第一章 知识图谱概述

一、知识图谱定义

（一）什么是知识图谱

尽管早在2012年谷歌便提出了知识图谱的概念，但截至目前，对于知识图谱这一概念尚无官方文档给出具体定义。知识图谱的描述大多还是引用谷歌所撰述的博客，即知识图谱是谷歌公司为增强其搜索引擎功能而开发的知识库，基于知识图谱的搜索引擎服务能够提升用户体验。

如上所述，知识图谱的提出最初是为了提升用户的搜索体验，已极大地改变了传统搜索引擎的模式。通过对查询语句的解析出用户需要查询的实体，捕获和理解用户搜索的需求，进而在准确获得待查实体信息的同时，获取与所查询实体关联的其他信息，从而更全面直观地呈现所查实体及其关联信息。

然而，伴随知识图谱与各领域的不断深入融合，金融、医疗、制造、交通、教育、农业、家居等领域应用快速积累并取得成效。知识图谱的内涵和外延也不断拓展，包括的知识内容和类别也在逐渐丰富，例如事件、规则、行为等。当前知识图谱的主要定义及其来源见表1-1，为了能够覆盖知识图谱的未来发展趋势，本文采用《信息技术 人工智能 知识图谱技术框架》（征求意见稿）中给出的知识图谱定义及相关术语，即：

知识图谱knowledge graph：以结构化形式描述的知识元素及其联系的集合。

其中，知识元素是指描述某一事物或概念的不必再分的独立的知识单位，可分为实体、概念（实体类型）、属性、关系、关系类型、事件、规则（推理逻辑）等。

表1–1 当前知识图谱的主要定义及其来源

序号	定义	来源
1	知识图谱是以结构化形式描述的知识元素及其联系的集合。 注：知识图谱将人类认知的信息转化为机器可读、可处理、可呈现的形式；提供了（图的形式）一种更好地组织、管理和理解海量信息的能力。	国家标准
2	知识图谱：（1）将海量知识及其相互联系组织在一张大图中，用于知识的管理、搜索和服务。（2）特指谷歌公司开发的知识图谱。	《计算机科学技术名词》（第三版）
3	知识图谱是谷歌公司为增强其搜索引擎功能而开发的知识库，基于知识图谱的搜索引擎服务能够提升用户体验。	维基百科
4	在知识表示与推理中，知识图谱是指应用图结构数据模型或本体论进行数据集成的知识库。	
5	知识图谱是以概念及其关系描述知识的数字结构。知识图谱包括可允许人与机器共同理解和推理的本体。	
6	知识图谱是以结构化的形式描述客观世界中概念、实体及其关系，将互联网的信息表达成更接近人类认知世界的形式，提供了一种更好地组织、管理和理解互联网海量信息的能力。	《知识图谱发展报告》（2018）
7	知识图谱作为一种知识表示形式，是一种大规模语义网络，包含实体、概念及其之间的各种语义关系。	《知识图谱：概念与技术》
8	知识图谱本质上是一种语义网络。其结点代表实体（entity）或者概念（concept），边代表实体 / 概念之间的各种语义关系。	百度百科
9	知识图谱（1）在图式（schema）中定义了实体的抽象类别和关系；（2）主要在图中组织并描述了真实世界中的实体及其关系；（3）包括了各种特定领域。	Paulheim, Heiko."Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods"
10	知识图谱是实体及其语义类别、属性和关系的网络。	Krötsch, Markus; Weikum, Gerhard. "Editorial of the Special Issue on Knowledge Graphs"

