基于外部 ID 的中文实体对齐分析*

——以中国科学院院士 Wikidata 数据子集为例

王瑞云　贾君枝

摘　要　本文尝试解决中文学者命名实体与外部知识库的实体对齐短缺的问题。通过 SPARQL 语义查询抽取维基数据子图——中国科学院院士的知识图谱子图,初步构建国内知识库的中文院士实体与 Wikidata 实体的对齐以及与外部 ID 对应的知识库的实体对齐。对院士实体的三个数量型特征对齐的外部 ID 个数(ids)、不同语种的 Wikipedia 站点个数(sites)、实体的全部陈述个数(states)与目标分类(有无 VIAF 实体对齐)的相关分析发现,目标分类与 ids 特征正向相关最强,直接 VIAF 实体对齐只存在 ids 高区的院士,占比偏低。因此,提出利用 LC、ISNI 等外部 ID,应用 VIAF 对重要来源库的重定向功能,构建间接的 VIAF 实体对齐的方法。本文为中文知识库进行外部实体对齐提供了可行的初步方案,提出的实验方法显著地提高了较小 ids 值(1-7)的院士拥有 VIAF 实体对齐的个数,最终通过实体对齐的 VIAF 信息集成增加了院士实体的 ids 数量,丰富了中文学者与外部知识库的实体对齐信息。图 4。表 5。参考文献 19。

关键词　外部 ID　Wikidata　VIAF　命名实体对齐

Analysis of Named Entity Alignment Based on External-ID

——Taking Data Subset of Wikidata for Academician of Chinese Academy of Sciences as an Example

Wang Ruiyun　Jia Junzhi

Abstract: This paper attempts to solve the shortage of Chinese named entity alignment with external knowledge base. By selecting Wikidata subgraph about academician of Chinese Academy of Sciences, initially mapping the Chinese academician entity in knowledge base in our country with the entity in the Wikidata, and mapping with the entity of external-ID in other knowledge bases, the paper analyses the distribution of three quantitative features which include quantity of external ID (ids), quantity of Wikipedia sites for different languages (sites), quantity of all entity statement (states), and then present visual graph of the relation between features and target classification VIAF ID. The paper initially finds that target classification is most positive correlated with the ids and the ratio of the academician with direct VIAF ID only existing in ids high-level area is relatively low. Therefore, this paper proposes that it is necessary to further exploit LC, ISNI and other external-ID, take advantage of the redirection function of VIAF to important source database, and find out a way to build indirect VIAF entity alignment. This paper not only provides an initial and useful method for external entity alignment of Chinese knowledge base, but also improves markedly the number of academicians with small ids (1-7) in having VIAF entity alignment with an experiment method. What's more, through integration with VIAF, we can add the quantity of ids of academician entity and enrich the entity alignment information of Chinese scholar with external knowledge base. 4 figs. 5 tabs. 19 refs.

Keywords: External-ID; Wikidata; VIAF; Named Entity Alignment

---

* 本文系国家社会科学基金项目“中文学术领域命名实体的知识图谱构建研究”(项目编号:18BTQ072)的研究成果之一。

DOI: 10.13666/j.cnki.jnlc.2020.0211

命名实体识别是人工智能自然语言处理领域的一项重要的基础研究,主要任务是识别出文本中的人名、地名和机构名[1,2]。而中文命名实体的识别比其他语言更为困难,近年有研究者利用最新的深度学习技术解决命名实体识别问题,采用深度学习框架和神经网络结构实现了命名实体特征的自动学习[3-5]。深度学习虽然能够高效地学习多粒度语言单元间复杂的语义关联,但对自然语言的深度理解亟需复杂知识的支持,需要实现从字面意思到言外之意的跃迁。

命名实体对齐是把自然语言文本中识别出来的实体与现有知识库的实体进行匹配,包括两个方面:一是识别不同来源同一名称的实体是否为实际语义上的同一个实体,二是识别有哪些不同的名称实体实际指代同一实体。2014 年开始在线运行的国际图书馆合作项目——虚拟国际规范档(VIAF),在多个国家图书馆之间实现了上述两方面的实体对齐工作。中国国家图书馆(NLC)和中国高等教育文献保障系统(CALIS)管理中心开展的名称规范档项目,已经实现在规范档记录内部对同一实体不同名称的聚集,但是还缺乏不同知识库间同名规范记录的实体对齐,且聚簇深层实体对齐需要领域知识的支持。基于实体属性关系构建知识图谱是当前实体对齐中的重要解决方案,已有一些重要成果投入应用,如维基百科(Wikipedia)、维基数据(Wikidata)和关联开放数据云(LOD)等。国外研究者利用 Wikipedia 的海量数据构建自动抽取的语料库,并利用其中的多语言知识映射来提高跨语言命名实体识别效果[6,7]。实体对齐技术包括基于文本相似度的对齐和基于关联数据语义的对齐,基于关联数据语义的实体对齐包括实体模式对齐和实例对齐两个层次。贾君枝等学者基于词表映射方法研究知识库间实体类和属性的实体模式层的对齐[8,9]。刘晓娟等构造基于 DBPedia 的中英文命名实体词典,利用关联数据的语义关系对候选命名实体进行对齐[10]。Qingheng Zhang 等学者基于知识图谱嵌入技术研究跨知识图谱间的实体对齐,改进传统知识图谱嵌入框架,在典型数据集 Dbpedia、Wikidata 和 Yago 上进行实验验证[11]。贾君枝和 Lingbing Guo 等学者分别利用 LOD 中跨知识库的 OWL:Sameas 关系研究跨知识库的实体对齐[12,13]。上述跨知识库间实体对齐研究都采用了 100,000 对以上的实体对进行对齐,但是研究的实体大部分是英文命名实体;而本文实体对齐研究的实体限制为中文学术命名实体这一实体子集,旨在建立和完善中文知识库实体与外部知识库实体的对齐。

