《智能信息处理》课程考试

知识图谱发展现状与展望

米富横

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 15 日

知识图谱发展现状与展望

米富横
（大连海事大学　信息科学与技术学院　大连　116026）

摘　要　为了充分展现国内在垂直知识图谱领域研究的现状，以垂直领域知识图谱为研究对象对其发展现状和趋势进行综述。 对垂直领域知识图谱的定义和分类、架构和关键技术的发展现状进行了详细论述；针对垂直领域知识图谱的具体应用进行了论述，并以学术信息知识图谱和医药卫生知识图谱为例进行了详细介绍。 最后对垂直领域知识图谱发展中存在的问题和对策以及未来的趋势进行了探讨。

关键词　知识图谱；垂直领域知识；知识获取；知识表达

Development status and prospect of vertical domain knowledge graph in China

Mi Fuheng
(Department of Information Science Technology, Dalian Maritime University, Dalian　116026)

Abstract　In order to fully demonstrate the current research status of vertical knowledge graph , his paper took vertical domainknowledge graph as the research object to summarize is development status and trends.Firstly, the paper discussed in detailsome key topics of the vertical domain knowledge graph ,such as the definition and classification , architecture and key technologies. Furthermore , it focused on the specific applications of knowledge graph in vertical field ,and took the academic informationknowledge graph and medical and health knowledge graph as examples.Finally , this paper pointed out the problems and futuretrends in the development of knowledge graph in vertical fields.

Keywords　knowledge graph;vertical domain knowledge;knowledge acquisition;knowledge expression

1 引 言

以互联网为代表的信息科技已经成为经济增长的强劲引擎，也成为了国与国之间竞争的焦点，其技术的强弱直接体现了国家综合国力的强弱。 互联网技术的发展历经了最初的文档互联、到后来的数据互联、再到现在的知识互联，以至于未来的万物互联，这种发展趋势充分体现了人们所认知世界的复杂性和量变引起质变的道理。 随着网络应用的深入，以往的文档或者数据已经无法满足网络智能化的需求，需要从新的角度去组织网络中包含的多源、海量、异构数据。从语义网的概念开始，知识互联成为网络技术发展的方向，而知识图谱则是知识互联发展的必然选择[1]。

知识图谱并不是一个全新的概念，可以看做是语义网络（semantic Web）深化发展的形态，其出现实际上是学术界和工业界对于互联网发展的一种探索的结果。2006年Berners-Lee等人[1]提出了数据链接（linked data）的概念，并呼吁对相关的技术标准进行完善和推广，以推进语义网络的发展。此后有一波语义网络研究和应用的热潮，而知识图谱正是这次技术热潮发展的必然结果，也是对其研究成果的升华与深化。

知识图谱最早由谷歌提出，并率先将其应用于搜索引擎中，取得了不错的效果。在谷歌搜索的示范效应下，国内外的企业界和学术界对知识图谱技术在多个领域的应用

进行了探索，并取得了丰富的成果。目前，知识图谱已经成为智能搜索、智能问答、个性化推荐、医疗卫生、金融安全等领域的关键支撑技术，是学术界和工业界的热点。本文将在上述基础上，聚焦于国内的垂直知识图谱研究领域，对其进行分析讨论。

2 知识图谱的定义与分类

知识图谱本质上是一种将世界实体和实体关系进行相互关联的语义网络，其中的节点表示实体（entity），边则代表实体之间的各种语义关系（relationship）；在学术论文中，则根据应用场景和技术背景等对知识图谱提出了很多不同的定义。总体来看，虽然知识图谱没有统一的定义，但是公认的知识图谱的概念应该包括如下几个基本要素：知识节点（从实际对象抽象而来）边（节点间的关系，由实际关系抽象而来）和对象的数量（节点和边的数量要足够大）[2]。

知识图谱可以从不同的角度将其分为不同的类型，如从构建方法、构建技术、使用方式等。 目前比较常用的分类方法是从应用目标出发，将其分为通用知识图谱和垂直知识图谱。通用知识图谱不面向特定的领域，强调的是知识的广度，包含了大量的常识性知识；而垂直知识图谱则面向特定领域，强调的是知识的深度，包含某个领域的特色知识。 两者之间的区别与联系如表 1 所示。