（二）知识图谱相关术语定义

《信息技术 人工智能 知识图谱技术框架》（征求意见稿）给出的其他知识图谱相关术语及其定义见表1-2。

表1–2 知识图谱相关术语及其定义

序号	术语	定义
1	知识 knowledge	通过学习、实践或探索所获得的认识、判断或技能。 [来源：GB/T 23703.2-2010, 2.1]
2	知识元素 knowledge element	描述某一事物或概念的不必再分的独立的知识单位。 注：可分为实体、概念（实体类型）、属性、关系、关系类型、事件、规则（推理逻辑）等。
3	知识单元 knowledge unit	按照一定关系组织的一组知识元素的集合。
4	本体 ontology	表示了实体、关系、属性类型和关联的一种模型。 注：又称本体模型。
5	图式 schema	本体模型的规范化表达。
6	实体 entity	现实世界中独立存在的对象。 [来源：GB40651-2021, 3.10]
7	属性 attribute	一类对象中所有成员公共的特征。 [来源：GB/T 40216-2021, 3.1.2]
8	实体类型 entity type	一组具有共有属性的实体集合的抽象。
9	事件 event	在某个时刻或时间段内发生或预计发生的事。
10	关系 relation	实体或实体类型间的联系。 注：关系可描述实体类型和实体类型、实体类型和实体、实体和实体之间的关联方式。
11	知识图谱供应方 knowledge graph supplier	使用数据、知识等构建知识图谱以满足特定需求，并提供基于知识图谱的基础工具或服务的组织。 注：基础工具和服务是指可基于其构建复杂的应用程序或系统的中间件。

12	知识图谱集成方 knowledge graph integrator	根据知识应用需求，将知识图谱、信息系统或服务进行整合，提供知识图谱应用系统及服务的组织。
13	知识图谱用户 knowledge graph user	使用知识图谱应用系统及配套服务支持以满足自身需要的组织或个人。 注：知识图谱用户可对外输出必要数据或知识。
14	知识图谱生态系统合作伙伴 knowledge graph ecosystem partner	为知识图谱供应方、集成方和用户提供知识图谱构建和应用所必需的信息基础设施、数据、工具、方法、标准和机制等的组织。
15	知识表示 knowledge representation	利用机器能够识别和处理的符号和方法描述人类在发现或理解客观世界时获得的知识的活动。
16	知识建模 knowledge modeling	构建知识图谱的本体及其形式化表达的活动。 注：知识建模活动可包括实体类型定义、关系定义及属性定义。
17	知识获取 knowledge acquisition	从不同来源和结构的输入数据中提取结构化知识的活动。 注：知识获取的数据源通常按数据组织结构的维度可分为结构化数据、半结构化数据、非结构化数据（如纯文本、音频和视频数据等）。
18	知识融合 knowledge fusion	整合和集成知识单元（集），并形成拥有全局统一知识标识的知识图谱的活动。
19	知识存储 knowledge storage	设计存储架构，并利用软硬件等基础设施对知识进行存储、查询、维护和管理的活动。 注：常见的知识存储方式分为：基于关系数据库的存储方式、基于图数据库的存储方式、基于RDF数据库的存储方式等。
20	知识计算 knowledge computing	基于已构建的知识图谱和算法，发现/获得隐含知识并对外提供知识服务能力的活动。 注：知识计算可分为统计分析、推理计算等。知识的统计分析是对知识图谱蕴含知识结构及其特征的统计与归纳；知识的推理计算是从已有的事实或关系推断出未知的事实或关系，实现知识图谱隐性知识的发现与挖掘。
21	知识溯源 knowledge provenance	在知识图谱全生命周期中追踪原始数据向知识转化的活动。
22	知识演化 knowledge evolution	随本体模型、数据资源等变化产生的新知识对原有知识的补充、更新或重组的活动。 注：通过知识计算得出的补全知识也可触发知识演化。

（三）知识图谱的构成

知识图谱根据包含内容的不同可以划分为本体层和实例层，如图1-1所示。其中，本体层，也称概念层，可包括实体类型、属性、关系类型、规则、约束等本体相关知识元素；实例层，也称实体层，是对本体层的实例化，可包括实体类别所对应的实体及其属性、实体间关系等实体相关知识元素。

