基于本体的科研机构画像标签体系构建方法研究*

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，随着网络技术的发展和中国对科技发展的重视，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各个科研机构的关联机构一直是科研界关注的问题，本体模型可从语义层面对科研机构的多种属性和关系进行定义和描述，因此本文探索基于本体的方法对科研机构的属性和关联机构进行揭示、画像和识别，主要选取表征科研机构活动领域的学科类别和行业类别标注来论述科研机构属性特征的标签化过程，以合作机构和对标机构的识别来论述在错综复杂的关系网络中识别关联紧密的机构过程。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，还可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 Scientific Research Institutionsattribution on the basis of Ontology Model

Abstract: Scientific research institutions are important objects of scientific research management and evaluation, and important units of resource organization. With the development of Chinesescience 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-makingand identify future partnersand competitors. The paper explored to profile and identify the attributes and related institutions of scientific research institutionsbased 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 selectedcooperative 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

0 引言

科研机构是以社会和经济需求为导向，制定有明确的研究方向和任务并持续有组织的开展相关的研究与开发活动的机构[1]。作为国家科学研究的主体，它们

*本文为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）成果之一。通信作者：曾建勋，Email:Zeng@istic.ac.cn, ORCID: 0000-0002-0432-961.

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知识组织、资源关联、科研管理和评价的重要基础，也是准确进行竞争情报分析，识别不同行业领域竞争对手的前提条件，因此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和显性揭示这些隐性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、行业等多种属性特征并且存在层级、合作、引用等多种语义关联关系，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特征进行定义和描述，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的需要，本研究不仅针对一级科研机构，更深入到对下属二三级机构的属性特征的标签体系构建。

1 国内外相关研究

科研机构属性标签化是指对机构属性特征和关联关系进行分析，提炼它的特色之处并进行显性揭示、描述和形成标签的过程[²]。科研机构是具有多种属性的社会实体，这些属性又可分为相对稳定的静态属性和随时间变化的动态属性，静态属性描述相对简单，动态属性较为复杂，如何对属性特征进行准确画像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。

1.1 基于本体的科研机构描述方法研究

对机构本体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、真实存在的、人为赋予的、独立存在的真实实体。名义主义理论认为机构由所拥有的成员及其成员之间的关系构建，具有复杂社会关系的独特实体，该理论构建机构本体的基础[3]。在名义主义理论基础上，学者们对机构本体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[4]。Scott 认为机构是保证社会生活稳定的规则、规范和文化认知结构[5]。Paul 与 Searle 认为机构是通过人的交流交互来创建和维持的，但独立于人类和人的信念而存在，认为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[6，7]。

随着机构本体认识的加深，学者们试图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。2010 年 Epimorphics 公司构建了政府机构本体[8]。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，发布了新的机构本体，旨在支持多个领域机构信息的关联数据发布[9]。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[10]。Rabab 等基于本体构建了机构知识记忆模型，对相关的人、资源、技术等进行了描述和定义[11]。Lorenzo 研究了机构本体中的属性类型与表征符号的关系[12]。Owen 等提出了支持不同信息架构的机构本体[13]。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[14]。金玉琴等探索了数字人文数据基础设施建设中的机构本体构建方法[15]。胡雪环等从科研机构的属性、关系以及演化路径、层级结构等几个特征来探索科研机构本体的构建方法[16]。

1.2 科研机构具体属性的描述方法研究

科研机构具有多种属性，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的名称属性构建了机构规范文档的语义化描述模型，设计了机构名称的属性及不同机构实体间的关系，并引入 Schema 词汇表对其进行语义描述[17]。Paul J 等提出了机构概念描述模型，从角色、规则、权利、责任和过程几个实体角度对机构进行描述，并定义了不同实体的描述准则，该概念模型不仅依赖机构本体，还需要其它本体的支持，如 REA 本体和服务本体等[18]。MaxwellA 等基于机构的变革过程理论和实施理论提出了机构描述发展模型，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[19]。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，偏重对社团发现、关系抽取等算法进行改进和创新[20]。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[21]。Galan 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[22]。JuhaKettune 等研究了与高等教育机构相关联对象的关系特征，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，研究表明关联对象通过相互合作可改善机构的质量保证体系[23]。

