

# 科研实体名称规范的研究与实践

张建勇 钱力 于倩倩 董智鹏 黄永文 刘建华 郭舒 王峰

**【摘要】**[目的]建立机构规范、作者规范、期刊规范、基金规范,为发现系统、科研实体分析评价等建立数据基础。[方法]以多源异构数据为基础,对数据进行汇聚和融合,形成具有唯一标识符的统一的结构化数据。依据名称规范元数据模型,对科研实体及实体间的关系进行抽取。针对不同的科研实体可获取的文献特征,制定不同的消歧规则集合,结合传统字符串匹配方法和深度学习方法进行文本相似度计算。[结果]形成包含260多万条数据的机构规范库、2300多万条数据的作者规范库、3万多条数据的期刊规范库和200多万条数据的基金规范库。以NSTL机构规范为例,与InCites机构规范进行对比,结果显示所遴选的美、英、中3个国家的6所高校,对标吻合度平均值达到86.8%。[局限]所提出的消歧规则和算法在处理文献特征表达形式多样性方面有待进一步细化和提升;需对具体数据源数据情况进行分析,以选择合适的算法模型。[结论]本研究提出了多源异构数据汇聚融合方法,设计了科研实体消歧规则和算法,能够有效实现名称规范数据库建设的规范性和全面性。

**【关键词】**名称规范;期刊规范;机构规范;基金规范;作者规范

**【作者简介】**张建勇,钱力,中国科学院文献情报中心,中国科学院大学图书情报与档案管理系(北京100190);于倩倩(通讯作者,ORCID:0000-0001-8777-1171),中国科学院文献情报中心,E-mail:yuqianqian@mail.las.ac.cn(北京100190);董智鹏,中国科学院文献情报中心(北京100190);黄永文,中国农业科学院农业信息研究所(北京100081);刘建华,上海科技大学图书馆(上海201210);郭舒,国家互联网应急中心(北京100029);王峰,中国科学院自动化研究所(北京100190)。

**【原文出处】**《数据分析与知识发现》(京),2019.1.27~37

**【基金项目】**本文系国家科技图书文献中心(NSTL)资助项目“名称规范数据库建设”(项目编号:科1817)、中国科学院文献情报中心青年人才领域前沿项目“基于深度学习的名称规范方法研究”(项目编号:G180171001)和中国科学院文献情报中心重点任务专项“科研人员研究方向和研究重点分析”(项目编号:院1643)的研究成果之一。

## 1 引言

名称规范问题由来已久。就海量的科技文献数据而言,由于作者、机构、期刊、基金等科研实体名称在出版形态中具有不同的表达形式,或相同的表达形式可能会指向多个科研实体,给科研实体的准确识别带来极大的困难。消除名称歧义,精确识别与定位科研实体,是信息组织与检索、科技评价、知识服务等的重要基础。随着大数据时代的到来,数字资源成为科技文献资源出版和利用的主流形态。通

过不同渠道获取元数据并集中汇聚成为一种新的资源整合和利用模式<sup>[1]</sup>,而传统的规范著录方式难以满足大规模数据的处理时效要求。这既为解决名称歧义问题带来了困难,也为名称规范建设带来了新的思路和机遇。以多源多类型数据为基础,以大数据环境下的计算能力为手段,从汇聚的大规模科技文献中抽取科研实体及实体间的关系。采取不同方式对不同类型的科研实体进行消歧归并,建立不同来源数据之间的同一和归一关系,从而实现科研实体

名称规范。这为以篇章为主的供给型知识服务向以科研实体为主的计算型知识服务转变奠定更加坚实的数据基础。

## 2 相关研究

本文从名称规范的理论、技术方法、算法、规范控制模型等方面对名称规范数据的研究现状进行探讨和分析。

### 2.1 理论研究

名称规范的建设思想经历了规范控制、访问控制、唯一标识符等阶段<sup>[2-3]</sup>。规范控制的主要思想是指定规范记录中的一个名称作为规范名称,将规范名称与书目记录中的其他名称进行关联。规范控制的优点是建立规范名称,其他名称指向规范名称。缺点是在检索规范名称时,可获取全部书目记录,但检索其他名称时,只能获取包含其他名称的书目记录。访问控制不设置规范名称,将所有的名称(全称、简称、真名、化名、译名等)视为同等的。优点是无须在书目记录中添加规范名称,仅需将书目记录中的名称添加到访问记录中。检索任意名称,均可获取全部的书目记录。缺点是多次检索会加重检索系统的负担。采用唯一标识符进行名称规范控制,一方面是希望更容易地将规范记录和书目记录关联起来,另一方面是希望能将规范数据与语义网关联集成。当前,唯一标识符的重要性越来越被认可。科研实体唯一标识符包括 ORCID、ResearcherID、AuthorID、ISNI、Ringgold、DOI 等。由于还未形成统一的全球解决方案,且在文献数据中应用程度低。据统计,WOS 数据库中 2000 年~2016 年的 1 900 多万篇文献中,具有 ORCID 的文献占比仅为 19%<sup>[4]</sup>。因此,目前利用唯一标识符解决科研实体名称歧义问题更多是愿景和辅助手段。

