基于本体的科研机构标签体系研究*

郭红梅[1] 曾建勋[1]

1 中国科学技术信息研究所，北京，100038

摘 要：科研机构是科研管理与评价的重要对象，是资源组织和关联的重要单元，随着网络技术发展和中国对科技的重视，科研成果呈指数增长，科研活动范围广泛，形式多样，如何从海量的科研成果中挖掘科研机构的特征，从复杂的社交网络中识别关联对象一直深受科学家关注。文中构建了科研机构本体模型，基于本体模型定义和描述科研机构的属性和关系，构建科研机构的特征标签库，并重点选取表征科研特征的性质职能、学科领域、行业类别、关联机构等来论述标签化过程。对科研机构属性和特征的标签化有助于快速了解机构全貌，促进机构在文献检索、关联聚类、分面导航、统计分析和定标比超等方面的应用，辅助科研规划和管理决策，快速识别未来的合作伙伴和竞争对手。

关键词：科研机构；机构本体；机构画像

Profiling Attribution on Label of Scientific Research Institutions Based on Ontology Model

Guo Hongmei and Zeng Jianxun

Institute of Science and Technical Information of China, Beijing 10038

Abstract: Scientific research institutions are important objects of scientific research management and evaluation, and important units of resource organization. With the development of Chinese science and technology, scientific research outputs have increased exponentially, and scientific research activities have a wide range and diverse forms. Excavating the characteristics of scientific research institutions from the massive scientific research results and identifying the related institutions from complex social networks has always been a concern of the scientific research community. Institutional profiling help to quickly understand the institution, assist scientific research planning and management decision-making and identify future partners and competitors. The paper explored to profile and identify the attributes and related institutions of scientific research institutions based on ontology model. It selected the subject category and industry category attributions to discuss the process of labeling the attributes of scientific research institutions. It selected cooperative and benchmarking institutions to discuss the process of identifying closely related institutions in the intricate network.

Keywords: Research institutions; Institutions ontology; Institutions profiling

分类号 G250.74

1 引言

科研机构是以社会和经济需求为导向，有明确研究方向和任务并持续有组织地开展相关研究与开发活动的机构。作为国家科学研究的主体，它们是科技资源和科学成果的主要创造者和发布者，在长期从事科学研究的过程逐步形成了各自特色，并建立了复杂的关联关系。科研机构丰富的属性特征和关联关系是进

*收稿日期：2021-04-01；修回日期：2021-08-31

基金项目：国家社会科学基金重点项目“基于知识组织的图书馆资源发现服务体系研究”（项目编号：17ATQ002）。

作者简介：郭红梅，女，1985，博士，馆员；曾建勋，男，1965，博士，研究馆员，博导，主要研究领域：信息资源建设、知识组织和数字出版，Zeng@istic.ac.cn。

行知识组织、资源关联、科研管理和评价的重要基础，如何从科研活动及其海量、多样化、非结构化的科研成果中挖掘隐含的属性特征和关联关系，提炼科研机构各自的特点并赋予相应的标签一直是科研界关注的重点。科研机构具有名称、性质、学科、行业等多种属性，而且存在层级、合作、引用等多种关联关系，这些多样的属性和复杂的关系具有本体特征，本体作为一种能在语义层面对知识进行描述的概念模型，能很好地对科研机构的属性进行定义和描述，并可基于知识推理来挖掘隐性的语义关系。因此，本文探索基于本体的方法和思维来构建科研机构画像标签体系。为满足更细粒度的机构索引和管理需要，不仅针对一级科研机构，更深入到对下属二三级机构的标签体系构建。

本文贡献主要包括三个方面：（1）对机构的精准画像可以支撑以机构属性标签为入口的检索和导航。机构画像过程中对机构的地域、性质、职能、学科、行业等多种属性进行了标签化，标签化后的机构不再是孤零零的名称，而是一个个内涵丰富的实体，可按照某种或某些标签对机构相关的资源进行检索和导航，对具有相同标签的机构进行关联检索，将具有某种或某些特征的机构同时检索并推荐出来，弥补了传统上仅按照机构名称对机构进行关联、对机构相关资源进行组织、检索和导航的不足；（2）支持更精细的统计分析和科学评价需求。机构画像不仅仅对一级机构，而是对其下属二三级等更细粒度机构的标签化，可实现从深层次对机构的理解和把握，支持按照某种或某些属性标签对处于不同层级的机构进行遴选、统计、聚类或对比分析，而且也可按照某种特征准确定位与某机构最相关、最细粒度的关联机构，满足多元化的信息支撑服务。（3）支持机构知识图谱的构建。在画像过程中构建了科研机构之间的合作、引用和层级关系，形成了复杂的关系网络，可通过机构知识图谱将科研机构丰富的特征标签和关联关系构建成一个完整的知识体系，当用户搜索某机构时，机构知识图谱可以提供对该机构最全面的摘要，让用户快速得到机构的科研、学科、行业、关联机构等详细标签，辅助用户在短时间内获取最想要的信息，深入和广泛地了解机构之间的关系。