图 1 通用知识图谱与垂直知识图谱比较

比较项目	通用知识图谱	垂直知识图谱
知识特点	常识性知识，突出知识广度	专业领域知识，突出知识深度
知识来源	来源广泛，主要是语言知识图谱、常识知识图谱、百科知识图谱等	来源相对狭窄，主要依赖专业领域书籍、专利、模型、经验等
应用领域	检索、娱乐、问答等	专业领域的检索、教育培训等
知识获取	自动化程度高，对知识质量要求不高	难以自动获取，对知识质量要求高
表现形式	形式固定，通常以文字、图片、视频等为主	形式依赖于内容，包括但不限于文字、图片、图表、模型（二维、三维）、视频以及特殊格式文件等
图谱受众	面向有知识检索需求的普通人群，受众面广	面向特定领域的专业技术人员，受众面窄
使用难度	几乎没有使用难度，具备基本的常识即可	有一定的使用难度，要求具有特定领域的专业知识
代表性工程	WordNet、HowNet、Cyc、ConceptNet YAGO、FreeBase、Dbpedia、Wikidata、zhishi.me、SSCO 等	SIDER、IMDB、MusicBrainz、电商、企业商情等
构建方法	具有成熟的构建流程，在知识获取、知识融合、知识推理等方面具有较多的研究基础和应用案例	尚无统一成熟的构建流程，在知识获取、知识融合等关键技术领域仍处在探索阶段，但在应用驱动下发展较快

从表 1 中可以看出，垂直领域本身具备知识图谱的所有特点，也应该吸收通用知识图谱的各种技术来促进自身的发展。但是必须看到，由于垂直领域本身的特点，导致垂直领域知识图谱从构建流程到关键技术都与通用知识图谱具有一定的不同，在研究和应用中应该引起重视。

3 知识图谱的架构

知识图谱的架构是构建和应用知识图谱的基础，良好的架构对于知识图谱的性能具有决定性的影响，因此知识图谱架构是进行知识图谱研究的首要问题。 知识谱的架构包括图谱本身的逻辑结构以及构建知识图谱所采用的技术架构[3]。

知识图谱的逻辑结构可以分为模式层和数据层两部分。模式层在数据层之上，存储的是经过提炼的知识，通常采用本体等技术来管理；模式层借助本体库对公理、规则和约束条件的支持能力来规范实体、关系以及实体类型和属性等对象之间的联系。 数据层则主要由一系列的事实组成，在知识图谱的数据层，知识可以用事实为单位进行存储，也可以采用“实体—关系—实体”或者“实体—属性—性值”的三元组作为存储方式。

技术架构主要包含构建流程和构建技术两方面的问题，这里主要讨论图谱的构建流程问题。知识图谱的构建方式可以分为自顶向下和自底向上两种。自顶向下的构建方式从结构化资源出发，通过从资源中抽取本体和模式信息不断地加入到知识库中；自底向上的构建方法则是从公开的资源中采取技术手段获取资源，并对资源进行人工审核后再加入知识库中。

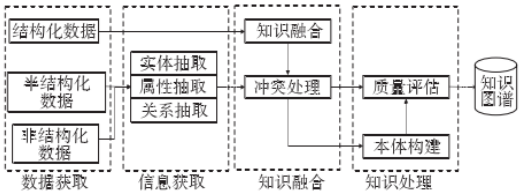
从目前的研究和应用效果来看，采取混合模式（即自顶向下和自底向上相结合）进行构建，并在构建中采取“小步快跑、快速迭代”的方法是较为合理有效的。

4 知识图谱的关键技术

无论图谱的构建模式如何（自顶向下或者自底向上），其中的关键技术都是相同的，图 2 展示了知识图谱中涉及到的关键技术及构建流程。从图 2 中可以看出，知识

抽取、知识融合、知识推理和知识应用等都是知识图谱体系中的关键技术。需要指出的是，垂直领域知识图谱是源于通用知识图谱且根植于特定行业，所以其构建中的关键技术与通用知识图谱相对比，既具有类似性也具有独特的特点。 下面将从垂直领域知识图谱构建的相关关键技术，如知识抽取、知识融合和知识推理等角度论述国内的发展现状与趋势等[4]。