国内外学者在通过构建本体、描述模型或挖掘算法来对机构的属性和关系显性化描述方面，积累了很多有益的理论和实践经验，不断丰富了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需求，针对科研机

构某些具体属性进行定性描述，忽略了在文献检索、分面导航、定标比超、统计评价分析等方面的应用场景，没有从应用需求的角度对科研机构所有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的描述揭示。（3）科研机构属性标签化的目的是支撑以机构为单元在文献检索、分面导航、统计分析等方面的应用，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属性特征和关联关系，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，不仅仅局限于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。

2 科研机构本体模型的构建

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、行为等属性外，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，此外彼此还建立了合作、引用等关联关系，属性和关系蕴含在与科研机构相关的社会实体之中，比如机构主体、机构物质实体、机构信息实体、机构权利实体、机构协议实体、机构功能实体、机构行为实体、科研成果实体等，这些实体相互作用，共同支撑机构的持续发展[24]。依据各实体在机构发展中的功能作用，采用自下而上的思想，在科研机构本体模型的构建过程中将实体分为物理层、特征层和约束层三个层次，如图1所示。最底层是物理层，主要包括科研机构所依赖的物质实体，比如人员、各种物理实体或物理行为，对特征层实体起支撑作用；中间层是机构特征层，描述机构的基本信息、科研成果、科研行为等属性特征；最顶层是约束层，与特征层实体进行交互并对其进行约束控制。机构主体是具有所有权、直接或间接以人类实体为基础的实体，通常指机构的法人和成员。机构信息实体以物质层的信息内容实体为基础来描述机构的基本信息，如机构简介、发展历程、联系方式等，通常利用文本或图像表示。机构物质实体是以物理实体为基础，描述机构的硬件设备等。机构行为实体以物理行为为基础，来描述科研活动中的行为。科研成果实体用于描述机构的产出。机构权利、协议、规则等实体用来规范和约束成员之间的相互交互。物理层和规则层实体通常不直接体现科研机构的特征，特征层的各个实体则用于描述揭示机构的不同特征面，每个机构实体都是通过相应的属性特征来描述和定义的如图1所示，但各实体并不是孤立存在，而是相互关联、相互作用共同对机构本体进行限定描述，比如机构的权利、功能和协议实体对机构行