2 国内外相关研究

对机构本质的认识经历了虚拟主义理论、现实主义理论和名义主义理论三个阶段，虚拟主义理论认为机构是由权利和义务相关对象组成的、独立存在的虚拟实体。现实主义理论认为机构是由不同的成员构成的、人为赋予的、独立存在的真实实体。名义主义理论认为机构是由所拥有的成员及其成员之间的关系构成的、具有复杂社会关系的独特实体，该理论是构建机构本体的基础[1]。在名义主义理论基础上，学者们对机构给出了更多具体的定义，Hodgson 认为机构是规范社会相互交互行为的、既定和普遍存在的社会规则系统[2]。Scott 认为机构是保证社会稳定的规则、规范和文化认知结构[3]。Searle 与 Paul 认为机构是通过人的交流交互来创建和维持的，但独立于人类的信念而存在，交流交互是机构存在的本质，并提出利用本体来描述和揭示机构的特征[4-5]。

随着机构本质理论体系的逐渐成熟，国内外学者也认识到机构是复杂的社会实体，由多种构成要素并具有自身特征，他们探索了多种方法来提炼机构的特色之处并赋予标签以实现画像的目的，按照画像对象的不同将相关研究细分为面向机构相关主体的画像和面向机构自身属性特征的画像。

2.1 科研机构主体特征的画像方法研究

对科研机构主体的画像又称为用户画像，用户画像概念最早是由 Copper A 提出，他认为画像是基于用

户真实行为数据而构建的虚拟模型，随着学者研究的深入开展，其内涵越来越丰富[6]。用户画像是实现大数据环境下精准化信息服务的重要工具，近年来在人工智能、数据挖掘、信息检索、图书馆、健康医疗、商业营销等领域得到广泛的研究和具体应用。

Eke 等总结了信息检索和推荐领域中用户画像最新的研究进展，包括用户特征的提取、画像的技术和方法、画像的过程以及画像的效果等[7]。曾建勋指出数字图书馆服务必须将用户需求与知识创造相结合，从多维度对用户的属性特征进行细分和描述，以实现在知识创造过程中提供精准服务[8]。刘海鸥等对用户画像的概念、构成要素、模型等进行总结，将用户画像方法分为基于行为的画像方法、基于兴趣偏好的画像方法、基于主题的画像方法和基于人格特性与用户情绪的方法[9]。Liang 等构建动态用户和词嵌入模型对 Twitter 上动态用户进行画像[10]。陈泽宇等在 LDA 主题模型和神经网络模型的基础上，采用森林分类算法对用户属性进行分类以实现用户画像[11]。Gu 等参照 MagicFG 画像模型，对大数据环境下 Web 用户的行为进行画像[12]。

2.2 科研机构自身特征的画像方法研究

科研机构是具有多种属性特征的社会实体，其属性可细分为相对稳定的静态属性和随时间变化的动态属性。静态属性在机构官网上均有介绍，容易识别，描述相对简单。动态属性较为复杂，如何对他们进行准确标签化，国内外学者进行了很多探索。本体作为重要语义知识描述工具，可实现对机构属性和关系的综合全面描述和关联揭示，学者们探讨了多种机构本体的构建方法。此外，为满足具体应用场景的个性化需求，学者们也深入探索了针对某些具体属性的描述方法。

2.2.1 科研机构的本体描述模型构建方法研究

学者们探讨了多种机构本体的构建方法，通过构建本体模型来对机构的属性及交流交互过程中形成的复杂关系进行定义、描述和揭示。马里兰大学构建了高校本体，定义了描述高校及相关活动的元素，如学生、教员、课程、科研成果等[13]。Rabab C 等基于本体构建了机构知识记忆模型，对相关的人、资源、技术等进行描述和定义[14]。Lorenzo PG 研究机构本体中的属性类型与表征符号的关系[15]。Owen E 等提出支持不同信息架构的机构本体[16]。2010 年 Epimorphics 公司构建了政府机构本体[17]。为促进数据的共享，增强互操作行，W3C 对 Epimorphics 机构本体进一步扩展，发布新的机构本体，旨在支持多个领域机构信息的关联数据发布[18]。叶壮壮将 Wikidata 和 DBpedia 两个知识库已有机构属性进行融合来构建科研机构本体[19]。金玉琴等探索数字人文数据基础设施建设中的机构本体构建方法[20]。胡雪环等从科研机构的属性、关系、演化路径以及层级结构等方面探索科研机构本体的构建方法[21]。

2.2.2 科研机构的属性描述方法研究

学者们针对某种或某类属性的描述方法进行了深入研究。曾建勋、贾君枝等针对科研机构名称构建了机构规范文档的语义化描述模型，并引入 Schema 词汇表对其进行描述[22]。Paul 等提出了机构概念描述模型，从角色、规则、权利、责任和过程角度对机构进行描述，并定义了不同实体的描述准则[5]。MaxwellA 等基于机构的变革过程理论和实施理论提出机构描述发展模型，用于评价分析机构在发展过程中的特征、相似性、差异性、劣势和优势[23]。孟琳等通过对多源知识进行数据获取、信息融合和挖掘后，对机构的核心成员、机构兴趣等动态属性进行抽取和画像研究[24]。Taneja G 等认为高校网站首页上不同标签字段的检

索浏览情况,可反映学生对高校的关注情况，从而辅助学生进行高校的选择，通过对国外高校网页元数据字段的浏览分析发现,学生更关注学校的研究领域、学术项目、地理位置和科研环境[25]。Galan M 等研究发现高校的课程设置、声望、评价评议、就业情况、学费等是学生在择校中比较关注的属性[26]。Kettune J 等研究了与高等教育机构相关联对象的特征，关联对象包括影响机构发展的其它组织、客户以及内部的员工和学生等[27]。