### 2.2 技术方法研究

名称规范的技术方法根据应用需要和语料基础的不同,可以分为基于文献数据库的实体名称规范、基于 Wikipedia 的实体名称规范、基于社会网络的实体名称规范、基于本体的实体名称规范、基于众包方式的实体名称规范等。基于文献数据库的实体名称规范采用文献特征进行实体消歧。Chavezaragon 等<sup>[5]</sup>采用作者姓名、文章题名和会议地点作为特征向量,构造相似性矩阵,通过欧氏距离计算相似度,对 DBLP 中

的作者进行消歧。很多学者提出利用 Wikipedia 作为名称规范的背景知识,因为它覆盖很多的概念,每篇文章中都包含实体或概念的信息,具有丰富的语义信息且内容实时更新。Fader 等<sup>[6]</sup>利用 Wikipedia 用户贡献的信息和新的消歧模型,组合先验信息和语境信息以提高消歧精度。基于社会网络的实体名称规范方法是利用人物社会关系关联构建社会网络,进而实现相应的名称实体消歧。郎君等<sup>[7]</sup>利用人物社会关系构建潜在社会网络,结合图谱分割算法和模块度指标实现人名检索结果的重名消解。基于本体的实体名称规范,主要利用本体丰富的语义信息,通过本体类和属性的匹配,促进对象实体的消歧和规范。朱小婷<sup>[8]</sup>提出一种人物本体实例树匹配算法框架解决中文人名歧义问题。基于众包方式的实体名称规范利用社会性网络,采用众包的方式共同开发规范控制服务。Phillips<sup>[9]</sup>尝试汇聚人类记忆机构的各类资源,利用社会性网络、采用众包的方式,共同开发规范控制服务。

### 2.3 算法研究

根据所采用的算法类型的不同,可以分为字符串相似度算法、传统机器学习算法、深度学习算法等。字符串相似度算法是名称相似度计算的常用方法,SimPack<sup>[10]</sup>、SecondString<sup>[11]</sup>、SimMetrics<sup>[12]</sup>等开源工具均提供了包括 Levenshtein 距离、L2 距离、Cosine 相似度、Jaccard 相似度等在内的多种相似度计算方法。然而由于其不能有效区分名称相似度很高却非同一实体的情况,因此常与实体特征、规则等结合使用。例如孙海霞等<sup>[13]</sup>考虑机构名称构词特点,提出基于规则和编辑距离的机构名称匹配策略,对作者单位数据进行规范处理。传统机器学习算法包括朴素贝叶斯、支持向量机、决策树、K-means、层次聚类、AP 聚类、LDA 等。Han 等<sup>[14]</sup>利用朴素贝叶斯和支持向量机算法,对文献信息中的作者进行消歧,解决同名(Synonyms)和名称变体(Polysemes)问题。在传统机器学习方法中,词向量通常为离散表示,表示维度大而且词汇之间具有鸿沟,这一问题在深度学习算法中得到改进。深度学习算法包括分布式词向量、序列卷积网络(CNN)、双向 LSTM、循环神经网络(RNN)等。汪沛等<sup>[15]</sup>针对特定领域提出一种结合词向量

(Word2Vec)和图模型的方法实现实体消歧。马晓军等<sup>[16]</sup>提出融合词向量(Skip-gram)和主题模型(LDA)的领域实体消歧方法。当前深度学习算法通常结合传统文本相似度算法共同发挥作用。

## 2.4 规范控制模型研究

规范控制模型包括规范数据的功能需求FRAD、书目框架BIBFRAME、RDA等。FRAD共定义16个实体,其模型基本原则是:书目领域里的实体可通过名称和(或)标识符认知。在编目过程中,这些名称和标识符是创建受控检索点的基础。黄艳芬<sup>[17]</sup>介绍了FRAD概念模型,并分析其与CNMARC在规范控制的内容、检索点、著录实体方面的异同。BIBFRAME规范控制模型定义了4个子类:代理(Agent)、地点(Place)、时间(Temporal)和主题(Topic)。王景侠<sup>[18]</sup>分析BIBFRAME模型从1.0到2.0的演进。RDA将规范控制引入其中,与书目描述一起构成了完整的编目规则。张璇<sup>[19]</sup>阐述了RDA的规范控制思想,并比较分析了RDA规范记录与传统规范记录的主要差异。

## 3 研究思路与框架

科研实体名称规范的建设流程如图1所示。

(1)对不同渠道来源获取的元数据进行装载、清洗、汇聚和融合;

(2)建立名称规范元数据领域模型,根据领域模型对汇聚后的科技文献元数据进行实体抽取和关系抽取;