为实体进行约束和限定，机构的物质实体和主体支撑机构的正常发展运行，科研成果实体来揭示行为实体的效果。

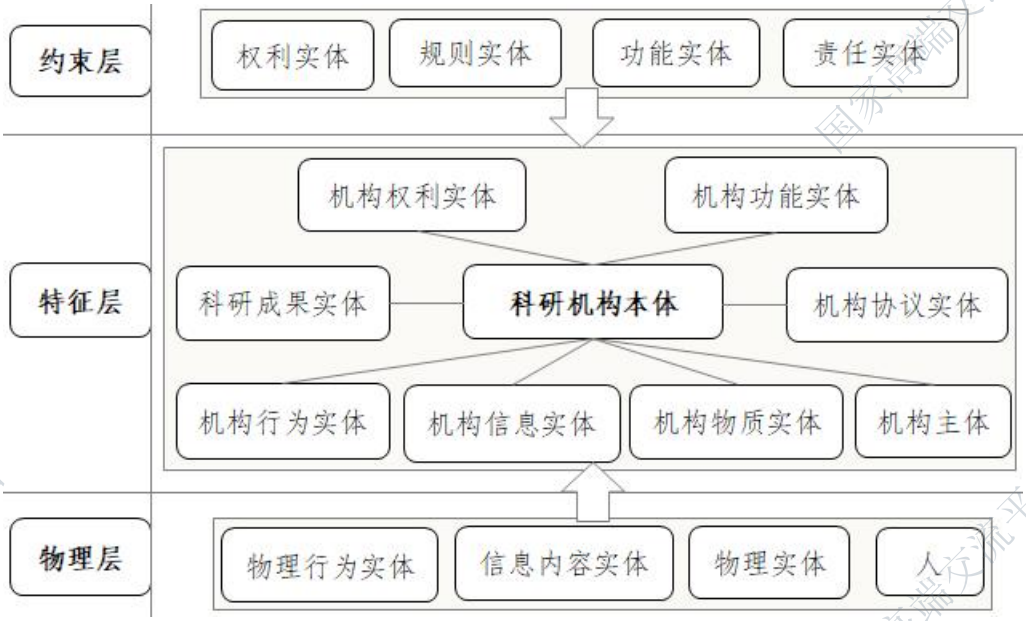

图 1 科研机构本体模型

表 1 科研机构相关实体的属性描述框架

实体类别	属性	描述字段	字段标识	必备性
机构信息实体	唯一性	唯一标识符	UnifiedID	必备
	机构名称	具体名称	Name	必备
		类型（规范名称、简称、别称、其它）	NameType	规范类型必备
		语种（中文、英文、日文等）	Language	必备
	地域归属	行政区域	City	必备
		区域代码	CADC	有则必备
		邮政编码	ZIP	有则必备
		详细地址	Address	有则必备
	组织架构	主管机构	Supervisor	必备
		下属机构	Sub	有则必备
	其它描述	历史简介	Description	有则必备
		机构网址	WebsiteURL	有则必备
机构功能实体	机构性质	性质类别（高校、研究所、医疗机构等）	Nature	必备
	机构职能	职能类别（业务部门、管理部门、党政部门等）	Function	有则必备
机构主体	主体类型	主体类别（员工、学生等）	Member	有则必备
	主体数量	主体规模	Num	有则必备
机构行为实体	人才培养	培养类型（本科、硕士、博士、博士后等）	Degree	有则必备

		培养方向（专业、学位类别、研究方向等）	Discipline	有则必备
	科研行为	活动领域	Field	必备
		合作伙伴	Co-partern	有则必备
		竞争对手	Competitor	有则必备
科研成果实体	成果类型	成果类型（论文、基金、专利等）	Type	有则必备
	成果规模	产出规模	OutputNum	有则必备
	影响力	影响力	Impact	有则必备
机构权利实体	权利类型	权利类别	Type	有则必备
	权利内容	具体内容	Text	有则必备
机构协议实体	协议类型	协议类别	Type	有则必备
	协议内容	具体内容	Text	有则必备

3 基于本体模型的科研机构画像标签体系研究

所谓科研机构画像，即机构信息的标签化，利用标签体系勾画机构的属性特征。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，包括特征标签、关系标签等，形成机构标签库，为科研机构标签体系的构建和画像奠定基础。

3.1 科研机构的画像标签内容研究

机构本体中科研机构是由多个相关的实体相互作用，共同来描述限定，因此对科研机构的描述属性是对不同实体的属性语义关联分析基础上抽象总结出来的。通过对各个实体属性和关系的总结凝炼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；基本信息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、发展历程等属性；社会关系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，如由于名称变更、拆分、合并等过程中产生的沿革关系和同一关系，科研成果的合作产生的合作关系以及机构组织架构过程中产生的层级隶属关系；科研活动是对科研行为的描述，包括产生的科研成果、主要活动领域、合作机构以及对标机构等属性。