国内外学者通过构建本体、描述模型或挖掘算法对机构的属性和关系、用户行为等进行显性化描述方面，积累了很多有益的理论和实践经验，不断丰富机构画像方法技术体系，但仍存在以下几点不足：（1）大多研究只是面向具体应用需求，针对科研机构某些具体属性进行定性描述，没有从整体上对科研机构的属性和关系进行综合全面的梳理，而且已有的研究主要集中在对一级机构属性和关系的描述揭示，很少涉及对下属更细粒度机构的分析。（2）对科研机构的行为特征描述揭示不够，已有的画像研究主要集中在对科研机构成员或具体科研用户行为特征的描述，很少有在用户之上对机构行为及其关联关系进行描述揭示。（3）科研机构画像的目的是支撑以机构为单元在文献检索、分面导航、定标比超、统计评价分析等方面的应用，但目前大多方法还处于理论探索阶段，缺乏对具体场景下应用效果的验证。因此本文以科研机构在知识组织、关联揭示和检索导航等应用场景的具体需求为导向，综合分析科研机构的特征和关联关系，基于本体思维构建一套能准确定义和描述科研机构属性和关系的标签化方法体系，不局限对一级科研机构的描述，还适用于对下属二三级机构的描述。

3 面向科研机构画像的本体模型构建

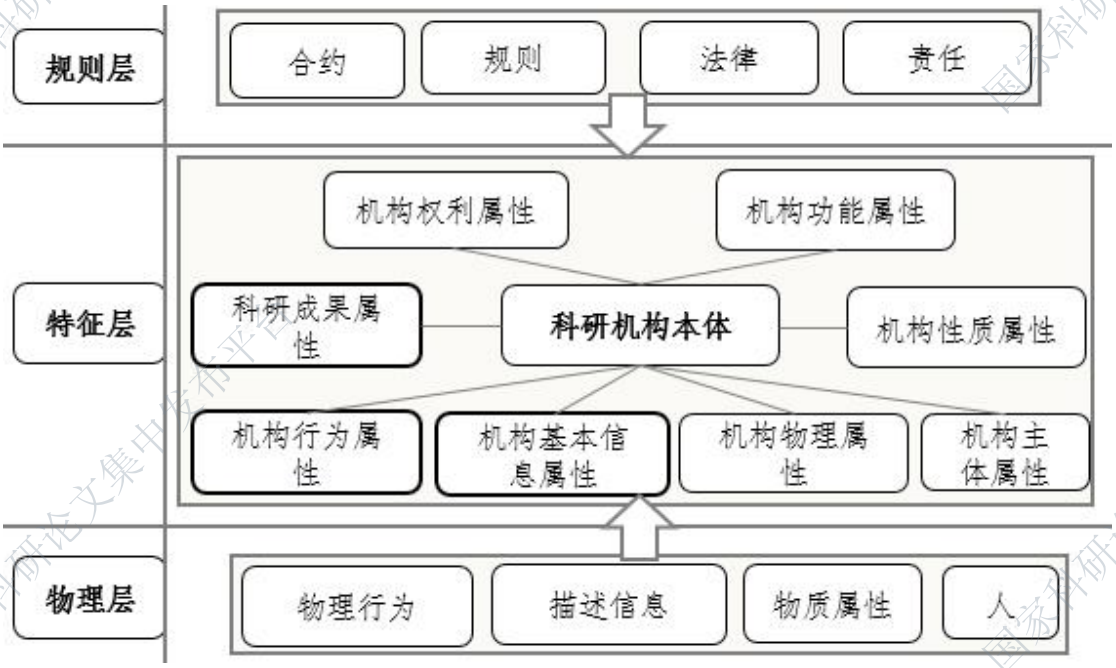
科研机构作为国家科学研究的主体，处于社会关系网络之中，除具有普通社会对象共有的经济、法律、行为特征等外，在从事科学研究的过程中也逐步形成了自身的科研特征，比如学科、行业、研究主题等。此外，科研机构之间彼此还建立了合作、引用等关联关系，这些特征和关系通过科研机构相关的属性进行揭示。文中借鉴 Paul J 等提出的概念模型[5]，采用自下而上的思想构建科研机构本体模型，根据各属性在机构发展中的作用将它们分为物理层、特征层和规则层（如图 1 所示），不同层的属性相互作用，共同支撑机构的持续发展。最底层是物理属性层，主要包括科研机构所依赖的物理主体、物质和行为，对特征层属性起支撑作用；最顶层是规则层，主要包括科研机构的所要承担责任和所有遵循的合约、规则、法律等文件，对科研机构进行约束控制。中间层是特征层，主要对科研机构的基本信息、科研成果、科研行为和机构主体等进行描述。科研机构主体通常指机构的法人和成员，基本信息属性主要包括机构简介、发展历程、联系方式等描述字段，通常利用文本或图像表示。物理属性用于描述科研机构的硬件设施等。行为属性来描述机构在科研活动中的行为。科研成果属性用于描述机构的产出特征。功能和性质属性主要对科研机构的性质职能特征。物理层和规则层通常不直接体现科研机构的特征，特征层的各个属性用于描述揭示