对中文命名实体的跨知识库实体对齐的研究还处于起步阶段,很多中文权威网站(如中国科学院官网、国家自然科学基金委员会官网等)处于“孤岛”的状态,关联数据发布应用还不充分,大量的实体语义知识掩藏在海量的异构文本中。这不利于机器理解和自动抽取,不方便外部知识库链入对齐。同时,中文网站也没有主动与外部知识库进行实体对齐,外文知识库对中文知识库的实体对齐占比极低,这些不足成为中文命名实体对齐研究的弱项。本文利用 RDF 自动查询工具 SPARQL 从 Wikidata 抽取中国科学院院士(个人实体的子类——顶层学者)的语义信息,在院士实例的细粒度层次,利用 Wikidata 现有的中文学者的关系链,初步构建中文院士实体到 Wikidata 的跨知识库的实体对齐;然后利用 Wikidata 给出实体的外部 ID,扩充构建中文实体与外部 ID 所属的知识库间的实体对齐;最后重点利用实体的外部 ID 特征及其相关特征的关系,尤其是院士与 VIAF 中实体对齐的关系,构建中文命名实体院士到 VIAF 的实体对齐,以期为中文学术领域命名实体知识库的构建以及中文命名实体跨知识库实体对齐提供参考方法。

1 基于码属性的实体语义表示

码属性引用了关系数据库的相关概念,是能唯一地标识一个实体的属性集。很多知识库为内部的每个实体分配一个永久 ID 作为该实体的码属性,可以在知识库内部唯一地标识这个实体,该 ID 称为实体的内部 ID。在跨知识库实体对齐时,

该 ID 称为引用知识库实体的外部 ID。

1.1 外部 ID 的起源与发展

外部 ID 对应知识库中的内部 ID。内部 ID 是知识库为每个实例分配的内部唯一标识符,相当于数据库中能唯一区别实体的码,在知识库内部能唯一地确定一个实体。内部 ID 与知识库的机构域名、路径一起定义该实体的统一资源识别符 URI。例如,华罗庚院士在 Wikidata 的内部 ID 为 Q590111,维基数据永久域名为 https://www.wikidata.org,存放实体类资源的路径 entity,构成维基数据华罗庚 URI 为 https://www.wikidata.org/entity/Q590111。其他知识库如 VIAF 等也采用同样的 URI 定义实体资源。在此基础上,跨知识库实体对齐时,外部用户和合作伙伴可以根据知识库内部 ID 自动构建引用该实例资源的 URI,方便外部知识库的实体对齐和实体引用。

外部 ID(维基数据中称 External-ID)是语义网关联数据技术的应用要求和发展成果。关联数据最高级标准要求,在 RDF 中尽可能提供到可用外部知识库实体 ID 的链接[14]。外部 ID 是内部 ID 经过实体对齐函数的映射结果,是内部 ID 代表的实体在外部知识库中对应实体的 ID。实际的同一实体在对齐的两个知识库之间形成对应的 ID 对(内部 ID、外部 ID),相当于 RDF 三元组(内部 ID TheSameAS 外部 ID)。如果知识库在每个实体 RDF 陈述三元组中给出该实体在外部知识库中的 ID(本文中的外部 ID 就是此种用法),以及外部知识库固定的域名路径,就可以自动构造出外部知识库中实体的 URI。外部 ID 是知识库的构建者(或委托的第三方)根据知识库内部海量知识和底层的本体逻辑关系,实时运行实体对齐算法后提供的实体实例级对齐运算结果,实体外部 ID 是知识库实体对齐的最重要信息。

基于内部 ID 和外部 ID 的属性应用促进了知识库间的实体对齐和互操作。首先,知识库中基于 ID 的实例 URI 定义格式与传统 URI 定义相比,更加持续和稳定,同时使用固定域名路径和 ID 的组合 URI 命名方式,更方便外部知识库进行引用;其次,知识库可以更好地参加知识库间协作项目,进行实时的知识库间实体对齐计算;最后,有利于采用 RDF 语法描述并发布实体实例与外部知识库外部 ID 的实例对齐结果。目前提供外部 ID 实体对齐的知识库主要有 VIAF、Wikidata 和 LOD 等。

1.2 国内外知识库实体语义表示的比较

关于本文研究的中国科学院院士数据的来源,国内本地知识库使用的是中国科学院官网、百度百科的中国科学院院士分类子集。这两个国内知识库在院士资源的权威性、正确性和更新的及时性方面与国外知识库相比有相当大的优势,但是在基于语义网的知识表示方面与国外知识库 Wikidata 存在一定的差距,具体如表 1 所示。

表 1 国内外三个知识库语义表示的比较

	中国科学院官网	百度百科	Wikidata
权威性	最高	较高(源于官网+自有信息+用户编辑贡献)	较高(源于集成的 Wikipedia 多种语言站点+用户编辑贡献)
语言	简体中文	简体中文	多语言,大部分包含简体中文标签
用户应用	登录,3—4 层目录超链接,不提供用户服务的工具	1. 按实体名称检索 2. 中科院院士列表超链接	1. 按名称检索 2. 按 ID 检索 3. SPARQL 检索
知识表示	HTML	HTML	以 RDF 为基础的多种数据表示
结构化	无结构化	半结构化	结构化

续表

	中国科学院官网	百度百科	Wikidata
实体识别	文本,无实体识别	多类型的实体识别,内部 ID	基于本体层次类的实体识别,内部全局实体 ID
实体对齐	无内部 ID,无外部 ID	无对齐的外部实体 ID	有实体对齐的外部 ID

中国科学院官网是院士信息的权威来源机构,在权威性、及时性方面有其他两个来源数据不可比拟的优势,是其他来源数据的参考来源和最终的验证标准。但是其知识表示没有采用先进的语义网技术,有着“数据孤岛”的局限。中国科学院官网中的院士信息访问是逐层超链接的:首页->院士->学部->院士列表->单个院士信息[15]。单个院士页面包括 URI、照片图像、个人简介文本三部分。个人简介描述是全文本,没有进行实体识别表示和实体对齐。