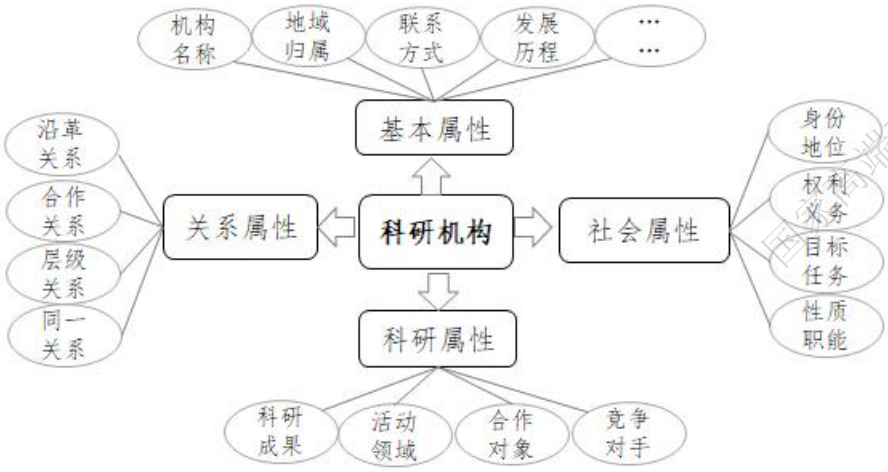

图 2 科研机构的属性特征

科研机构属性的标签化过程就是对机构在从事科研过程中所形成的社会属性、关系和活动进行显性描述的过程。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成三类，包括描述信息标签、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 2 所示，科研机构的基本属性和社会属性在机构官网上均有介绍，相对容易识别和赋予标签，因此本研究重点对科研行为相关的属性和关系进行标签化研究解决科研活动相关的活动领域、合作机构以及对标机构的识别和标注。

表 2 科研机构的画像标签库

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 科研机构属性标签化流程研究

科研机构具有静态属性和动态属性，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的属性，并随着内容扩充和时间推移而变化，比如机构的活动领域、相关机构等。静态属性大部分可以从机构官网上得到，获取方式较为简单，而动态衍生而来的属性和关系则需要基于机构行为、科研成果和已有的静态属性综合推理得到，分析过程相对复

杂，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，从而实现对机构的全面画像，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、发展历程等属性可以通过本地收割或远程采集等方式从已构建的机构规范库、文献及相关成果库和机构官网等来源获取。其次在已标注属性的基础上，通过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以构建不同粒度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 科研机构画像标签化流程

4 科研机构主要属性的标签化方法研究

在科研机构的众多属性中，科研活动领域的标签尤为重要，它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活动领域的学科类别和行业类别为例来论述科研机构属性的标签化过程。目前科研机构的画像、排名和评价研究中多是对一级机构的分析，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，并不能满足从更细学科粒度上进行科研管理的需要，因此本文构建的标签体系主要针对下属二三级机构的特征进行画像，对机构特征的分析更专指、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 科研机构活动领域的标签化方法研究

4.1.1 主要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。机构名称是机构创建时所赋予的，它能标识出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，所以本文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布局新的研究领域，本文将其称为动态学科，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。

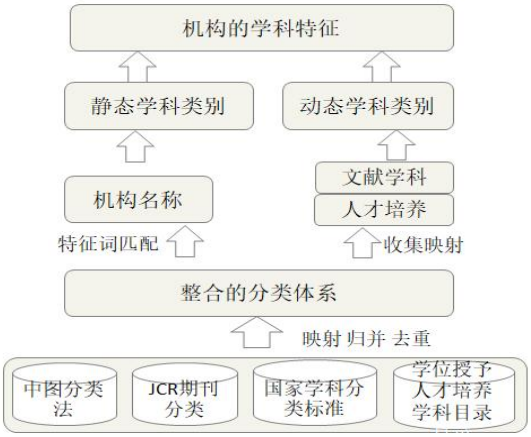

图 4 机构学科特征的标注

对于学科范畴存在多种分类标准和体系，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》、标识文献和期刊等类别的《中国图书分类法》和 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，需要将几种学科体系进行合并融合，由于 WoS 数据库已将 JCR 期刊分类体系与《学位授予和人才培养学科目录》进行了映射[25]，本研究采用去重取并集的方法将《学位授予和人才培养学科目录》、《学科分类与代码国家标准》和《中国图书分类法》三种体系进行映射合并，为了保证标注的学科层次上具有可比性，每种分类体系只取一级和二级学科进行合并。