机构的不同特征面，它们并不是孤立存在，而是相互关联、相互作用共同对机构本体进行限定描述。

图 1 科研机构本体模型

3.1 科研机构本体的标签体系研究

科研机构本体是由多个属性相互作用共同来描述限定，通过对各个属性和关系的总结分析，凝炼出科研机构在社会关系、社会属性、科研活动等方面需要描述的属性特征，如图 2 所示。社会属性主要是科研机构作为社会实体所具有的身份地位、权利义务、目标任务和性质职能等；基本属性主要包括机构的通用描述信息，如机构名称、地域归属、联系方式、发展历程等；关系属性是指科研机构在参与科研活动过程中，与其它社会实体产生的关联关系，例如由于名称变更、拆分、合并等过程中产生的沿革关系，科研成果合作产生的合作关系以及机构组织架构中得到的层级隶属关系；科研属性是对科研行为的描述，包括产生的科研成果、主要活动领域、关联机构等。



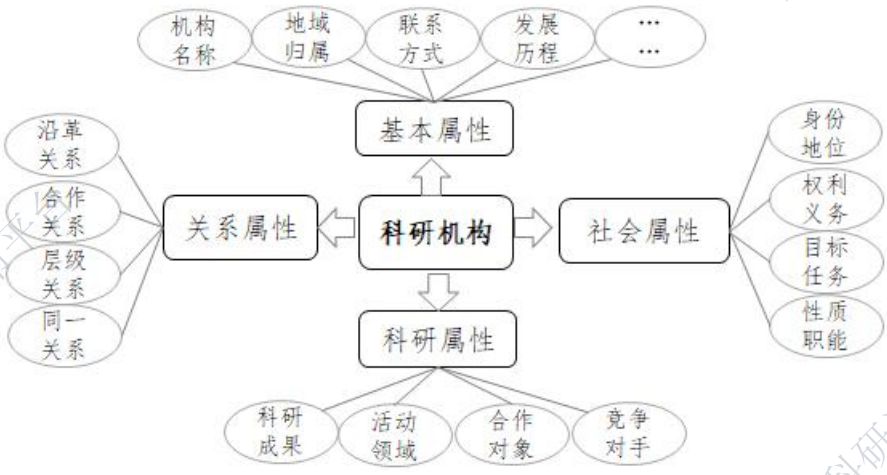


图 2 科研机构本体的属性特征

科研机构标签化就是利用标签体系勾画它在从事科研过程中所形成的社会属性、关系和领域的过程，精准、细粒度且结构化的标签体系是机构画像的基础，其广度和粒度对机构画像的精确性有较大影响，因此首先要提炼科研机构的标签，形成机构标签库，包括特征标签、关系标签等。对科研机构本体中各实体的属性和关系的抽象凝练得到科研机构在社会、关系、科研等四个方面的主要属性特征，按照各属性特征在机构画像中的作用和关系将它们分为三类，分别是描述信息标签、关联关系标签和关联机构标签，从三个维度构建标签体系，如表 1 所示。

表 1 科研机构本体的标签体系

一级标签	二级标签	标签内容来源
描述信息类标签	基本特征标签	各种变体名称、地域归属、联系方式、发展简介、组织架构等
	社会特征标签	身份地位、权利义务、性质职能、任务目标等
	科研特征标签	活动领域、成果类别、成果数量、影响力等
关联关系类标签	层级关系	主管部门、上级机构、下级机构等
	沿革关系	更名、合并、拆分、重组等变革关系
	合作关系	文献、基金、专利等成果中的合作
	同一关系	同一机构不同分类体系下的名称
关联对象类标签	合作机构	合作强度较大的机构
	对标机构	规模、研究内容和水平等方面相当的机构

3.2 科研机构本体的标签化流程研究

科研机构的静态属性相对稳定，比如机构名称、地域信息、联系方式、创立时间等，动态属性是由静态属性衍生而来，并随着内容扩充和时间推移而变化，比如机构的活动领域、关联机构等。静态属性获取方式较为简单，而动态属性标注过程相对复杂，需要基于机构行为、科研成果和已有的静态属性综合推理得到，因此，在机构属性标签化过程中按照获取的难易程度分层次进行标注，具体流程如图 3 所示。首先获取机构的基本属性信息，它们是识别和构建机构关联关系的基础，也是对科研活动进行描述的基础，机构名称、地域归属、联系方式、发展历程等可以通过本地收割或远程采集从已构建的机构规范库、文献及相关成果库和机构官网等获取。其次基于已标注属性和机构本体中不同实体之间的关联和作用，识别机构间的关系，例如对机构名称变更过程的分析可以得到机构实体的沿革关系，对机构主管、主办单位属性的分析可构建机构的层级隶属关系，对科研成果的参与机构分析可构建机构间的合作关系，对科研成果研究主题的分析可得到机构间的学科、行业或研究兴趣的相似性关系等。最后基于构建的关系数据，利用主题分析、规则和知识推理的方法识别主要关联机构，并计算每个关联机构的关联强度，从而为某机构推荐相关或相似的机构，实现机构间的科研合作和定标比超。