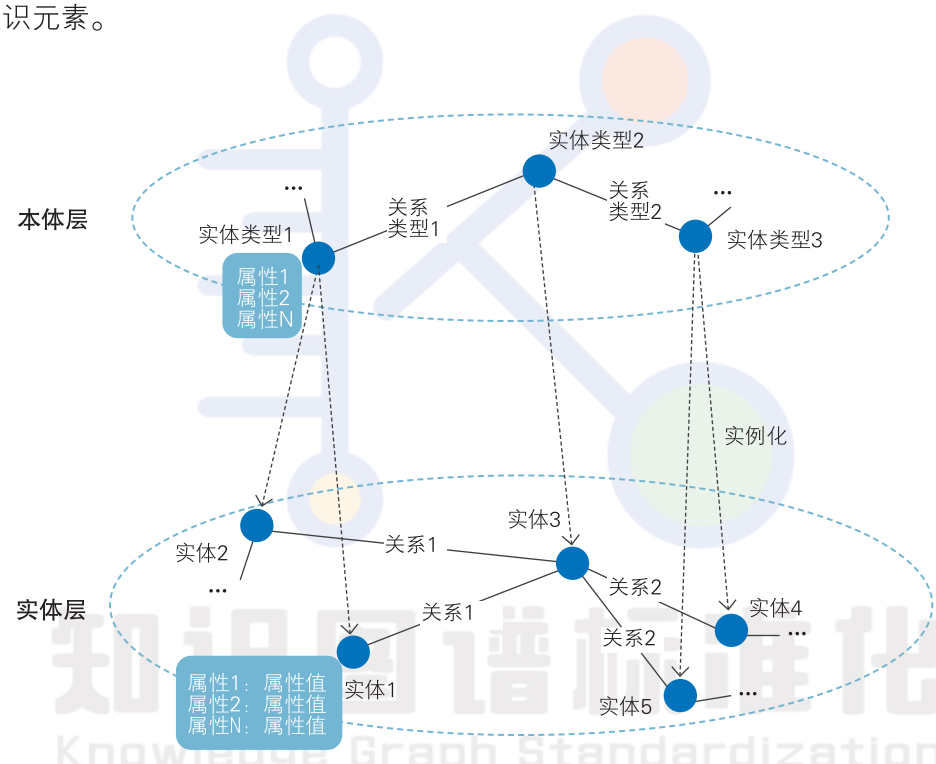

图1-1　知识图谱的构成示意图

二、知识图谱的定位

（一）知识图谱与语义网络、本体等概念的关系

语义网络是用于描述和关联万维网数据的一系列技术标准，使得网络

上的信息具有计算机可以理解的语义。本体是一种规范化和形式化的知识分类体系，通过对概念的严格定义和概念之间的关系来确定概念的精确含义，表示共同认可的、可共享的知识。知识图谱与语义网络、本体概念既有区别也有联系，它们解决的问题域是一致的，在发展历程上各自独立又互有融合，在不同的历史阶段都扮演着重要的角色。

知识图谱较传统语义网络具有如下特点：

1、规模巨大，知识质量高。

知识图谱是大数据时代下的产物，强调对知识的覆盖度，数据规模巨大。通用知识图谱通常会拥有亿级数量的实体与关系，领域知识图谱的实体与关系的数量也会达到百万乃至千万。此外，知识图谱数据来源异构多样，通过使用交叉验证的方式可验证知识图谱中事实的准确性，也可结合相关大数据质量验证算法及众包方式进行事实的验证。基于多维度的质量管控方式，极大地提升了知识图谱中知识的质量。

2、构建和维护的数据结构友好，知识语义丰富。

知识图谱通常以三元组进行表示，这是典型的图结构，同时也可以借助语义网络的RDF表示。面向图数据或RDF数据，已发展出有效的管理办法，使知识图谱的知识表示更灵活、更友好。此外，语义的丰富程度体现在语义种类多和建模多样化等方面，知识图谱通过构建问题域内的各类语义关系，并通过涵盖不同语义关系的知识图谱间融合，可以覆盖问题域内的基本语义关系。此外，在语义关系建模中结合权重或概率设置，还可进一步更精准地表达语义内容。

3、容易工程化，应用范围广。

知识图谱友好的结构、灵活的表示及良好的可视化能力，极大地降低了知识库的构建成本。各行各业可以快速构建行业内知识图谱并为下游应用提供推荐、问答、搜索、预测等智能服务。虽然相较语义网络而言，

其基于逻辑规则的推理能力较弱，但随着深度学习的快速发展，知识图谱链接、补全等基于低维稠密向量的推理技术弥补了知识图谱推理能力的不足，并取得了较好的应用效果。

4、聚焦内容存在差异，领域知识处理能力更强。

知识图谱继承了传统语义网络的部分特性，可以现实世界中的事物为节点，以其间的语义关系为边，构建有向图以形成直观的知识表达。与此同时，知识图谱更加聚焦所包含知识本身的一致性、准确性、完整性等特征，不仅关注知识中语义信息的丰富表达，还关注知识的全生命周期质量，如大规模知识的融合、更新、溯源、组织、管理、维护等，进而获得了更强的领域知识处理能力。