对于静态学科类别的标注，由于不同机构命名没有特定规则，比如经济领域

的院系名称有经济管理学院、金融学院、财经学院等，因此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 4300 个高被引机构作为种子数据，对其下属的 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，然后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，表 3 列出了部分学科所标注的特征词。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 部分学科类别的学科特征词

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，机构科研成果的学科分布可反映机构关注的领域，揭示机构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，因此本研究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，为了兼顾两种视角的结果，选取它们共同出现的学科作为该机构科研产出的学科属性。对中文文献按照期刊中图分类法的学科类别进行统计，对于英文文献按照 JCR 期刊分类体系进行统计。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、发展策略和所处的研究领域，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。通常认为从机构名称中识别出的静态学科权重更大，更能标识机构的活动领域。随着科学交叉融合，很多科研机构逐步成为综合性的研究机构，涉及多个学科领域，需要综合多种数据来标注机构的学科领域。

4.1.2 主要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。行业是以机构主要从事的经济活动来确定的，当机构只从事一种经济活动时则按照该经济活动确定机构的行业，当机构从事两种以上的经济活动时则按照主要经济活动来确定。对于科研机构来说，专利是最主要的成果转化形式，在一定程度上能体现机构的服务领域，所以本文利用专利领域来表征机构的行业属性。首先统计不同层级机构专利的领域分布，然后利用国际专利分类与国民经济行业分类参照关系表将专利领域映射到国民经济行业分类上。此外，部分机构名称中也有表征所属行业类别的特征词，从机构名称和相关简介信息中抽取表征机构领域或者行业的术语，根据行业分类的说明为不同层级的行业类别标注特征词，将抽取的术语与标注的特征词进行对比，根据结果不断优化比对算法，调整和扩充各层级的特征词。

国民经济行业分类体系较为详细的描述了不同的行业，种子库中 40 多万个不同层级机构的行业类别进行初步统计发现，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同一可比的层级上，本文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 关联机构的标签化方法

除属性之外，科研机构作为社会对象，在参与科研活动过程中与其它机构建立不同的关系，比如层级隶属关系、发展沿革关系、合作关系和引用关系等，在错综复杂的关系网络中识别关联紧密对象也是对科研机构特征进行描述的重要内容。

关联机构是指在与某机构关联比较紧密的机构，主要体现在两机构的科研活

动或科研成果的交互程度以及科研活动或科研成果的领域相似程度，关联机构可帮助识别和推荐已有的或潜在合作伙伴和竞争对手，辅助科研管理和决策。

4.2.1 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。此外在文献、专著和专利数据中，除了作者署名机构字段外，部分还具有基金项目字段，本研究将标识同一基金项目的文献、专著和专利中的科研机构也作为合作机构看待。对于科研机构来说，基金项目资助数量较少，申请难度较大，而且部分项目的申请对合作机构还有一定的限制，因此合作难度较大。专利数量和申请难度次之，文献数量最多，合作概率较大，因此在计算合作强度时为基金项目、专利和文献分别赋予不同的权重，具体见公式 1。

$$\mathrm{Co}_{Intensity=}\sum_{1}^{i}\left(a_{project}*n_{proiect}+a_{patent}*n_{patent}+a_{paper}*n_{paper}\right)$$

公式 1

其中$a_{project}$为项目合作强度权重，a_{patent}为专利合作强度权重，a_{paper}为文献合作强度权重，i 指不同学科领域。分别计算机构与每个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。