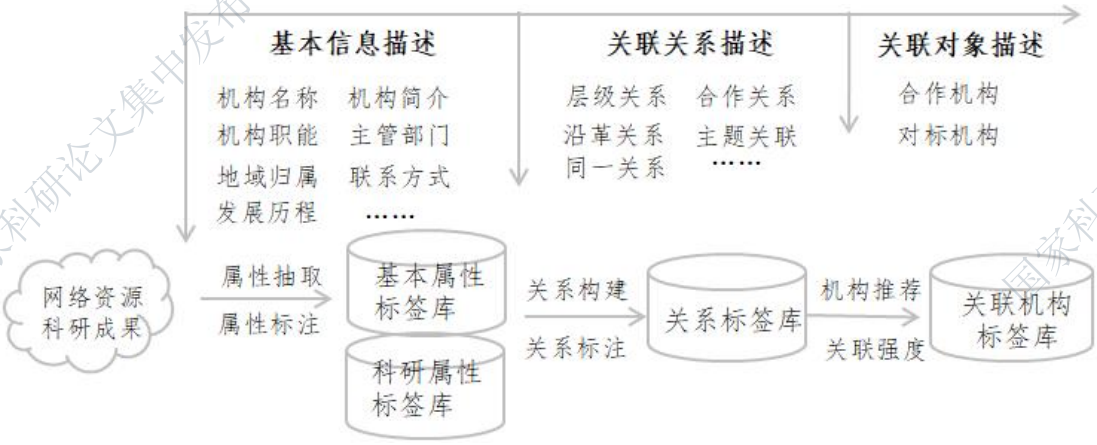


图 3 科研机构本体的标签化流程

4 科研机构特征的标签化方法研究

4.1 基本属性的标签化方法

重点选取能够揭示机构科研特征的、相对复杂的动态属性进行标签化。机构的性质和职能决定机构的社会责任和发展方向，对机构的发展有指引导向作用，是标注机构学科和行业的基础。活动领域标签是进行科研管理评价、统计分析、识别竞争对手和合作团队的前提和基础，而且随着科学的发展，机构的活动领域也在不断的调整和扩充，远超越了创建之初的设想，所涉及的学科和行业范围越来越广，因此本文以表征科研机构性质职能和活动领域的学科类别和行业类别以及关联机构为例，来论述科研机构属性特征的标签化过程。目前科研机构的画像、排名和评价研究大多针对一级机构，由于一级机构多是综合性机构，所赋值的活动领域特征标签粒度较粗，不能满足从更细学科粒度上进行科研管理的需要，因此本研究构建的标签体系主要针对下属二三级机构的特征进行描述，更专指、更具体，满足从更细的学科和层级粒度对科研机构进行评价和管理。

4.1.1 性质职能的标签化方法

由于机构在发展历程中新建、更名、拆分、合并等现象频繁发生，根据机构存在的时效性将其分为连续体和非连续体，连续体是指在较长一段时间内持续稳定存在，具有实体形式的机构组织，比如某所高校或研究所等。非连续体是由于社会发展需要，在一定时期内存在，一般需要依赖其它实体机构而存在，比如国家重点实验室。根据不同层级机构间的关系和是否有独立法人地位，又将连续体分为独立体和依赖体，比如某高校是独立体，而它下属的院系需要依赖高校实体而存在，属于依赖体，具体见表 2。

国家标准《组织机构类型（GBT20091-2006）》体系按照机构的功能和性质，将其分为企业、机关、事业单位、社会团体以及其他组织机构五大类，科研机构主要分布在企业和事业单位中，从实际应用需求的角度出发，并结合科研机构所从事的重点业务，对一级机构及下属机构分别进行性质特征描述。将一级机构独立体划分为高等院校、科研院所、医疗机构、企业、学协会等，将独立体下属的依赖体划分为管理部门、业务部门、服务部门等。

表 2 机构性质职能标签的特征词

连续性	独立性	性质职能	特征词

连续体	独立体	高等院校	大学、学院、学校等
		科研院所	研究所、研究院、研究中心等
		公司企业	公司、集团、厂等
		医疗机构	医院、疾病控制、预防等
		其它	学会、协会、编辑部等
	依赖体	管理部门	人事、财务、办公室、行政、党政等
		业务部门	学部、学院、系、教研室、研究中心等
		服务部门	服务部、后勤、指导中心、校医院等
		其它部门	图书馆、档案馆等
非连续体	依赖体	国家重点实验室	实验室
		国家工程中心	工程中心
		基地中心	示范基地、培训基地、教育基地等

4.1.2 学科属性的标签化方法

科研机构的学科类别通常体现在机构的名称、科研成果和人才培养三个方面。机构名称是创建时所赋予的，它能标识机构最初设置的目标和研究方向，不少高校和研究所名称中就存在标识学科类别的词语，比如中国医科大学（医学）、中国药科大学（药学）、中国政法大学（法学）、中国科学院化学研究所（化学）、中国科学院声学研究所（声学）等。由于机构的名称相对固定，不会轻易更改，文中将从机构名称中得到的学科类别称为静态学科。此外，在机构发展过程中所从事的研究领域也会随着需求进行调整，比如为满足社会或科技需要，或为了发展机构特色，或为了追求国际热点等，机构布局新的研究领域，文中将其称为动态学科，通常体现在科研成果和人才培养的学科方向。静态学科和动态学科从不同角度揭示了机构的学科布局，因此机构学科类别的标注应综合静态学科和动态研究领域两方面的特征，如图 4 所示。