(3)分析数据特征,根据不同规则和算法对不同类型的科研实体进行归并和消歧;

(4)得到科研实体的规范数据,包括优选名称、其他名称、实体关系、规范关系数据等。如果得到的科研实体规范数据字段较少,如缺少别名或其他描述数据,可通过第三方数据源进行丰富。

### 3.1 多源数据汇聚融合

科技文献的来源渠道可包括数据库商、出版商和服务商等,不同来源的元数据遵循的标准有所差异,数据描述粒度有所不同,数据质量也优劣不等。

(1)对不同来源元数据的质量进行评估,根据评估结果依次对数据源进行装载汇聚。数据质量评估标准包括数据描述字段的准确程度、丰富程度、覆盖的时间范围等。通常认为描述字段越丰富即数据描述粒度越细、越准确(即描述字段错误率越低)、覆盖时间段越广,数据质量越好。数据质量好的数据源



图1 科研实体名称规范的建设流程



数据会被优先装载。

(2)将不同来源的数据统一转换映射为同一文献元数据标准格式,形成具有统一描述标准的数据,有利于后续的数据交互和数据规范。既可选择其中一个数据源所遵循的元数据标准,作为其他数据源转换映射的标准。此种做法节省了制订新的元数据标准的成本,但对已有元数据标准要求较高,需满足其他数据源数据字段的描述需求。也可以通过对不同数据源遵循的元数据标准进行分析,并结合国际主流标准,建立统一文献元数据标准,作为所有数据源转换映射的标准。此种做法增加了不同数据源数据转换映射的灵活性和可靠性,但成本跟前者相比较高。

(3)对同一数据源内的数据进行查重,查重依据为国际通用的唯一标识符和关键字段,例如科技论文的查重依据为DOI、题名+作者等。在数据源内查重后,对不同数据源数据进行查重。

(4)根据查重结果及不同数据源的元数据字段分析结果,对不同数据源的重复数据进行字段级融合,有利于将其中“薄数据”的内容进行数据丰富化,从而形成“厚数据”以便于进行科研实体数据抽取和科研实体名称规范。需要注意的是,在数据格式转换过程中,为每篇文献中的科研实体对象(期刊、论文、作者、作者机构、基金项目等)赋予唯一标识符,作为后续科研实体抽取和回溯依据。

### 3.2 名称规范元数据模型设计

通过实体分析技术,对科技文献中的科研实体进行分析,并参考国际主流规范控制模型如FRAD、BIBFRAME等;建立以论文为中心,连接期刊、机构、基金和作者的名称规范元数据领域模型,如图2所示。

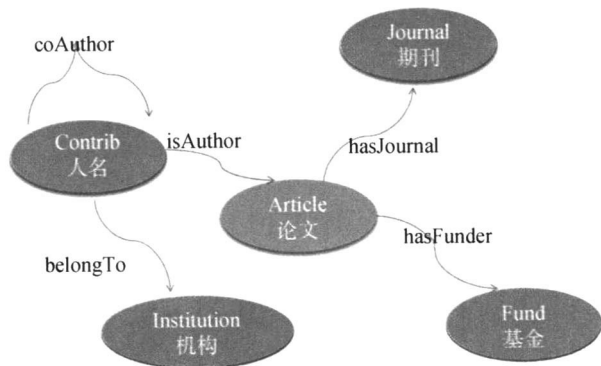


图2 名称规范元数据领域模型

该模型支持实体识别和关联关系描述。以期刊论文科研实体为例,一篇期刊论文发表于一本期刊上,一篇期刊论文由一个或多个作者创作,一个作者属于一个或多个机构,一篇期刊论文由一个或多个基金资助。其中,机构既可为作者所属机构,也可为出版机构和基金资助机构。

(1)作者规范元数据字段包括作者标识符、优选名、其他名、姓、名、姓名前缀、姓名后缀、学位、个人简历、性别、国籍、电子邮箱、电话号码、外部链接、职称职务、专业、研究方向、出生日期等。

(2)机构规范元数据字段包括机构标识符、优选名、其他名、机构简介、电子邮箱、电话号码、研究方向、外部链接、地址信息描述、国家、城市、州或省、邮政编码等。

(3)期刊规范元数据字段包括期刊标识符、优选名、其他名、ISSN、eISSN、出版年、出版者、卷、期、创刊日期等;

(4)基金规范元数据字段包括基金项目标识符、优选名、其他名、基金项目日期、资助金额、资助者、资助说明、摘要、主题词、分类号、关键词等<sup>[20]</sup>。

### 3.3 消歧规则和算法设计

将期刊、机构、基金、作者4类实体数据从汇聚融合后的数据中,按名称规范元数据字段进行分类抽取,形成期刊、机构、基金、作者的不同形式名称、属性以及与其他实体间的关联关系集合。由于同一作者、机构可能发表多篇文章,同一基金可能资助多篇文章,同一期刊会刊载多篇文章,因此抽取的科研实体会在多篇文章中出现,即会存在重复现象。同时,由于同一科研实体在出版形态中可能具有不同的表达形式,因此科研实体名称规范过程实际为查重过程和消歧过程。不同的科研实体所具有的文献特征不同,根据文献特征设计消歧规则并结合相关算法实现名称消歧和归一。