知识图谱较本体具有如下特点：

1、知识图谱既关注实体类型也关注其实例。

本体侧重于表达认知的实体类型框架，表达实体类型、实体类型间关系、约束、规则等。但框架需要与具体实例结合才能完整描述现实世界中的事实，机器也需要同时结合结构化的本体与其实例完成对于现实世界或者某个特定领域事实的理解。

2、关注规模化知识的获取和关联。

本体构建需明确涉及的概念、术语及其关系等，精确性要求较高，往往专家参与构建和审核过程，涉及的数据量级较更小。知识图谱作为传统知识工程在大数据时代的延续，力求知识覆盖的完整性，目的是将事实性知识结构化，不但包括本体中涉及的概念、术语及其关系，还包括具体实体、属性及其关系。因此，知识图谱更加侧重规模化地获取事实性知识。如何在可接受准确性的前提下，自动化高效地获取事实性知识是知识图谱研究的重要领域。

3、关注规模化知识的存储。

本体涉及数据规模较小，可通过树形结构等形式展现，对底层存储的逻辑结构及物理结构要求较低。相比之下，知识图谱对大规模知识的存储有强烈的要求，特别是对其下游应用的支撑，使其不论在逻辑还是在物理层面都需要较高存储性能。

（二）知识图谱与图数据库、传统数据库的关系

图数据库一般泛指采用各种图数据模型进行数据存储和查询的数据库管理系统。图数据模型的要素通常包括顶点（代表实体）和边（代表关系），通过边将顶点连接在一起，顶点和边上可以附加属性信息，图数据模型的数学基础是图论。知识图谱的主要数据表达形式包括属性图和RDF图等，也是常用的图数据模型。

目前大多数图数据库支持属性图；支持RDF图的图数据库则一般称为三元组库。图数据库与图计算引擎不同：图数据库一般用于图数据的联机事务处理（OLTP）任务，包括查询处理与事务管理；图计算引擎主要用于图数据的联系分析处理（OLAP）任务，包括各种图分析算法。图数据库是知识图谱的主要数据管理工具，知识图谱与图数据库的主要关系还表现在：

1、图数据库是实现知识图谱存储的方式之一。

图数据库是实现知识图谱存储的主要方式，但不是唯一方式。知识图谱还可存储在关系数据库、键值数据库、文档数据库或其他NoSQL数据库中。

2、复杂知识图谱有赖于多种存储形式的混合应用。

复杂知识图谱是指除包含基本图结构外还含有本体信息、属性文本、多媒体数据等多模态类型数据的知识图谱。复杂知识图谱往往需要采用图数据库、关系数据库、NoSQL数据库、分布式文件系统等多种存储形式实现存储管理功能。

3、知识获取等其他环节与图数据库协同支撑知识图谱构建及其应用。

图数据库目前主要支持知识图谱的存储管理和查询应用。然而，知识图谱的全生命周期还包括知识获取和知识融合等环节。目前，这些环节的任务需采用其他工具完成。因此，图数据库需与其他工具通过接口功能调用与交互，以完成知识图谱的构建与应用任务。

此外，知识图谱数据管理也可基于传统关系数据库，特别是针对已有遗留系统的知识图谱数据管理任务。目前，包括Oracle、IBM DB2在内的主流商业关系数据库管理系统均具有支持知识图谱数据管理的扩展模块。但，采用关系数据库管理知识图谱的主要问题在于需要进行图数据模型与关系数据模型的转换，影响知识图谱数据管理任务的性能。

（三）知识图谱与人工智能的关系

知识是各行业重要的生产要素，是大量劳动和研究成果的凝练，也是推动人类不断进步的重要基础。构建知识图谱的过程，实现了机器对知识的读取、处理和呈现，为机器具备获取知识、理解知识、应用知识等能力提供了支撑，也为后续机器形成认知能力奠定了基础。同时，也为人类与机器协作实现海量知识的智能化组织、管理、沉淀和再利用给出了一种可选技术路径。

1、知识图谱推动人工智能系统对知识的自动化获取。

自动化知识获取是知识图谱构建的核心之一。早期的知识获取使用人工移植的方法，依靠人工或辅助工具将人的知识移植到机器的知识库中，从而使机器获取知识。随着在机器学习领域的不断突破，人工智能系统使用机器学习等方式在运行过程中，通过学习，获取知识并进行知识积累，对知识库进行完善与更新。