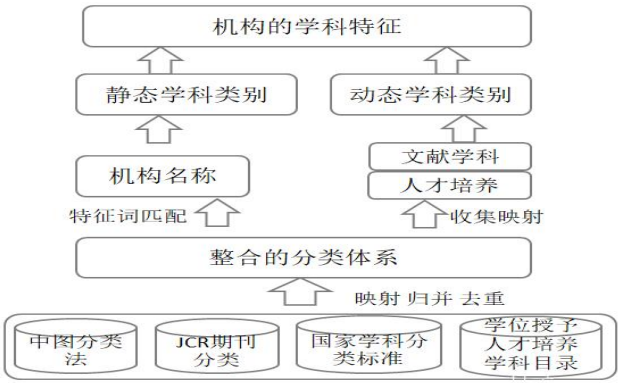


图 4 科研机构学科属性的标签化

目前存在多种学科范畴分类标准和体系，常用的有《学科分类与代码国家标准（GB/T 13745-2009）》、学位委员会和教育部颁布的《学位授予和人才培养学科目录》、《中国图书馆分类法》和 Web of Science(WoS)数据库的 JCR 期刊学科分类体系。由于使用目标不同，不同分类标准中所设置的学科粒度也存在差异，比如《学位授予和人才培养学科目录》和《学科分类与代码国家标准》相比，前者在医学和管理学领域划分的比较详细，而后者在经济学和语言学领域划分的比较详细。为兼顾不同的分类体系和中国机构的学科特

点，将几种学科体系进行映射、合并融合。

由于不同机构命名没有特定规则，对于静态学科类别的标注，需要预先构建不同学科领域的特征词典，然后依据机构名称中的特征词来标注机构的学科类别。为充分准确构建不同学科下的特征词典，选取各领域 4300 个高被引机构作为训练数据，对 40 多万条二三级机构名称进行预处理，抽取能表征机构学科类别的词语映射到相应学科中，构建各学科的特征词典，表 3 列出了部分学科所标注的特征词。将机构名称与已构建的学科类别词典进行匹配，实现不同层级机构静态学科的标注，对于无法按照特征词映射上的机构，分别与四种分类体系的最细粒度层级进行比对，如果匹配上则取其上级类值。

表 3 学科属性标签的特征词

学科类别	特征词
电子科学与技术	电子技术、电子科技、电子科学、电子信息
公共卫生与预防医学	防治、公共卫生、疾病预防、卫生管理；预防控制、预防医学
环境科学与工程	环境保护、环境工程、环境监测、环境科学
经济学	保险、财经、财贸、财政、金融、经济、经贸、商贸
图书情报与档案管理	档案、计量、情报、图书、文献
新闻传播学	传播、传媒、新媒体、新闻
信息科学与系统科学	系统科学、信息管理、信息科学、信息系统
信息与通信工程	通信工程、信息工程、信息通信
体育学	体育、运动
社会学	人文、社会、社科
计算机科学与技术	大数据、计算机、人工智能
航空、航天科学技术	航空、航天、宇航
城乡规划学	城市规划、城市建设、城乡规划、城镇规划

科研成果是机构参与科研活动的主要产物，科研成果的学科分布可反映机构关注的领域，揭示研究主题随着时间的演化和转移，文献是科研成果的主要形式，因此以文献资源为核心来分析机构的动态学科特征。文献的学科类别可以分别从发文期刊和施引期刊的学科获取，发文期刊的学科是机构主动选择的，而施引期刊的学科是外部学者对文献的理解，是客观自发的行为，二者从不同角度揭示机构的研究主题分布，可以相互验证和补充。此外，科研机构担负着人才培养的责任，所设置的学科和专业也可反映机构的特色、发展策略和研究领域，因此收集不同层级机构所设置的本科专业、授予的硕士、博士研究生学位方向，补充文献的学科领域。

4.1.3 行业属性的标签化方法

科研机构在从事科研活动、服务社会和支撑国民经济发展的过程，也会产生一定的社会经济效益，通常体现在不同的行业类别中，对机构行业类别的标注有助于对比机构科研成果的应用效果或服务社会的成效，尤其是一些以技术为主的科研机构，在成果转化过程中为不同行业带来了较大的社会效益。科研机构所涉及的行业主要集中在教育业、科学研究和技术服务业、信息传输、软件和信息技术服务业、卫生和社会工作等类别中。由于国民经济行业分类在不同行业的分类详细程度存在差异，比如制造业较为详尽，而

在科研机构比较集中的教育和科学研究和技术服务业分类较为粗略，为了准确标注各机构的行业，并尽量保证各机构的行业在可比的层级上，按照实际需求对不同大类下的行业类别进行层级调整，比如将 Q841 医院（Q 卫生和社会工作）与 C27 医药制造业（C 制造业）调整为同一层级，尽量保证不同行业分类体系保持在相同粗细粒度上进行标注和对比。

4.2 关联关系的标签化方法

机构间层级关系、发展沿革关系、科研合作关系和科研引用关系等多种。层级关系通常体现在机构的组织架构和科研成果的机构署名中。沿革关系用于描述机构发生变更前后，新旧机构之间的替代与被替代关系，通常包含两种情况，一是由于机构自身的变化，主要包括普通更名、改制更名、升格更名、转设更名等；二是涉及多个机构的名称变更，主要包括合并更名、合并转设更名、拆分更名等。合作关系构建主要基于科研成果，如果两个或多个机构同时参与一项或多项科研成果中（科技文献、专利、基金项目等），则这些机构两两之间具有合作关系。此外，将标识同一资助项目的科研成果的机构也视为合作关系。引用关系主要反映在科研成果的参考文献中，一般分为直接引用关系、共被引关系和耦合关系，引用关系越强的机构之间研究主题越相似。

4.3 关联对象的标签化方法

科研机构在长期从事科研活动过程中形成了自身特关联机构是指在与某机构关系比较紧密的机构，主要体现在两机构的科研活动或科研成果的交互程度，集中在合作或引用关系较强的机构，因此将合作强度和引用强度较大的机构均视为关联对象。关联对象的标注是识别合作伙伴和竞争对手的基础，二者存在交叉重叠，通常合作密切的机构也是同领域内科研势力相当的机构，存在竞争关系。