图 2 知识图谱关键技术



知识抽取是指从数据中获取实体、关系和属性等知识，以供知识图谱的构建所用。知识抽取的对象主要是公开的链接数据，通过抽取技术可以获得相关知识的实体、关系以及属性要素。在此基础上，可以进一步地形成知识图谱构建所需的各种知识资源。从数据的结构化程度上可以将链接数据分为结构化数据、半结构化数据和非结构化数据三种，这三种形式的数据在各个领域都是广泛存在的，对于知识图谱的构建来说都是必需的。对于结构化较好的数据，如数据库中的数据等，可以采用 D2R（relationonal database to RDF）映射的方法将其转换为图谱中的数据；对于半结构化的数据，如各种技术报告、表格数据等，采用封装器（wrapper）来抽取，并将其形式化后存储到图谱中；对于非结构化的各种文本数据，则需要采用监督 / 半监督的智能学习算法来抽取数据。 目前研究的重点是面向结构化和半结构化的知识进行自动化抽取，对于非结构化知识的抽取还处于探索状态。

通过知识抽取技术可以获得大量的相关知识，但是这些知识存在着质量不齐、知识间关联不清甚至知识间存在冲突等现象，因此必须对知识进行融合，使其在一个统一的框架下进行整合、加工、消歧等操作，才能获取高质量的知识；然后将相关的知识通过构建本体、评估知识质量等步骤，才能形成高质量的知识库。

常见的知识融合主要包括具体知识的融合和知识库融合两个层面。 具体知识的融合是指对两个或多个具体的知识实例进行融合，消除其中的实体冲突、关系冲突或者指向不明等问题，主要通过实体对齐等来实现；知识库融合则是指将外部的数据库部分或者整体地融合到本地数据库中，以实现知识的高质量、低成本导入，主要通过知识库融合来实现。

根据使用目的的不同，领域知识图谱可以用来执行分类、查询、预测等传统的数据挖掘任务，也可以用来实现问答、智能推荐等智能型任务，上述任务的执行依赖于图谱的推理功能。知识推理指的是从已知知识库的实例出发，通过各种计算得到实体间新的关系的过程。 知识推理一方面可以挖掘隐含的知识，丰富知识库；另一方面可以基于知识库解答问题，是知识图谱体系中的核心功能之一。 常见的知识图谱采用的推理算法包括基于逻辑的推理、基于图的推理和跨知识库的知识推理等。

5 国内垂直领域知识图谱现状

知识图谱作为一种基础性资源，在促进国民经济各个行业的知识化方面具有重要的意义[5]。目前国内对于知识图谱的应用种类繁多，本文在相关文献的基础上，从电商平台、企业信息、科技情报、创业投资、农林科技、医疗卫生、工业应用、影音娱乐等不同领域搜集了部分相关的应用案例，如图 3 所示。

图 3 垂直领域知识图谱应用案例

应用领域	核心问题	应用场景	应用案例
电商平台	消费热点发现、消费偏好分析、仓储物流优化	产品推荐、配送路线优化	阿里巴巴、美团大脑
企业信息	企业关联关系分析、企业数据挖掘、垂直领域知识发现	投资尽职调查、企业风险预警	量子魔镜、天眼查、启信宝、企查查、达观数据、文因互联
科技情报	情报资源内蕴关系挖掘、复杂关系可视化	科学研究热点发现、家谱/合作关系可视化	CiteSpace、上海图书馆家谱服务平台
创业投资	投资风险规避、企业发展潜力评估	风险投资	棱镜平台、智言科技、因果树
农林科技	农林知识与关系抽取、可视化表达	知识查询、科普培训	中国农科院水稻知识图谱
医疗卫生	医学知识获取、形式化表达、领域知识推理	疾病辅助诊断、中医配方辅助分析、医学知识科普	中医药知识图谱、医学知识图谱综述、乳腺肿瘤知识图谱、中医养生知识图谱
军事科学	多源异构信息获取、信息间关系表达、信息推理	战略决策辅助分析、军事知识科普	武器装备、作战体系、战场目标、军事知识
工业应用	领域知识获取、领域知识推理	特定领域知识发现、专业知识培训	明略数据、电磁数据处理、项目管理
文化娱乐	领域知识获取、复杂关系可视化	音乐推荐、领域知识科普	佛学、音乐知识图谱