百度百科知识库整体上采用维基技术,提供两种查询应用方式:一是查询“中国科学院”机构实体,在内容网页找到分学部的院士姓名列表并连接到个人实体资源;二是直接按院士个人姓名查询个人主页。每个个人实体资源以个人姓名作为内部 ID,增加域名路径构建。如百度百科华罗庚 URI 定义为 https://baike.baidu.com/item/华罗庚;实体资源包含实体标识和链接的超文本内容[16],如下所示:

华罗庚(1910.11.12—1985.6.12),出生于江苏常州金坛区,祖籍江苏丹阳。数学家,中国科学院院士,美国国家科学院外籍院士,第三世界科学院院士,联邦德国巴伐利亚科学院院士。中国第一至第六届全国人大常委会委员。

学术著作:《堆垒素数论》《优选学》《高等数学引论》《从杨辉三角谈起》

……

其中江苏常州、金坛区、丹阳、数学家、中国科学院、美国国家科学院、第三世界科学院院士、第六届全国人大常委会等文本标注了百度百科知识库的实体实例;《堆垒素数论》《优选学》《高等数学引论》《从杨辉三角谈起》等著作实体可以利用关联数据技术与图书馆机构知识库中的著作实体关联。上述标注出的每个实体都关联到百度知识库内部的一个实体条目,如“https://baike.baidu.com/item/堆垒素数论”是堆垒素数论的实体关联的实体条目[17]。但这些关联都是知识库内部的自链,没有到外部知识库的外部 ID 的跨知识库对齐,甚至没有列出到百度自己的“百度学术”知识库的跨库实体对齐。而国内的 CALIS 知识库中引用了到外部百度百科的实体 URI 对齐,国外的 Wikidata 中甚至有少量到百度学术的外部 ID 引用的实体对齐。由于得不到百度的回应,外部知识库到百度知识库的外部 ID 引用数量一直处于很低的水平。

Wikidata 是基于维基百科的开放关联数据项目,其中所有实体都分配唯一持久的以 Q 开头的数字序列 ID 为内部 ID。每个实体内部包含大量的 RDF 三元组(h 实体,关系,t 实体),其中 h 实体和 t 实体都采用 Wikidata 内部实体 ID,构建 Wikidata 的内部实体关系。Wikidata 适应关联开放数据的发展要求,在 Wikipedia 的基础上增加了外部 ID 映射,描述知识库内各实体与外部实体的对齐关系,尤其是到著名的规范控制知识库的实体对齐。双向促进作用下,这些外部著名知识库也随后给出自己知识库的实体与 Wikidata 的 ID 对齐关系,极大地促进了跨知识库的实体对齐。

2 基于外部 ID 的实体对齐研究

2.1 Wikidata 中国科学院院士数据选取

中国科学院院士是中国科学研究水平最高的学术团体,在国内外有很高的知名度和引用率,国内外多个知识库都包含大量的中国科学院院

士实体。但国内知识库的院士实体表示缺少语义关系,无法进行语义推理。学术成果仓库 VIAF 中的实体包含大量的学术实体对齐关系,但机构实体数量和语义关系偏少,且又缺乏基于语义的查询工具,所以本文选取 Wikidata 知识库抽取研究数据。

Wikidata 有基于本体的实体语义层次框架,其中"中国科学院"机构语义信息丰富,机构下有更细粒度实体为中国科学院的学部。学部通过部分属性关联到中国科学院,并与每个院士通过院士的会员属性关联,形成语义关系链(实体—会员关系—学部会员,学部会员—部分关系—中国科学院院士),根据该关系链构建 Wikidata 的中国科学院院士子图。Wikidata 为方便用户使用提供了 SPARQL 查询应用,可以根据用户的具体需求设计图模式的查询方案,定制用户感兴趣的个性化数据。所以本文选择 Wikidata 的 SPARQL 应用,研究中国科学院院士的外部 ID 信息,尤其是学术领域知识库 VIAF 的外部 ID。

经过多次调试,本文精心设计对中国科学院院士的 Wikidata 的 SPARQL 查询方案,包括匹配模式和结果显示两部分。匹配模式为三个条件的交集:a.实体类型属于个人实例;b.实体的会员属性为某个学术机构;c.该学术机构是中国科学院的学科分部(而不是分支机构)。结果选取实体 ID、实体标签、学部 ID、学部标签、出生时间、去世时间、外部 ID 个数、链接维基百科站点个数、陈述个数、外部 VIAF ID、博士生导师、受教育机构等属性和关系。具体的 SPARQL 查询代码如下,查询结果得到了 2235 行 12 列的数据[18]。