#### (1)机构名称消歧

对机构信息中的机构名进行匹配,由于机构名存在多种变体形式,如Chinese Academy of Sciences和CAS均指中国科学院,而一些从字符串层面来看相似度非常高的机构名却并非指向同一个机构实体,如“Newpark Mall Sears Outlet”以及“Newpark

Mall Gap Outlet”,因此单从字符串本身进行模糊匹配是不够的,需要引入其他信息,构成匹配规则集。可用于机构匹配的文献特征包括机构名称、国家、城市、邮编、地址,制定规则集合:若所在国家、城市相同,且机构名(或机构别名、机构地址或邮编)之间模糊匹配(根据字符串相似度达到一定阈值)则为同一机构。机构名称消歧框架如图3所示。

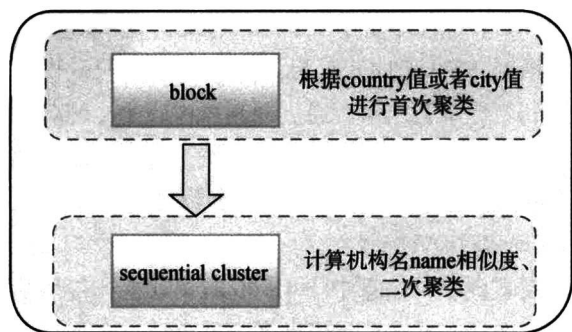


图3 机构名称消歧框架

在机构名称消歧框架中,block模块将所有记录根据country值和city值进行首次聚类,从而减少记录两两匹配的次数,以提高计算效率。sequential cluster<sup>[21]</sup>是一种非常简单的聚类方法,其思想是顺序选择检测样本并比较与其他待检测样本的相似性,有相似类则归类,无相似类则为检测样本建立新类,直到所有样本检测完毕。由于每个待检测样本只比较一次,因此大大减少了比较次数。sequential cluster重点是对如何判断“机构名匹配成功”的实现,对机构名的比较主要分为简称与全称、简称与简称、全称与全称三种情况。将不包含空格且全为大写字母形式的字符串认为是首字母简称。简称与全称的匹配采用一种启发式的简称抽取方法,对全称删除“the”“in”“of”等无意义的介词后,直接提取首字母形成全称的伪简称形式,然后利用编辑距离进行匹配。首字母简称之间的匹配采用“string identity”的方式,直接判断字符串是否相等。全称之间(包括前缀简称的情况)的比较采用Jaro-Winkler以及Jaccard值进行计算。

基于字符串的相似度计算如编辑距离、Jaro-Winkler和基于词条的相似度计算如Jaccard系数将文本视为字符或词串的集合,并未考虑词与词之间的语义。在深度学习技术兴起后,利用分布式表示的词向量进行文本相似度计算成为研究热点。分布

式向量的训练方式更加简洁,所得的词语向量表示的语义可计算性进一步加强。因此,可在前述基础上融合词向量的语义关联优势。基于SimNet框架<sup>[22]</sup>搭建基于词向量的匹配学习算法,如图4所示。

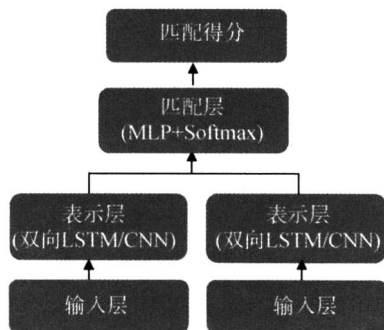


图4 基于SimNet框架的匹配学习算法

SimNet是一种有监督的神经网络语义匹配模型,输入层为词向量,表示层包括序列卷积网络(CNN)、双向LSTM等多种表示技术。匹配层利用文本的表示向量进行匹配度计算。将基于词向量的匹配得分与基于字符串或词条的相似度计算得分进行权值加和,得到最后的匹配得分,得分高的名称为同一机构。

## (2)作者名称消歧

文献信息是大部分现有作者名称自动化消歧方法主要选择的计算特征,通过选择合适的特征、相似度计算方式以及特征合并方法,往往能取得很好的消歧效果。Zehnalova等<sup>[23]</sup>利用DBLP等数据库中作者历年所发表文献中的关键词的使用程度随时间的演变关系,发现研究者在一段时期内也会有比较稳定的研究方向和研究主题。Newman<sup>[24-26]</sup>对Los Alamos、Medline、SPIRES、NCSTRL等数据库中的文章数据进行作者合著网络研究,发现合作者网络具有“小世界”特性以及聚集性,研究者在很长一段时期会有比较稳定的合作人员。因而可选取的特征包括:作者名称、文献标题、出版物名称、关键词、分类号、机构名、邮箱、合作者关系等。制定规则集合:

### ①弱规则

Rule0:待消歧作者具有相同作者名简称。

### ②强规则集合

Rule1:待消歧作者具有相同的唯一标识类型和唯一标识内容,或者待消歧作者具有相同的邮箱;

Rule2:待消歧作者具有至少一个相同的合作者

简称且其合作者机构相似;

Rule3:待消歧作者具有相似的机构;

Rule4:待消歧作者具有相似的研究领域。

弱规则用于解决同一个作者实体存在多种不同的名称表达形式即名称变体问题,强规则集合用于解决不同作者实体同名冲突问题。基于规则集进行作者名消歧的方法先利用弱规则解决名称变体问题,从而将作者名歧义问题转化为同名问题,再利用强规则集进行同名消歧。其中强规则集按照可信度从高到低排序,在对待消歧作者进行同名消歧时,依次进行判断,若满足其中一条,则判断为同一作者实体。

对于弱规则中的作者名称,科技文献中一般都会保留姓氏的全称形式,而名字部分形式多样化。可以按照一定规则(如保留姓氏,名字提取首字母)将不同人名形式归一到统一规范的作者名简称中,再利用作者名简称进行第一次聚类,合并不同形式的作者名称。例如,“Abascal J. L. F.”与“Abascal Jose L. F.”的作者名简称均为“Abascal JLF”,因而在第一次聚类中被初步判断为可能为同一作者实体。

国外期刊文献作者的署名主要4种情况:

①大部分国外期刊将作者的姓名完整列出,名在前,姓在后,如Carmencita Pilapil;

②有些期刊会将作者的名缩写在前,姓在后,如C Pilapil;

③个别期刊也会将作者姓名完整列出,但姓和名之间用逗号隔开,逗号前是姓,逗号后不论是全拼还是缩写,均为名,如Pilapil, Carmencita;

④姓在前,名字缩写在后。如Pilapil C。

对于强规则结合中的唯一标识符或邮箱特征,根据唯一标识符或邮箱名(不考虑大小写)是否相同进一步对作者聚类。对于合作者关系,根据合作者集中的合作者名简称是否相同且合作者的机构是否相似进一步对作者进行聚类。对于文献标题、出版物名称、关键词特征,将标题、出版物名称、关键词中出现的所有词项(除去停用词)合并为一个主题词集合,利用Jaccard相似度算法进行文献主题相似度计算,判断待消歧作者的研究领域的重合程度。对于机构名特征,根据待消歧作者的机构名是否相似进一步对作者聚类。由于机构名也存在歧义问题,即

机构名称形式不统一、机构变迁,因而需要有针对性地解决。可以利用已经消歧的机构数据,通过作者所在机构的机构唯一标识,在机构规范库中判断机构是否相同或相似。

### (3)期刊名称消歧

期刊名称消歧过程是利用文献的特征属性对待消歧期刊进行聚类的过程。可选取的文献特征包括期刊名称、ISSN、期刊DOI、论文DOI、年、卷、期、论文题名等。制定规则集合如下:

①待消歧期刊具有相同的ISSN或具有相同的期刊DOI;

②待消歧期刊具有相似期刊名称;

③待消歧期刊具有相同的论文DOI;

④待消歧期刊具有年、卷、期,且论文题名相似。

根据ISSN或DOI匹配,若完全匹配,可获得准确的期刊名称聚类结果。通过期刊名称或论文题名的相似度匹配,可采用编辑距离的方法进行计算,若名称完全匹配,也可获得准确的期刊名称聚类结果。

### (4)基金名称消歧

在基金项目信息中,可获取的文献特征包括基金项目标识符、基金项目资助者、基金项目名称。基金项目标识符如NSFC、NSF等的基金号,通常具有唯一性,但也有少量基金号在全球范围内并不唯一。因此,首先通过基金项目的唯一标识符和基金项目资助者判断关联,其次根据基金名称的相似度判断。相似度计算方法可采用编辑距离的方法。

## 4 应用实践与结果分析

NSTL在“十三五”发展规划中提出优化国家科技文献资源保障体系,拓展元数据资源采集方式。从单纯加工扩展到加工、采集、交换、赠予、呈缴和购买等多渠道获取元数据资源,构建中国科技信息资源的“大”元数据体系<sup>[27]</sup>。NSTL从多渠道获取的数据中包含有大量科研实体,这些科研实体面临缺少规范、缺少唯一标识、缺少关联等各种问题。在NSTL的服务转型期,迫切需要以NSTL数据集为基础,长期、稳定地推进科研实体名称规范库的建设。为科研实体名称建立唯一标识体系和关联,提高数据的规范质量。与国际相关规范库链接,从而提升NSTL数据的整体价值。对NSTL知识发现系统、信息服务