通过上述分析可以看出，目前国内真正实现落地的知识图谱研究，更多的是集中在可以迅速产生经济效益的领域，如电商、金融、创投等。这一方面说明知识图谱的确可以创造经济效益，也获得了市场的认可，这对于知识图谱的研究发展是非常有利的；另一方面，这会加剧知识图谱人才和研究热点的不均衡，对于一些基础性的、非盈利性的知识图谱研究可能会陷入无人问津的境地，这对于行业的均衡发展是不利的。 需要学术界承担起拓展知识图谱研究领域的责任，也需要国家相关政策的倾斜和支持。

6 垂直领域知识图谱存在的问题与对策

纵观知识图谱发展的历程，其间经历了多次起伏，总体来说，目前处于一个新的高潮期，并在落地方面取得了令人瞩目的成绩；同时也应该注意到，一些瓶颈问题如果不能解决，知识图谱的应用广度和深度将会受到极大的限制，不能达到预期的高度[6]。

从垂直领域知识图谱的技术和应用角度来看，目前存在的问题有两个方面，根基问题（即构建图谱的基础数据）和技术问题（即构建图谱的知识融合、推理和数据安全等）。

解决高质量的数据源问题，目前更多的依赖于行业的发展和技术积累。随着“中国制造 2025”和“两化融合”等政策的实施，我国企业的管理和技术水平都在快速发展，能够提供的数据质量也在不断提升。 可以预见在不远的将来，将出现更多的高质量的数据源以满足垂直领域知识图谱发展的需要。

知识图谱中的知识抽取、知识融合和知识推理等技术在前面已经进行了详细的论述，值得注意的是，作为一门横跨专业领域、计算机工程和管理工程的技术，相关领域的技术对于知识图谱具有非常明显的促进作用，比如图数据库是一种基于图论算法的新型数据库，非常适合处理大量复杂交互多变的数据，属于计算机科学的研究课题。随着垂直领域知识图谱复杂程度的提升，传统的关系型数据库已经难以满足需要，图数据库则很好地满足了知识图谱中对于知识频繁查询的需求。

在数据安全方面，区块链技术是目前呼声最高、也是最有可能满足需求的技术。区块链技术主要是基于分布式存储和共识技术，具有全程留痕、不可窜改等特点，非常适合于保护数据安全。 目前区块链在数字货币方面取得了瞩目的成绩，同时许多研究人员在其他方面也进行了一些应用探索。区块链技术的去中心化和不可窜改的特性非常适合保护垂直领域知识的知识产权，具有广阔的应用前景，也将是下一步研究的热点。在知识图谱的具体技术方面，一方面要善于从相关学科中吸取新的技术和方法促进知识图谱的发展；另一方面还要未雨绸缪，针对未来可能的技术需求提前做好预研，以保证知识图谱技术能够满足经济社会发展的需求。

7 结 论

国民经济的各个垂直领域对于知识图谱的需求都是非常强烈的，知识图谱可以应用的范围也是非常广阔的，因此建设垂直领域知识图谱对于经济社会发展是非常必要的。 目前在垂直领域，知识图谱应用的范围还是较窄，集中在一些可以迅速见到效益的领域（如电商、搜索等），且应用的深度不够，多是集中在图谱的构建上，还未在行业中真正形成使用和建设的良性循环。 在未来，知识图谱应该作为与国家标准一样的基础知识资源进行建设，建立相应的国家、行业规范。 在实际的应用场景下，各个垂直领域更多地关注于本行业的知识资源建设与使用，而不必关心知识图谱的表现形式和底层技术，这样必将能极大地提高各个行业的知识化水平，为我国经济社会发展提供更强的助力。

参 考 文 献

[1] 胡泽文，孙建军，武夷山．国内知识图谱应用研究综述［Ｊ］．图书情报工作，2013,57(3);131-137,84.

[2] 徐增林，盛泳潘，贺丽荣，等．知识图谱技术综述［Ｊ］．电子科技大学学报，2016,45(4);589-606.

[3] 刘峤，李杨，段宏，等．知识图谱构建技术综述［Ｊ］．计算机研究与发展，2016,53(3);582-600.

[4] 阮彤，王梦婕，王昊奋，等．垂直知识图谱的构建与应用研究［Ｊ］. 知识管理论坛,2016(3): 226-234.

[5] 王静，柯登峰，胡茜，等．一种医学知识图谱构建方法及装置：中国，CN 201910599431. 7[P].2019-10-22

[6] 李春华．基于机器学习模型与众包的知识融合方法研究［Ｄ］．苏州：苏州大学，2017.