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIXwdt: <http://www.wikidata.org/prop/direct/>
PREFIXwikibase: <http://wikiba.se/ontology#>
PREFIXbd: <http://www.bigdata.com/rdf#>
SELECT ?item ?itemLabel ?mem ?memLabel ?
birthLabel ?deathLabel ?ids ?sites ?states ?VI-
AFIDLabel ?docLabel ?edu
WHERE
{?item wdt: P31 wd: Q5 .
?item wdt: P463 ?mem.
?mem wdt: P749 *wd: Q530471.
?item wikibase: identifiers ?ids.
?item wikibase: sitelinks ?sites.
?item wikibase: statements ?states.
optional {?item wdt: P569 ?birth.}
optional {?item wdt: P570 ?death.}
optional {?item wdt: P214 ?VIAFID.}
optional {?item wdt: P184 ?doc.}
optional {?item wdt: P69 ?edu.}
SERVICEwikibase: label { bd: serviceParam
wikibase: language " [AUTO_LANGUAGE],
en" }
}order by ?mem desc(?VIAFID) ?itemLabel
```

以上获取的数据与中国科学院官网数据进行比较验证,发现自动抽取的 Wikidata 存在以下细小的错误:院士的学部分类比官网多了中国工程院院士和中国哲学社会科学院院士两个分类。这两个分类在历史上与中国科学院院士有很深的渊源,但将其与现在的中国工程院院士和中国社会科学院官网的院士信息对照,召回率 R 分别为 23.8%和 23.7%,召回率太低。而中国科学院其他六个学部数据的召回率 R 为 88.1%,没有覆盖的院士基本上是 2015 年和 2017 年当选,这些院士的信息没有在 Wikidata 中及时更新院士属性。但这也基本覆盖了中国科学院院士,基本上可以接受该数据的可信性。本文对照中国科学院官网的所有院士,选取 Wikidata 查询到的中国科学院六个学部的院士实例为研究对象数据集,投影消除与外部实体对齐关系较小的列,去除冗余后得到本文后续的研究数据为 1307 行 10 列的 CSV 数据集。

通过维基数据的实体标签和中文知识库的实体姓名的文本相似性比较,本文初始阶段构建中文知识库的 1307 位院士到 Wikidata 知识库的外部 ID 对齐,初步对齐的院士占总体人数的 88.1%。下面进一步以 Wikidata 为中介,分析 Wikidata 的关系子图,建立到其他知识库的外部 ID 对齐,再进一步重点构建到 VIAF 的外部 ID 对齐。

2.2 Wikidata 中国科学院院士数据特征

上述获取的院士信息包括三部分：第一部分是基本信息，包括实体编码、名称标签、学部编码、学部标签、出生日期、去世日期；第二部分包括三个数量型特征：院士实体对齐的外部 ID 个数（ids）、不同语种的 Wikipedia 站点个数（sites）和实体的全部陈述个数（states），此部分数据是本文研究的重要内容，其定义见下文 2.3.1 节；第三部分是实体对齐到外部知识库 VIAF 的外部 ID，它本身是外部 ID 的一个特殊实例，是本文最重要的目标分类变量数据。

VIAF 知识库中的实体资源实际上是一个实体对齐的聚簇信息，包括了众多到国家图书馆的规范名称的关联映射，如图 1 所示的著名院士华罗庚的 VIAF 图书馆实体 ID 聚簇信息[19]包含了 20 个到国家图书馆的外部 ID 实体对齐的关联。VIAF 与 Wikidata 实体的外部 ID 交集有 10 个，仅出现 VIAF 而不在 Wikidata 的有 12 个，只在 Wikidata 而不在 VIAF 聚簇的有 14 个，两个知识库的实体对齐具有很大的互补性。由于 VIAF 包含了大量的到图书馆知识库的实体对齐关联，本文后续在 Wikidata 众多外部 ID 个数（ids）中选取 VIAF ID 作为目标分类研究院士信息的外部关联实体对齐，不足之处在于本文初始的检索结果中有 VIAF 关联的只有 226 位比较知名的资深院士，占院士总数量的 17.3%，占比偏低。后文将详细分析有 VIAF 链接的院士的 Wikidata 特征，以及如何通过重定向推理计算构建间接的 VIAF 实体对齐。

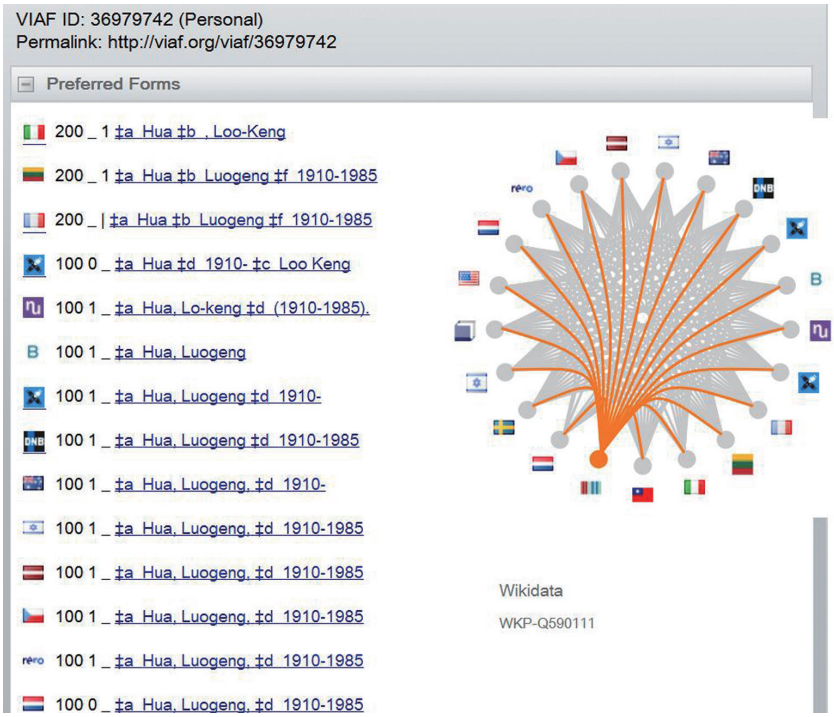

图 1　华罗庚院士的 VIAF 关联数据

2.3 三个数量型特征分析

2.3.1 特征的定义与解释

在全体数据集上，不考虑院士的基本属性，分析三个数量型特征与目标分类的相关关系，这三个数量型特征的定义和描述如下：

ids：Wikidata 中的 External-ID 的个数。每个 External-ID 将实体对齐到一个外部知识库的 ID，相当于一个实体基于 owl：theSameAs 的实体对齐