平台等不同层次的服务形成核心支撑能力,提高NSTL信息服务的水平,促进我国文献信息事业的发展和理论研究水平的提高。

#### 4.1 应用实践

##### (1)多来源数据集成融合

鉴于NSTL数据集合包含本地加工数据(历史库、加工库)、外部加工数据(中国图书进出口公司加工)、Web of Science数据、出版社数据(如剑桥大学出版社CUP, Wiley, 牛津大学出版社OUP等)、开放资源数据等多来源异构数据,且已达到千万级数据体量。建立一个多来源数据集成管理流程尤为重要。不同来源异构数据经过解析、清洗、转换、集成、融合等流程,形成具有唯一标识符的结构化数据。

##### ①数据格式转换

根据数据规模、数据质量、数据接收时间等,数据装载入库顺序依次为加工库、历史库、WOS、CUP、中国图书进出口公司加工数据等。由于不同数据源遵循的数据标准有所不同,如加工库、历史库遵循文献数据加工规范<sup>[28]</sup>;WOS数据遵循自定义标准<sup>[29]</sup>;CUP遵循JATS标准<sup>[30]</sup>等,因此需要对数据进行同构处理。目前,多来源数据采用NSTL制定的统一文献元数据标准<sup>[31]</sup>作为数据映射与数据转换依据,通过XML数据格式转换成统一格式的元数据进行存储,同时保留原始格式数据以便溯源。

##### ②篇级数据查重

各数据源之间的数据存在一定的交集,需要对数据进行查重处理。数据查重规则分为两类:一类使用唯一标识符,如DOI、ISSN等;另一类使用基于规则建立的唯一标识,如出版年、题名、第一作者、首页等组成唯一标识。以期刊论文查重为例,WOS源内查重规则为UID,加工库源内查重规则为[期刊Id+出版年+起页+题名]或[第一作者+出版年+起页+题名]。两者源间查重规则顺序为:DOI;ISSN+卷+期+起页;期刊名称+卷+期+起页;ISSN+卷+期+题名;期刊名称+卷+期+题名。

经过查重处理的数据将建立各数据源之间的关联关系。建立关联关系的数据通过基于元素或元素集的融合规则和各来源数据映射表进行数据的内容融合。可进行部分字段新增,如将WOS数据的其他

形式缩写补充NSTL数据;也可将整个层级替换,如将WOS的作者与机构数据替换现有CUP数据。数据查重融合后,作为科研实体名称规范的数据基础。当前主要采用WOS数据及与WOS数据查重融合的NSTL加工库、历史库、CUP数据,进行名称规范建设。

##### (2)科研实体信息抽取和消歧

根据名称规范元数据设计,从集成融合的文献中抽取期刊信息、机构信息、基金信息和作者信息。由于科技文献出版所限,描述字段有限,可针对性抽取。期刊信息能抽取到的字段包括期刊标识符、刊名、其他名、ISSN、出版者、年、卷、期、单篇文献唯一标识符等。机构信息能抽取到的字段包括机构标识符、机构名、国家、城市、地址、与作者关联的内部序号、单篇文献唯一标识符等。基金信息能抽取到的字段包括基金项目标识符、基金项目资助者、资助说明、单篇文献唯一标识符等。作者信息能抽取到的字段包括作者标识符、姓名、姓、名、所在机构唯一标识符、单篇文献唯一标识符等。此处科研实体标识符更多是系统赋予的内部标识符。对于未从科技文献中抽取到的字段,可通过第三方数据源后续进行丰富补充。

从以上描述可知,每类科研实体抽取信息都包含科研实体标识符。对科研实体的合并消歧,实际是对相应科研实体标识符的聚类。并且,在抽取的信息中都包含单篇文献唯一标识符,不同的科研实体可通过同一单篇文献唯一标识符进行关联。同时,在机构和作者间,还可以通过统一文献元数据标准所设置的与作者关联的内部序号,实现机构与作者的准确关联。科研实体消歧规则如表1所示。

表1 科研实体消歧规则

科研实体	消歧规则
期刊	ISSN   期刊DOI   期刊名称   论文DOI   年+卷+期+论文题名
机构	机构名称+国家+城市
基金	基金号+基金项目资助者   基金项目名称
作者	ORCID   ResearcherID   邮箱   作者名称+已规范的机构ID

在实际消歧过程中,需要结合实际情况对数据进行处理。以WOS机构名称消歧为例,WOS数据中提供机构名称的pref字段,即增强机构名称。既对名称规范提供助力,也对名称规范造成一定障碍。助力在于对于Peking Univ这样的机构,可以直接得到增强

机构名称,即机构全称为Peking University。障碍在于对于Univ Calif Berkeley、Univ Calif Davis这样的机构,都是对应两个增强机构名称,分别为University of California Berkeley和University of California System,以及University of California Davis和University of California System。由此可知,如果将University of California System认为是机构全称,则会将Univ Calif Berkeley和Univ Calif Davis视为一个机构,而实际上此为两个机构。