4.2.2 对标机构的标签方法

科研机构之间除了明确的层级、合作、引用关系外，在科研评价中也需要识别某个科研实体的对标机构，是指综合实力与本机构水平相当的机构。对标机构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选择，可避免以往研究中只对一级机构进行比较。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，进而实现信息检索系统中相关机构的推荐功能。科研产出通常利用科研成果论文量来衡量，又可细分为 SCI 论文数量、中文核心论文数量、第一作者或通讯作者论文数量；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。关联强度计算公式如下:

$$co_{BenchMark} = \frac{n_{person1}}{n_{person2}} \sum_{1}^{i} \left(\frac{n_{paper1}}{n_{paper2}} + \frac{n_{citation1}}{n_{citation2}} \right)$$

i 指不同学科领域,$\frac{n_{person1}}{n_{person2}}$的比值约接近 1 说明两个机构科研规模约接近，在同一学科领域中$co_{BenchMark}$值的越接近 2，说明两个机构的产出和影响力水平越一致，是对标机构的可能性越大。

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，对标机构的识别是与领域相关的，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 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。

完善的科研机构标签体系还可实现按照某种或多种特征标签对机构进行查询、筛选和关联检索，支撑以机构特征标签为单元的异构科研成果的关联组织分析，提高对机构知识管理的效率，丰富以科研机构为入口检索的多样性，根据标签特征准确定位机构，提升以机构为单位检索的准确度和精确度。此外还支持从多个维度对机构层次关系进行分析，实现机构知识图谱的构建。

参考文献

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

[2]Baldrige Organizational Profile[EB/OL].[2020-08-13].

https://www.nist.gov/baldrige/baldrige-organizational-profile

[3]Lowell Thomas Vizenor.CorporateBeing:A Study In Realist Ontology.2006

[4]Hodgson, G.M.: What are institutions? Journal of Economic Issues.2006, 40(1) :1-24.

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

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

[7]Paul J, Maria B, Owen E. Institution Aware Conceptual Modelling.Proceeding of the ER Forum and the ER 2017 Demo track. 2017.

[8]http://epimorphics.com/public/vocabulary/org.html

[9]https://www.w3.org/TR/vocab-org/

[10]http://www.cs.umd.edu/projects/plus/SHOE/onts/univ1.0.html

[11]Rabab Chakhmoune.Building corporate memories in collaborative wayusing ontologies.3rd International Conference on Next Generation Networks and Services. 2011. DOI:10.1109/NGNS.2011.6142544

[12]Lorenzo PasseriniGlazel. institutionalontologyas an ontology of types. https://www.researchgate.net/publication/277279339_Institutional_Ontology_as_an_Ontology_of_Types

[13]Owen Eriksson, Paul Johannesson, Maria Bergholtz.Institutional ontology for Conceptual Modeling.2018 .https://doi.org/10.1057/s41265-018-0053-2

[14]叶壮壮. 基于 Wikidata 的机构本体构建研究[D]. 2019.

[15]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.图书馆论坛. 2020,44(4)

[16]胡雪环. 科研机构本体的构建方式研究[D]. 2016.

[17]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

[18]Paul J, Maria B, Owen E. Institution Aware Conceptual Modelling.Proceeding of the ER Forum and the ER 2017 Demo track. 2017.

[19] Organization Development Models: A Critical Review and Implications for Creating Learning OrganizationsMaxwell. A. Asumeng and, Judith AnsaaOsae-Larbi

[20]孟琳. 多源信息融合的机构画像的方法研究[D]. 2018.

[21]TanejaG.How are higher education institutions defining their meta-description tags?International Journal of Educational Management. 2018. 32(7):1293-1306. https://doi.org/10.1108/IJEM-08-2017-0201.

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

[23]JuhaKettunen.Stakeholder relationships in higher education, TertiaryEducation and Management.2015,21(1):56-65. DOI: 10.1080/13583883.2014.997277

[24]Hohfeld WN. Some fundamental legal conceptions as applied in judicial reasoning. Yale Law J. 23(1) (1913) 16–59

[25] https://help.incites.clarivate.com