RDF 陈述。ids 值越大,从 Wikidata 到外部知识库同一实体对齐的个数(外部关联数)越大。例如已故院士华罗庚外链知识库 ID 个数有 26 个,包括:VIAF ID,美国国会图书馆 LC ID,德国 GND ID,中国的 Calis ID、NLC ID、ISNI,数学谱系项目编码和 Freebase 数据库编码等 26 个规范名称知识库 ID。上述外部 ID 都是国家级的规范名称数据库的编码,具有相当高的规范性和权威性,用户可信度高。在跨知识库的关联中,有着较大的链入和链出数,可以通过这些外部 ID,间接计算得到与其他知识库的实体对齐。

sites:不同语种的 Wikipedia 站点个数,每个站点链接该实体条目不同语言的资源网页,例如上述华罗庚院士的维基站点有 14 个,包括德语、英语、法语、西班牙语、日语、简体中文等不同语种的华罗庚院士的维基百科实体关联。

states:实体的全部陈述的个数。每一个陈述可以用 RDF 三元组表示实体的一个属性或关系的事实。例如上述华罗庚院士实体的全部陈述个数为 65 个,包括如表 2 所示陈述。表 2 中每行表示一个陈述,每个元素都是一个 URI,第二列"主体"表示华罗庚院士的 Wikidata 编码,第三列"谓词"是 Wikidata 的属性或关系,第四列"值"大部分都是 URI,表示与主体有关系的维基数据内部实体 ID 或外部知识库实体 ID。为便于理解,第三、四列给出标签代替 Wikidata 属性或实体 ID 编码。陈述个数越多,实体的属性关系语义越丰富。

表 2　华罗庚的陈述 RDF 三元组

序号	主体	谓词	值
1	Q590111 (华罗庚)	Instance of	个人
2	Q590111	member of	中国科学院数学物理学部
3	Q590111	educated at	清华大学
4	Q590111	educated at	剑桥大学
5 *	Q590111	VIAF ID	36979742
…	…	…	…

注:带 * 的第 5 行第 4 列,RDF 陈述的值 36979742,即维基数据 Q590111 实体的外部 ID

上述第 5 条 RDF 陈述描述一条谓词为 VIAF ID 的三元组,等价于两个知识库 Wikidata 和 VIAF 的实体对齐三元组:

wd:Q590111 owl:theSameIAs VIAF:36979742

内部 ID 冒号前面的字符串为实体对齐两个知识库域名 URI 中固定部分域名路径的简写前缀,前缀为知识库域名的简短略写,用户可以参考维基数据 SPARQL 界面提供的标准前缀。

2.3.2　特征的分布

对数据集中所有实体的上述三个数量型特征进行统计分析,结果可视化数据箱图如图 2 所示,并在图 3 中给出三个数量特征概率分布的密度图。

从图 2 和图 3 可以看出,这三个特征都是左偏分布的,均值都远大于中位值,均值受到极端大值影响比较大。左偏最严重的为 ids,最小值、下四分位甚至与中位数重合,都为极端小值 0;上线异常值分界点(上四分位+1.5 * 四分位间距 = 1+1.5 * 1)为 3,大于 3 的异常大的数据个数 232,占 17.8%,是具有高研究价值的数据。

从图 3 可以看出,上述三个特征值大致服从指数分布,只有右端尾部少数节点的三个特征值较大,大多数节点的三个特征值都很小,分别为 0、1、9。陈述个数 states 值 9 远小于百度百科中的内容,甚至下四分位以下的院士属性个数少于中科院官网可以提取的陈述数。产生上述问题的原因在于我国当前与国外知识库的协作还处于初始阶段,国外知识库的中文个人名称只能间接来自境外网站,而不是来自中国本地的知识库,所以很多院士的陈述数属性偏少,而跨知识库的实体对齐的 ids 更少。但这种情况会随着中文命名实体知识图谱构建的完善和中国对外交流的发展而逐步改善。

图 2 中三个特征值处于异常大值的实体是拥有丰富外部 ID 的质量高的实体,可以为建立中文学者命名实体对齐提供有参考价值的数据。表 3 给出了按照外部 id 个数排序前 15 名的中国科

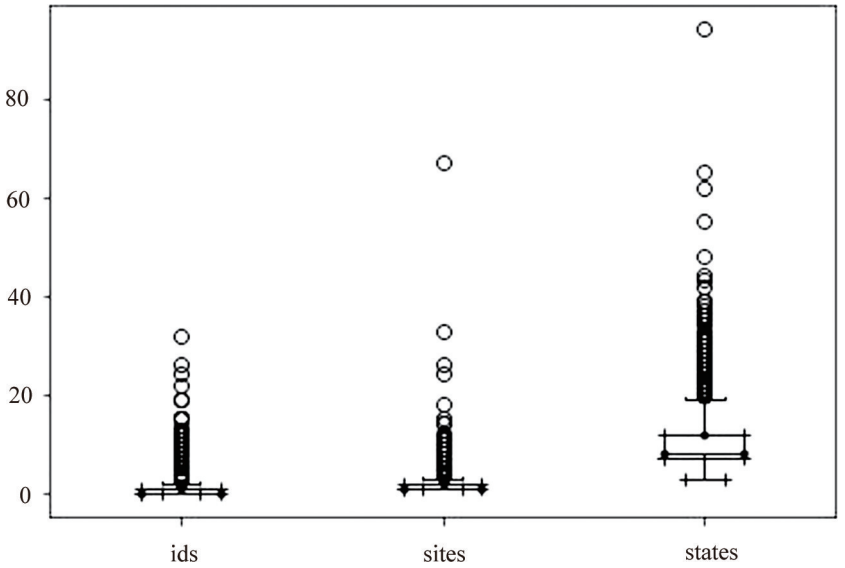

图 2　总体样本三个数量特征

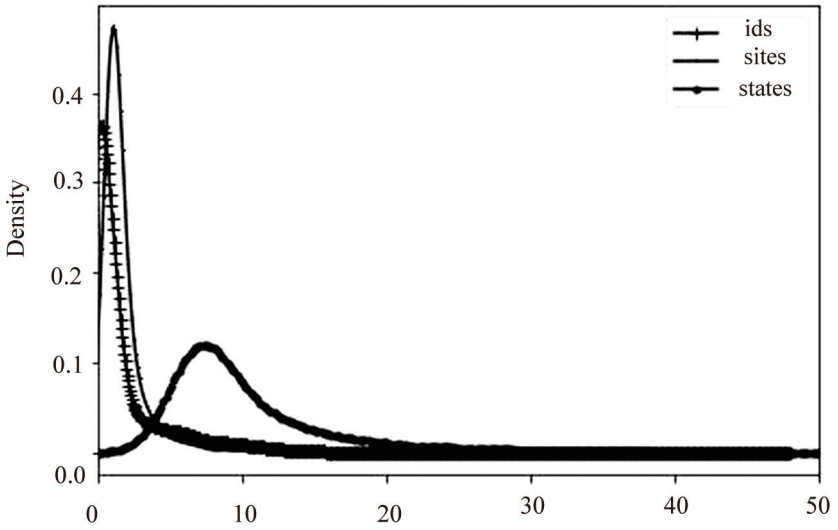

图 3　外部关联 id 个数、维基站点个数、陈述个数分布密度

学院院士名单(包括官网中的已故院士,Wikidata 中这些院士所属学部为生前官网所属的学部),可以看出前 15 位院士的三个特征值都同时处于异常大值部分,都拥有 VIAF ID。根据基本属性分析,这些院士都是学部委员级的资深院士,其中 10 名是已故院士,且排名大都靠前;一共只有 5 名(33%)健在院士,排名第一的杨振宁为外籍院士。上述分析说明 Wikidata 中著名资深院士的数据质量远高于新当选的年青院士,中国科学院官网及国内第三方机构应尽可能提供关联数据格式,发布实时更新的本地数据,以便更好地发布及时权威的本地信息,推动与国外知识库的对齐和互操作。

表3　External-Id 个数排名前 15 位的中国科学院院士信息

ID	姓名	学部 ID	出生日期	去世日期	ids	sites	states	VIAF ID
Q181369	杨振宁	Q46139495	1922/10/1		32	67	94	222662938