4.3.1 合作机构的标签化方法

合作机构的识别主要基于科研成果中的署名机构来判断，出现在同一科研成果中的机构即为合作机构，合作的科研成果越多，两机构的合作关联强度越大。本文主要基于公开发表的文献、专利和基金项目中的署名机构来识别合作机构。除了作者署名机构字段外，部分文献、专著和专利数据中还具有基金项目字段，本研究将标识同一基金项目的科研成果的署名机构也视为合作机构。分别计算某机构与各领域中其它机构的合作强度，强度较高的即为该领域内所识别出的合作机构。

4.3.2 对标机构的标签化方法

对标机构通常是指综合实力与本机构水平相当的机构，它的识别需要权衡科研机构的活动领域、人员规模、科研产出、学术影响力和国际地位等各方面的属性特征，运用知识推理的方法，依据综合性评判结果来确定，并不局限在同层级机构中。活动领域相同是指两个机构在相同分类体系下，学科或行业领域一致。科研人员规模相当是确保两个机构体量一致，具有可比性和公平性。在科研人员规模相当的情况下，通过科研产出指标和学术影响力指标来测度不同领域中的对标机构。科研产出通常利用科研成果论文量来衡量，学术影响力利用引文数量来衡量，其它科研合作指标和社交媒体指标等可以作为辅助，在必要情况下使用。对标机构的识别是与领域相关的，按照机构所属的科研领域可将机构分为专业领域机构和综合性机构，对于某综合机构如果查找某具体领域的对标机构，则推荐出的对标机构可能是单领域机构，也可能

是综合机构的下属子机构。如果要推荐某综合性机构的对标机构，不关联某具体学科，则推荐的对标机构也应该是综合机构，按照领域分别计算与某综合性机构的相关性，然后将各领域相关性进行综合来推荐相关机构。

5 实证研究

对科研机构来说，活动领域和合作机构是最重要的两个属性，因此文中重点选取这两个属性进行机构特征标签的实证研究。选取 2019 年中国高被引分析报告中物理学领域的高被引机构天津大学和清华大学为示范机构，对它的活动领域和合作机构进行识别和标注。

2011-2018 年，物理学领域的 64 种期刊上共发表学术论文 62682 篇，其中天津大学第一作者发文 625 篇（截止到 2019 年被引 1443 次），清华大学第一作者发文 935 篇（截止到 2019 年被引 1416 次）。从第一作者高发文期刊来看，天津大学发文主要集中在物理学学报、光学学报、光谱学与光谱分析等期刊上，清华大学发文主要集中在物理快报、物理与工程等期刊上。从天津大学发文期刊细分领域可知，它在物理学领域的主要活动领域是光学，清华大学则以力学和工程为主。

表 4 天津大学和清华大学在物理学领域第一作者高发文期刊

	天津大学		清华大学	
序号	期刊名称	发文量	期刊名称	发文量
1	物理学报	52	中国物理 B（英文版）	75
2	光学学报	50	中国物理快报（英文版）	63
3	光谱学与光谱分析	43	物理学报	57
4	中国物理 B（英文版）	40	中国科学:物理学 力学 天文学（英文版）	36
5	中国激光	31	物理与工程	35
6	激光与光电子学进展	22	光学快报（英文版）	30
7	中国光学快报（英文版）	16	中国物理 C（英文版）	27
8	光子学报	14	物理	25
9	纳米技术与精密工程	14	中国激光	18
10	中国物理快报（英文版）	13	中国科学（物理学 力学 天文学）	18

从表 5 可知，天津大学和清华大学在物理学领域的主要合作机构上存在差别，二者没有交叉重叠。天津大学的主要合作机构是南开大学、天津师范大学、中国科学院半导体研究所等，清华大学则主要与西北核技术研究所、中科院物理所、中国工程物理研究院等机构合作。此外即使同一个机构，从他与不同机构合作发文的期刊来看，他们合作具体的研究主题也存在差别，例如天津大学与南开大学和中科院半导体研究所的合作范文集中在光学和激光，与天津师范大学和河北工业大学则集中在光谱学，清华大学与西北核技术研究所和中国工程物理研究院的合作发文则集中在激光领域。

表 5 天津大学和清华大学主要合作机构与合作发文期刊

	天津大学		清华大学	
序号	合作机构	合作论文的主要发表期刊	合作机构	合作论文的主要发表期刊
1	南开大学	光学学报、中国激光	西北核技术研究所	强激光与粒子束、核电子学与探测技术
2	天津师范大学	光谱学与光谱分析、中国物理快报	中国科学院物理研究所	低温物理学报、中国物理学报
3	中科院半导体研究所	中国光学学报、发光学报	中国工程物理研究院	强激光与粒子束、中国激光

4	天津理工大学	光电子激光、功能材料	中国科学技术大学	光学技术、中国光学学报
5	河北工业大学	光谱学与光谱分析、物理学报	北京大学	物理快报、理论物理通讯

从天津大学和清华大学在物理学领域的活动领域分析可看出，在较粗学科分类粒度上机构的研究领域虽然相同，但在细粒度研究主题上却存在很大差别。从合作机构来看，不仅他们合作的机构有差异，即使与同一个机构合作，他们的合作主题也会存在很大差异。因此，必须构建机构全面的、细化的标签化体系，才能对机构进行准确地描绘和客观地评价。