鉴于此,NSTL机构名称消歧分为三步:使用机构名、国家、城市进行一次聚类,并给予聚类形成的簇集合进行编号;簇集合中WOS增强信息与机构名标准化后采用编辑距离进行相似度匹配,选取相似度最高的增强信息作为簇集合的规范名称;利用簇集合的规范名称二次聚类,选取相似度最高的作为最优规范名称,其他阈值大于等于0.8的为机构其他名称。

基金消歧因各来源数据基金信息较少,所以在消歧中采用基金项目中的特征信息丰富本地数据进行处理。利用基金项目标识符和基金项目资助者对数据进行一次聚类。将缺少基金项目标识符的数据,利用表达模式规则库从基金表达说明中提取基

金项目号和采集丰富的基金项目号二次聚类,形成更多的关联关系,并选取最优规范名称。

### (3)科研实体规范关系揭示

通过实体识别与规范关系揭示,提取文献数据中的科研实体和实体之间的关联关系,挖掘规范科研实体价值,形成数据深度聚合,提高信息服务支撑能力。NSTL科研实体名称规范服务系统,是对NSTL文献数据进行深度清洗、实体抽取、数据丰富、数据规范后,形成的高质量规范化的服务系统。通过该系统对科研实体规范关系进行揭示。揭示关系以论文为桥梁,将规范后的期刊、机构、基金、作者这4类实体规范元数据与之间的规范关系抽取,动态形成各种以规范名称为核心发布的热点科研动态、科研合作网络等。

以科研机构 and 科研项目为例,科研机构可揭示以机构规范名称关联的各种数据分析,如机构规范名称、机构其他名称、机构历年发文趋势、机构历年项目趋势、机构期刊发文TOP10、机构合作网络、机构主题词云、机构关键词共现网络,并以可视化展示,如图5所示。



图5 机构规范详情



科研项目可揭示以基金规范名称关联的各种数据分析,如科研项目名称、项目关联发布的项目成果和项目的详细信息。相关项目成果可通过 DOI 等唯一标识链接到原始文献。项目还与机构和期刊进行关联揭示资助机构国家、地区等信息。

#### 4.2 结果分析

以机构数据抽样分析为例,因目前数据集成融合中主要存在的数据来源为 NSTL 本地加工数据与 WOS 核心集数据,所以采用 InCites 数据对机构名称规范数据进行抽样对比分析。InCites 中包含 WOS 对机构数据进行规范的结果。

对比同一个机构在 InCites 和规范库中的论文集合(机构 A 在 InCites 中的论文集合为 PA,机构 A 在名称规范库中关联的论文集合为 PB)。将 PA 集合与 PB 集合进行吻合度计算。优先遴选 University of California Berkeley(美国)、Stanford University(美国)、University of Cambridge(英国)、University of Oxford(英国)、Tsinghua University(中国)、Peking University(中国)6 所高校作为抽样,从抽样对比结果中可看出,与 InCites 的机构规范对标吻合度均在 80% 以上,对比结果如表 2 所示。

针对对比结果,分析 PA 中未出现 PB 中出现的集合。篇数较少逐项排查,篇数较多则抽样排查,发现有部分机构是因为论文中通信地址是二级机构地址,在 WOS 增强信息中提供了一级机构名称;而 PB 中未出现 PA 中出现的集合,主要现象为通讯作者机

构未解析纳入计算,WOS 中提供的增强信息利用不准确,二级机构信息未纳入计算。所以针对机构规范还可调整机构算法,以进一步提高规范精度。

#### 5 结语

在大数据环境下,数据存储能力、计算能力的提升,为多源异构的科技文献数据汇聚融合,进而形成统一描述标准的科技大数据提供了可能。解决数据孤岛问题,打通数据壁垒是数据深层次挖掘和利用的必然趋势。科技大数据的形成及数据细粒度描述的发展倾向,为名称规范建设提供新的思路和方法。以多源数据为基础,依据文献特征形成科研实体消歧规则集,并结合字符串相似度计算和深度学习方法实现消歧。同时在名称规范建设中引入科研实体唯一标识符的思想,通过唯一标识符将不同科研实体名称关联,并建立同一科研实体的规范关系,提高数据关联的准确性和数据交互的便捷性。

本文梳理了面向多源数据的科研实体名称规范建设流程,包括多源异构数据汇聚融合、依据名称规范元数据抽取科研实体和实体间的关系、采用不同的规则和算法对不同类型的科研实体进行消歧归一等。并以 NSTL 名称规范数据的建设为例进行分析和说明。以期科研实体名称规范相关系统的建设提供思考和借鉴。利用多源数据建设名称规范,能够保证数据建设的全面性和权威性,为发现系统的建设、科研实体的检索定位以及分析评价打下良好的数据基础。

表 2 抽样对比结果

对比机构	A 中有 B 中无	A 中无 B 中有	A、B 吻合	吻合度
University of California, Berkeley	17861	43	144520	89.01%
Stanford University	20156	46	176216	85.91%
University of Cambridge	32673	842	156696	82.74%
University of Oxford	38978	23	166552	81.03%
Tsinghua University	13125	1188	107740	89.14%
Peking University	7313	12	97931	93.05%

#### 参考文献:

[1]程颖.资源发现系统元数据的问题与思考[J].图书情报工作,2015,59(9):104-110,126.  
[2]Niu J. Evolving Landscape in Name Authority Control[J].