Q590111	华罗庚	Q46139495	1910/11/12	1985/6/12	26	14	65	36979742
Q71874	梁思成	Q46151167	1901/4/20	1972/1/9	24	14	62	82960963
Q333500	钱学森	Q46139495	1911/12/11	2009/10/31	22	33	55	79422641
Q926125	吴文俊	Q46139495	1919/5/12	2017/5/7	19	11	42	46730008
Q15916846	郝柏林	Q46139495	1934/6/26	2018/3/7	15	3	31	9004075
Q534717	吴征镒	Q46146709	1916/6/13	2013/6/20	15	11	37	8306343
Q699428	杨钟健	Q46148731	1897/6/1	1979/1/15	15	15	39	307170346
Q323594	苏步青	Q46139495	1902/9/23	2003/3/17	15	5	39	15868981
Q10884754	任咏华	Q46145115	1963/2/10		13	4	29	76632538
Q2601552	王元	Q46139495	1930/4/29		13	4	27	61608445
Q704575	裴文中	Q46148731	1904/12/3	1982/9/18	13	12	35	53026642
Q707276	李四光	Q46148731	1889/10/26	1971/4/29	13	9	44	39740861
Q4843133	白春礼	Q46145115	1953/9/26		12	5	36	78620387
Q552928	路甬祥	Q46151167	1942/4/28		12	7	35	78445914

2.4　基于 VIAF ID 的目标分类特征分析

2.4.1　基于 VIAF ID 分类的确定

由于 VIAF 是全球著名图书馆的名称规范档的关联知识库，是学术领域比较著名的知识库，它对著名学者及其著作和代表作品进行名称规范化关联。VIAF 本身关联了全球 200 多个图书馆机构的规范档，所以与 VIAF 的同一实体对齐就意味着与多个著名图书馆的规范档的实体对齐。所以本研究选取 Wikidata 中的 VIAF ID 同一实体对齐为目标分类，根据每个实体包含的 VIAF ID值是否为空分为两类：如果 VIAF ID 有 ID 值，分类值为 1，表示该实体可以关联到图书馆规范名称库，获取图书馆的作品信息；否则分类值为 0，表示当前没有直接的 VIAF ID 的关联。

2.4.2　目标分类与数值特征的关系

本研究中目标分类与上述 2.3 部分三个数量特征的关系，分别可视化为散点图，如图 4 所示。图 4 中左图为 sites 和 ids 平面分布散点图，右图为 states 和 ids 散点图。这两个散点图都体现出有 VIAF 关联的分类样本居于分布图的右上方，没有 VIAF 关联的处于左下方，中间有少量的交叉区域。

从图 4 的两个散点图可以看出，y 轴 ids 对目标分类的正向相关作用明显比 x 轴上的维基站点或陈述个数两个特征要强，y 轴上部数据更可能有 VIAF 同一实体链出。ids 个数与其 VIAF 链出的关系，如表 4 所示。

从表 4 可以看出，ids 值为 0 的区域，没有到 VIAF 直接链接，也不具备推导计算的基础，这部分实体的对齐问题本文暂时不考虑；ids 在 8 及以上，全部具有直接 VIAF 链接，不需要推导计算，可以直接利用 VIAF 的信息。中间 ids 值为 1—7 的分区出现两个分类的混合，具体分析如下：ids 值为 1 的分区，直接 VIAF 链接数极小，这符合 2.2 中 VIAF 知识库多个关联聚簇特征；ids 值为 2—3 的分区，直接 VIAF 处于中间状态，具备聚簇关联基础，预估有很大可能推导出 VIAF 间接

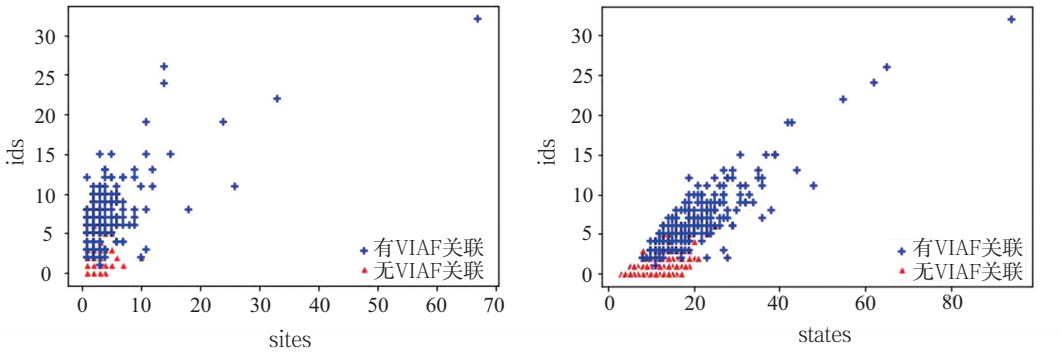

图 4　三个特征向量与 VIAF 关联分类的分布关系

表 4　ids 与直接 VIAF 分类个数的关系

ids 值分区	样本院士数	直接 VIAF 个数	直接 VIAF 占比
0	704	0	0%
1	315	1	0.3%
2—3	100	40	40%
4—7	80	74	92.5%
大于等于 8	77	77	100%
汇总	1307	226	17.3%

关联;ids 值为 4 到 7 的区间,大多数有直接 VIAF 链出,预估全部能直接或间接推导出 VIAF 关联。

2.4.3　基于重定向实现与 VIAF 链接

VIAF 知识库为了照顾来源地用户对本地图书馆的使用习惯,提供了利用来源知识库 ID 重定向到相应的 VIAF 聚簇的功能。如:Wikidata 中,Q10897600 刘敦桢院士只有一个 External-ID 为 ISNI:0000 0004 5301 9688,通过 VIAF 数据源重定向构造到 VIAF 的实体对齐 URI:http://viaf.org/viaf/sourceID/ISNI|0000000453019688,重定向自动进入 VIAF 实体 URI:https://viaf.org/viaf/19680591/,不仅能得到 VIAF ID 实体对齐,还可以得到 VIAF 中的 17 个来源知识库的外部 ID。这些关联链接包括 VIAF 到 Wikidata 的反向实体对齐,另外还有到其他著名国家图书馆的实体对齐链接:美国 LC-n81053811、德国 DNB-135553784、法国 BNF-11913144、日本 NDL-00536020、挪威 NTA-162467427 等的个人实体对齐关联,同时还可以得到 VIAF 从上述 17 个来源知识库动态实时收集的不同语言出版的 35 本著作的信息。