6 结语

本体模型从语义层次上对科研机构的概念、属性及关联关系进行全方位的定义和描述，不仅揭示了科研机构的学科、行业等属性和科研行为关联，还通过简单的知识推理形成语义化的关系网络，满足语义环境下检索和导航等服务应用需求，是揭示科研机构复杂属性和关联关系的最优工具。以科研机构本体为基础的机构画像可在对机构属性特征和关系进行知识推理和关联挖掘的基础上，提炼各个机构的特征，构建更细粒度和广度的标签化体系，可辅助用户快速直观了解某个机构特色、发展水平、活动领域等，从一个更为全面客观的角度提供对机构的信息挖掘和分析，对具有相同特征标签的机构进行分析，便于机构与机构之间的比较，辅助宏观决策和预测科研机构的发展趋势，识别潜在合作伙伴和竞争对手等。文中重点以科研本体模型和标签体系的构建理论研究为主，实证部分仅选取物理学领域高被引机构天津大学和清华大学进行学科领域和合作机构属性特征的标注，下一步将在不同场景进行大批量数据的应用示范研究。并根据具体的应用需求对方法体系进行调整和优化，构建适用不同应用场景的科研机构标签体系。

7 参考文献

[1] Lowell TV. CorporateBeing:A Study In Realist Ontology[D]. Buffalo: The State University of New York, 2006.

[2] Hodgson, GM. What are institutions?[J]. Journal of Economic Issues, 2006, 40(1) :1-24.

[3] Scott WR. Institutional carriers: reviewing modes of transporting ideas over time and space and considering their consequences[J]. Industrial and Corporate, 2003, 12(4): 879-894.

[4] Searle JR.What is an institution?[J] Journal of institutional economics, 2005, 1(1): 1-22.

[5] Paul J, Maria B, Owen E. Institution Aware Conceptual Modeling[EB/OL]. [2021-3-25]. http://ceur-ws.org/Vol-1979/paper-15.pdf

[6] Cooper A. The Inmates are Running the Asylum: Why High-Tech Products Drive Us Crazy and How to Restore the Sanity[M]. Sams Publishing. 2004: 157-189.

[7] Eke CI.A Survey of User Profiling: State-of-the-Art, Challenges and Solutions[J].IEEE Access, 2019, 4：1-19.

[8] 曾建勋.精准服务需要用户画像[J]. 数字图书馆论坛, 2017, (12):1-1.

[9] 海鸥,孙晶晶,苏妍螈等.国内外用户画像研究综述[J].情报理论与践, 2018, 41(11):155-160.

[10] Liang S, Zhang X, Ren Z, etal. Dynamic embeddings for user profiling in twitter[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, 2018: 1764-1773.

[11] 陈泽宇,黄勃.改进词向量模型的用户画像研究[J]. 计算机工程与应用, 2020, 56(1):180-184.

[12] Gu XT, Yang H, Jie Tang et al. Profiling Web users using big data[EB/OL]. [2021-3-25]. https://doi.org/10.1007/s13278-018-0495-0

[13] University Ontology[EB/OL]. [2021-3-25]. http://www.cs.umd.edu/projects/plus/SHOE/onts/univ1.0.html.

[14] Rabab C. Building corporate memories in collaborative way using ontologies[EB/OL]. [2021-3-25].https://www.researchgate.net/publication/239764861_Building_corporate_memories_in_collaborative_way_using_ontologies_Case_study_of_a_SSII.

[15] Lorenzo PG. Institutional ontology as an ontology of types[EB/OL]. [2021-3-25]. https://www.researchgate.net/publication/277279339_Institutional_Ontology_as_an_Ontology_of_Types

[16] Owen E, Paul J, Maria B. Institutional ontology for Conceptual Modeling[EB/OL]. [2021-3-25].https://doi.org/10.1057/s41265-018-0053-2

[17] Epimorphics[EB/OL]. [2021-3-25].http://epimorphics.com/public/vocabulary/org.html

[18] W3C[EB/OL]. [2021-3-25].https://www.w3.org/TR/vocab-org/

[19] 叶壮壮. 基于 Wikidata 的机构本体构建研究[D]. 太原: 山西大学, 2019:11-16.

[20] 金家琴, 夏翠娟. 数字人文数据基础设施建设中机构本体的构建:研究和应用[J].图书馆论坛, 2020, 40:34-43.

[21] 胡雪环. 科研机构本体的构建方式研究[D]. 北京:中国科学技术信息研究所, 2016:13-35.

[22] 曾建勋, 贾君枝.机构名称规范数据的语义模型构建[J].大学图书馆学报, 2019, 1:42-47.

[23] Maxwell A, Judith A. Organization Development Models: A Critical Review and Implications for Creating Learning Organizations[J]. European Journal of Training and Development Studies, 2015, 2(3):29-43.

[24] 孟琳.多源信息融合的机构画像的方法研究[D]. 北京: 北京邮电大学, 2018:28-46.

[25] Taneja G. How are higher education institutions defining their meta-description tags?[J]. International Journal of Educational Management, 2018, 32(7):1293-1306.

[26] Galan M, Lawley M, Clements M. “Social media’s use in postgraduate students’ decision-making journey: an exploratory study”[J]. Journal of Marketing for Higher Education, 2015, 25(2):287-312.

[27] Kettunen J. Stakeholder relationships in higher education[J]. Tertiary Education and Management, 2015, 21(1):56-65. DOI: 10.1080/13583883.2014.997277.