Cataloging & Classification Quarterly, 2013, 51(4): 404-419.  
[3]胡小菁.规范控制:从名称选择到实体管理[J].数字图书馆论坛,2018(1):2-7.  
[4]Youtie J, Carley S, Porter A L, et al. Tracking Researchers and Their Outputs: New Insights from ORCID[J]. Scientomet-

rics, 2017, 113(1): 437-453.

[5]Chavezraron A, Cruz J F R, Reyesgalaviz O F, et al. An Algorithm to Tackle the Name Authority Control Problem Using Semantic Information[C]//Proceedings of the 2009 Mexican International Conference on Computer Science. IEEE, 2010:176-179.

[6]Fader A, Soderland S, Etzioni O. Scaling Wikipedia-based Named Entity Disambiguation to Arbitrary Web Text[C]//Proceedings of the 2009 IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy. 2009.

[7]郎君,秦兵,宋巍,等.基于社会网络的人名检索结果重名消解[J].计算机学报,2009,32(7):1365-1374.

[8]朱小婷.基于本体的中文人名消歧[D].上海:华东师范大学,2013.

[9]Phillips L B. The Temple and the Bazaar: Wikipedia as a Platform for Open Authority in Museums[J]. The Museum Journal, 2013, 56(2): 219-235.

[10]Kiefer C. SimPack Project Page[EB/OL].[2018-11-11]. <https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/index.html>.

[11]SecondString Project Page[EB/OL].[2018-11-11]. <http://secondstring.sourceforge.net/>.

[12]UK Sheffield University. SimMetrics[EB/OL].[2018-11-11].<http://sourceforge.net/projects/simmetrics/>.

[13]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.

[14]Han H, Giles C L, Zha H, et al. Two Supervised Learning Approaches for Name Disambiguation in Author Citations[C]//Proceedings of the 4th ACM/IEEE Joint Conference on Digital Libraries. 2004: 296-305.

[15]汪沛,线岩团,郭剑毅,等.一种结合词向量和图模型的特定领域实体消歧方法[J].智能系统学报,2016,11(3):366-374.

[16]马晓军,郭剑毅,王红斌,等.融合词向量和主题模型的领域实体消歧[J].模式识别与人工智能,2017,30(12):1130-1137.

[17]黄艳芬.FRAD概念模型与CNMARC规范控制[J].图

书情报工作,2009,53(12):125-128.

[18]王景侠.书目框架(BIBFRAME)模型演进分析及启示[J].数字图书馆论坛,2016(10):67-72.

[19]张璇.RDA对规范控制思想的阐释及实践革新探析[J].图书馆研究与工作,2017(10):31-37.

[20]名称规范元数据标准[EB/OL].[2018-11-11].<http://spec.nstl.gov.cn>.

[21]Kainulainen J J. Clustering Algorithms: Basics and Visualization[EB/OL].[2018-11-11].<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.1490>.

[22]Baidu NLP[EB/OL].[2018-11-11].[https://www.sohu.com/a/149089880\\_465975](https://www.sohu.com/a/149089880_465975).

[23]Zehnalova S, Horak Z, Kudelka M, et al. Evolution of Author's Topic in Authorship Network[C]//Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining(ASONAM 2012). IEEE Computer Society, 2012.

[24]Newman M E J. Scientific Collaboration Networks. II. Shortest Paths, Weighted Networks, and Centrality[J].Physical Review E, 2001, 64: 016132.

[25]Newman M E J. Scientific Collaboration Networks. I. Network Construction and Fundamental Results[J]. Physical Review E, 2001, 64: 016131.

[26]Newman M E J. The Structure of Scientific Collaboration Networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2000, 98(2): 404-409.

[27]彭以祺,吴波尔,沈仲祺.国家科技图书文献中心“十三五”发展规划[J].数字图书馆论坛,2016(11):12-20.

[28]张建勇,曾燕.文献数据库数据加工规范[M].北京:知识产权出版社,2009.

[29]Web of Science Core Collection Schema[EB/OL].[2018-10-22]. <http://ipscience-help.thomsonreuters.com/wosWebServicesExpanded/wosSchemaWoSCCGroup/wosSchema.html>.

[30]Journal Archiving and Interchange Tag Set Versions[EB/OL].[2018-10-28].<https://jats.nlm.nih.gov/archiving/versions.html>.

[31]沈仲祺,张建勇.文献元数据设计指南和实践[M].北京:科学技术文献出版社,2017.