在上述 2.2.2 节中发现 ids 值在 1—7 的部分,有可能通过上述功能计算构建间接的 VIAF 实体对齐。为了进一步探索外部 ID 情况,再次应用 SPARQL 查询其 External-ID 的具体情况。第二次检索除了上述的 VIAF ID 外,又增加了两个外部 ID,分别为美国国会图书馆 LC、国际标准名称索引 ISNI。利用这些外部 ID,就可以使用重定向构建间接的 VIAF 实体对齐,其结果见表 5。

表 5　利用重定向构建间接 VIAF 的统计

ids	样本院士数	直接 VIAF 个数	直接 VIAF 占比%	构建间接 VIAF 个数	VIAF 占比%
1	316	1	0.3%	32	10.5%
2	55	13	23.6%	26	70.9%
3	45	27	60%	11	84.4%
4—7	80	74	92.5%	6	100%

表 5 将 ids 值分为 4 种情况,可以看出直接 VIAF 占比和总体 VIAF 占比都是 ids 值的增函数;4 个 VIAF 值的分区都通过重定向构建了间接的 VIAF 关联,VIAF 的个数有了显著的增加,4 个分区的 VIAF 间接关联符合 2.4.2 节的预估:

ids 在 4 以上的院士,100%都有 VIAF 的实体对齐。

ids 在 2、3 部分,VIAF 的实体对齐的占比也有很大提高,分别提高到 70.9%和 84.4%。

ids 值为 1 部分,利用 LC 或 ISNI 重定向构建了 32 个间接 VIAF,占 10.5%,另外有 252 个实体的外部 ID 是到中国名人录 CV 的实体对齐,但是 CV 知识库中的实体没有与外部知识库的关联。该部分外部 ID 需要注意的还有到中国国家图书馆和百度学术知识库的外部实体 ID 对齐 14 个,这两个是本地中文数据库,具有大量的本地权威规范信息,遗憾的是现阶段本地知识库没有提供到外部知识库的实体 ID 对齐。还有一些 ids 值为 1,提供到专业知识库的实体对齐,虽然这些专业知识库不再提供进一步的实体对齐信息,但在建立中文实体对齐中也是非常有用的信息。

ids 值为 0 的部分,没有外部 ID 可供借鉴,但很多院士实体提供了到 Wikipedia 多语言站点的外部链接;602 个院士只有一个到中文 Wikipedia 站点链接,该站点由国外机构创建,本文获取的可利用信息较少;拥有两种语言站点链接的有 95 项,可以利用其中的英语 Wikipedia 站点链接构建到其相关站点 Dbpedia 的外部 ID 链接。

总体来说,在本文院士数据集中,首先直接包含 VIAF 外部 ID 的有 226 个,在总体数据集中占比 17.3%;再通过 VIAF 知识库对来源知识库重定位功能,另外计算出 75 个间接到 VIAF 的实体对齐关系,为这些实体增加了与 VIAF 的实体对齐。无论是直接还是重定向计算得到与间接 VIAF 的实体对齐,都可以构建院士实体到这些国家图书馆的命名实体对齐关系,丰富与院士实体关联的书目作品信息。

3 结果与讨论

通过本文第 1 和 2 部分的分析,我们得到中国科学院院士的实体对齐总体可分三个层次:

第一层次,进行国内知识库间的实体对齐。在百度百科和中国科学院的院士知识库进行高准确且覆盖率为 100%的院士实例层的一对一的实体对齐,其首要工作是中国科学院官方机构(或委托的第三方)将每位院士资源的 URI 持续固定化,方便其他知识库引用;然后对院士的文本信息进行结构化和语义化转变,方便其他知识库的实体对齐;最后,百度百科需要利用自身资源完善与百度学术的显式实体对齐,并与中国科学院官方网站进行一对一的实例层实体对齐。国内其他知识库对其收录的每一位院士构建到百度百科或中国科学院官方院士知识库的实体对齐和实体引用。

第二层次,国内知识库可以利用 2.1 节语义查询方法,构建与 Wikidada 的一对一的实体对齐。本文中该实体对齐的准确率很高,依赖 Wikidata 的院士实体的 member of 属性值的准确率,属性值准确率可以通过 Wikidata 的用户编辑和专家委员会审核得到保证;院士实体对齐的覆盖率为 88.9%,可以接受。覆盖率的提高可以通过 Wikidata 中中国科学院院士信息的及时收录更新和 member of 属性的准确赋值实现。本层次的初步实体对齐以很高的准确率和较高的召回率实现了中文知识库命名实体与第一个国外知识库的实体对齐;更深层次可以 Wikidata 知识库为中介,进行第三层次的与国外知识库的对齐。

第三层次,利用本文重定向的外部 ID 对齐方法,在第二层次构建中文知识库院士实体与 Wikidata 实体的一对一对齐的基础上,为 603 位院士增加了两个以上跨知识库实体对齐,更重要的是可以为 301 位院士构建到 VIAF 的一对一的实体对齐,占总体院士数量的 23%,这些实体对齐结果显著地提高了中文知识库的实体对齐成效。

