兼容性、正则表达式通用性及成功率等方面还有优化空间，接下来计划继续提升算法可靠性与架构可扩展性，发现并利用更多中外文语料库，提高成功率。

参考文献

[1] 吴庆晏，巫元琼. 汉语人名英译应重视汉语语言文化特征 [J]. 出版发行研究，2015（4）：68-71.

[2] 陈忠. 跨文化传播的中外认知方式调适与对接 [J]. 山东师范大学学报（社会科学版），2021，66（1）：147-156.

[3] Item 987654321/47834 [EB/OL]. [2022-01-17]. http://ir.lib.ncu.edu.tw/handle/987654321/47834.

[4] A 130-dB CMRR Instrumentation Amplifier With Common-Mode Replication [EB/OL]. [2022-01-09]. https://www.webofscience.com/wos/woscc/full-record/WOS:000733233600001.

[5] Agile big man：The flexible marketing of Yao Ming [EB/OL]. [2022-02-09]. https://www.webofscience.com/wos/woscc/full-record/WOS:000225150800001.

[6] International Olympic Committee. Ming YAO [EB/OL]. [2022-01-17]. https://olympics.com/en/athletes/ming-yao.

[7] 付媛，朱礼军，韩红旗. 姓名消歧方法研究进展 [J]. 情报工程，2016，2（1）：53-58.

[8] 侯长来. 《SCI》中国文献被引著者同名异址的辨识 [J]. 现代情报，2005（3）：162-163.

[9] 何春建. 从WOS地址字段提取二级机构数据的半自动数据清洗方法 [J]. 新世纪图书馆，2017（8）：56-58，70.

[10] 高营. 基于WOS API辅助准确检索学者论文程序的设计与实现 [J]. 河北科技图苑，2020，33（1）：83-87.

[11] 孙源. 基于Word2Vec的SCI地址字段数据清洗方法研究 [J]. 情报杂志，2019，38（2）：195-200.

[12] 何涛，王桂芳，马廷灿. 基于类中心向量的论文作者归属机构自动识别方法研究 [J]. 情报学报，2019，38（7）：716-721.

[13] 霍朝光，司湘云，王婉如. 基于Doc2vec和SVM的作者姓名消歧研究——以PubMed Central为例 [J]. 情报科学，2021，39（7）：91-98，107.

[14] 盛晓光，王颖，钱力，等. 基于图卷积半监督学习的论文作者同名消歧方法研究 [J]. 电子与信息学报，2021，43（12）：3442-3450.

[15] 邓启平，陈卫静，嵇灵，等. 一种基于异质信息网络的学术文献作者重名消歧方法 [J/OL]. 数据分析与知识发现：1-12 [2022-01-22]. http://kns.cnki.net/kcms/detail/10.1478.G2.20211126.1059.002.html.

[16] 侯海东，洪腾龙，徐建良. SCI论文作者自动识别方法研究 [J]. 软件导刊，2018，17（8）：57-60.

[17] 昌宁，窦永香，徐薇. 基于多源数据的科技文献作者同名消歧研究 [J]. 情报科学，2021，39（6）：108-116.

[18] 刘玮辰，史冬波，李江. 基于职业经历和引文网络的华人姓名消歧算法 [J]. 信息资源管理学报，2020，10（6）：82-89，100.

[19] WAQAS H，QADIR A. Completing features for author name disambiguation（AND）：an empirical analysis [J]. Scientometrics，2022，127（2）：1039-1063.

[20] ZHANG L，HUANG Y，YANG J Q，et al. Aggregating large-scale databases for PubMed author name disambiguation [J]. Journal of the American Medical Informatics Association，2021，28（9）：1919-1927.

[21] REHS A. A supervised machine learning approach to author disambiguation in the Web of Science [J]. Journal of Informetrics，2021，15（3）：101166.

[22] KIM J，OWEN-SMITH J. ORCID-linked labeled data for evaluating author name disambiguation at scale [J]. Scientometrics，2021，126（3）：2057-2083.

[23] 百度指数 [EB/OL]. [2022-01-17]. https://index.baidu.com/v2/main/index.html#/trend/researchgate?words=researchgate.

[24] Google. researchgate-探索-Google趋势 [EB/OL]. [2021-07-17]. https://trends.google.com/trends/explore?q=researchgate&geo=JP.

[25] Pharmacological activity of phenolic compounds isolated from Psoralea cortlifolia [EB/OL]. [2022-01-09]. https://www.webofscience.com/wos/woscc/full-record/WOS:000245997000383.

[26] 武永亮，赵书良，李长镜，等. 基于TF-IDF和余弦相似度的文本分类方法 [J]. 中文信息学报，2017，31（5）：138-145.

[27] 薛征. 基于改进TF-IDF的文本信息热点话题发现 [D]. 武汉：武汉邮电科学研究院，2009.

作者简介

朱玉强，男，1978年生，硕士，研究馆员，研究方向：图书情报技术，E-mail：zzyqq@qq.com。
江涛，男，1988年生，博士研究生，馆员，研究方向：教育学、图书馆学、阅读推广、读者服务。
李翼飞，女，1999年生，硕士研究生，研究方向：情报分析与用户研究。

Practice of Author Name Disambiguation in Chinese-English Translation of Foreign Language Database

ZHU YuQiang[1]　JIANG Tao[2]　LI YiFei[1]
(1. Library of Shandong Normal University, Jinan 250014, P. R. China; 2. Library of Hainan Medical University, Haikou 571199, P. R. China)

Abstract: Aiming at the situation where there are multiple records of the names and addresses of the authors in English-to-Chinese translations of foreign language databases pointing to the same person or the same records pointing to different people, the article simulates manual sorting, integrating multi-source data, academic social networks, knowledge encyclopedias, and online translation websites and other corpora. Use the automatic operation of web document objects, regular expressions, short text similarity calculation and other technologies to compile programs to carry out the practice of disambiguation from English to Chinese name and address. The results show that the algorithm architecture is stable and effective, with strong scalability, and the success rate is recognized by practitioners. It provides new ideas and new methods for data preprocessing and cleaning.

Keywords: Name Disambiguation; Address Disambiguation; Data Governance; Foreign Language Database

（收稿日期：2022-01-25）