然而当前中国科学院院士与 VIAF 对齐的覆盖率还远低于国外同类实体,原因如下:一是国内图书馆还没有与 VIAF 协商完成正式加入 VIAF,而且有些中国科学院院士没有著作被国外图书馆收录,这超出了本文研究的范围;另一个原因,也是更重要的原因,虽然 Wikidata 中显式的外部 ID 可以为跨知识库的实体对齐提供重要的准确信息,但如同 2.4 节那些重定向推导得到的新的 VIAF 实体对齐,说明 Wikidtata 的外部 ID 并不完备,还缺失了某些隐含的 VIAF 外部 ID,那么就可能存在某些院士实体缺失了更多隐含的外部 ID 所表示的实体对齐关系。分析这些缺失隐藏的

VIAF 实体对齐信息的院士,发现他们基本上在本文表 4 中 ids 值为 3、2、1 的分区,论文下一步将对这部分院士按照 ids 值从大到小的顺序,采用基于标签、出生日期和更多语义属性对齐的方法构建院士与 VIAF 中的个人实体对齐。

解决第三层次第二方面的问题是本文下一步的研究方向,着力提高国内知识库院士与 VIAF 实体对齐的覆盖率,为更多院士实体构建到 VIAF 的一对一的实体对齐;另外还需要逐步扩大研究学者实体的范围到拥有高级奖项和荣誉等具有较高知名度的学者及其相关的机构和成果。

参考文献

1 刘浏,王东波.命名实体识别研究综述[J].情报学报, 2018(3): 329-340.

2 Alberto Boschetti, Luca Massaron.数据科学导论-python 语言实现[M]. 于俊伟,靳小波,译. 北京:机械工业出版社,2018:153-157.

3 柯昊,等. 数据缺失时基于 BP 神经网络的作者重名辨识研究[J]. 情报学报, 2018(6): 44-53.

4 张海楠,等.基于深度神经网络的中文命名实体识别[J]. 中文信息学报,2017(4): 28-35.

5 王蕾,等.基于神经网络的片段级中文命名实体识别[J]. 中文信息学报,2018(3):84-90.

6 Nothman J, et al. Learning multilingual named entity recognition from Wikipedia[J]. Artificial Intelligence, 2013, 194: 151-175.

7 Ni J, Florian R. Improving multilingual named entity recognition with Wikipedia entity type mapping[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2016: 1275-1284.

8 赵洁,贾君枝.基于映射方法对比的受控词表互操作质量影响因素研究[J].情报理论与实践,2018(4):117-122

9 贾君枝, 李捷佳. 基于关联数据的语义互操作研究[J]. 情报理论与实践, 2017 (8): 135-138..

10 刘晓娟,等. 基于关联数据的命名实体识别[J].情报学报, 2019(2): 191-200.

11 Qingheng Zhang, et al. Multi-view knowledge graph embedding for entity alignment [R]. In: International Joint Conference on Artificial Intelligence (IJCAI), 2019.

12 贾君枝,李晓. 关联数据中 owl:sameAs 网络分析[J]. 数据分析与知识发现,2017(10): 77-84.

13 Lingbing Guo, et al. Learning to exploit long-term relational dependencies in knowledge graphs [R]. In: International Conference on Machine Learning (ICML), 2019 .

14 W3C.Linked data glossary[EB/OL].[2019-03-10].https://www.w3.org/TR/ld-glossary/#x5-star-linked-open-data.

15 中国科学院. 中国科学院-院士信息-数学物理学部 - 王元[EB/OL]. [2019-03-15]. http://casad. cas. cn/sourcedb _ ad _ cas/zw2/ysxx/sxwlxb/200906/ t20090624_1792000.html.

16 百度百科.华罗庚-百度百科[EB/OL].[2019-03-11].https://baike.baidu.com/item/华罗庚.

17 百度百科. 堆垒素数论-百度百科[EB/OL].[2019-03-11].https://baike.baidu.com/item/堆垒素数论.

18 Wikidata. Wikidata query service [EB/OL].[2019-05-11].https://query.wikidata.org.

19 OCLC.VIAF-36979742[EB/OL].[2019-05-14].http://www.viaf.org/viaf/36979742.

(王瑞云 讲师 山西大学经济与管理学院, 贾君枝 教授 中国人民大学信息资源管理学院)

收稿日期:2019-07-25