2、知识图谱对可解释人工智能至关重要。

知识图谱实现机器认知智能的两个核心能力：“理解”和“解释”。机器理解数据的方式之一是建立从数据到知识库中实体、概念、关系的映射；机器解释数据的方式之一是利用知识库中实体、概念、关系解释现象的过程，即将知识库中的知识与问题或者数据加以关联的过程。通过知识图谱与机器学习等技术的融合呈现数据训练或推理路径，可以为人类直观利用可视化方式或符号形式解读人工智能系统的行为提供技术支撑。从而，避免深度学习黑盒问题的不可解释性，增加了人工智能系统的可信赖程度。

3、知识图谱推动基于知识的智能应用。

从基于知识的智能应用层面，知识图谱已在关联分析、辅助决策、搜索、问答、推荐等场景中取得成效，细分行业的应用见第2章：

a) **关联分析**

基于构建的知识图谱，可通过最短路径、链路预测、随机游走等图算法，深度挖掘实体间复杂的网络关系。如：通过基于判决书、口供等信息，结合知识图谱平台库构建的走私案件知识图谱，实现一个人与人，人与团伙、走私物、走私渠道方式等关系分析。

b) **辅助决策**

以决策主题为重心，基于知识图谱构建决策主题研究相关知识库、分析模型库和情报研究方法库，建设并不断完善辅助决策系统，可以为决策主题提供多方位、多层次的决策依据，达到辅助决策者的目的。

c) **知识搜索**

通过语义分析技术，准确理解用户查询的真正意图，精准搜索并排序后返回符合用户需求的搜索结果。搜索内容可包括原始文本数据及图谱内容等。目前谷歌、百度搜索引擎中的知识图谱成效印证了本应用的价值。

d) **知识问答**（Knowledge Base Question Answering, KBQA）

通过分析用户自然语言问题的语义，进而在已构建的知识图谱中通过检索、匹配或推理等手段，获取正确答案。知识问答可以回答部分包含较复杂的关联关系的问题，从而为用户提供更精确、简洁的答案。

e) **知识推荐**

基于知识图谱特征和路径，可对相关知识内容进行智能推荐。例如，通过知识图谱中携带的大量属性、关系等特征信息，构建人员画像，对相关的知识进行推荐。同时，基于知识图谱中的路径查询，结合关系分析能力发现显性或隐性关联知识进行推荐。

4、知识图谱促进人工智能系统与大数据、数字孪生、物联网等技术的融合发展。

随着移动互联网的发展，万物互联成为可能，互联所产生的数据也在爆发式地增长，数据的种类和关系繁多，数据要素的深度应用需求旺盛，大数据、数字孪生、物联网等技术是数据要素流通中的重要技术。人工智能系统中知识图谱的构建可以提供更为灵活的数据接入和深度处理能力，进而降低了人工智能系统与现有大数据、数字孪生、物联网相关系统的集成难度。

以数字孪生为例，其以数字化的方式构建物理实体的虚拟模型，并对物理实体在现实环境中的行为进行模拟，借助虚实交互反馈、数据融合分析、决策迭代优化等手段，扩展物理实体的能力，为客户提供更实时、更智能的服务。复杂物理实体的数字孪生建设，涉及大量动态实体、实体间的关系及其多维度的指标属性，通过知识图谱保存动态的多维度指标状态，可以有效促进数字孪生系统的动态特性。

传统物联网通过有线和无线网络，实现物-物、人-物之间的相互连接。围绕当前物-物、人-物、物-人、人-物-服务之间的连接和数据互通需求，知识图谱在物联网数据接入、管理、分析等方面可以为物联网相关系

统提供支撑，进一步提升其数据链接能力。例如，阿里云将电力领域设备说明、操作规程等复杂技术文档，应用知识图谱支持操作人员快速进行操作查询、故障诊断、维修指导、业务学习，同时也方便业务文档的管理、迭代、沉淀、传递，知识图谱逐渐成为电力领域专业知识管理应用的基石。华为云基于油气勘探开发过程中会产生多种形式的海量数据，有效聚合这些多源异构数据，促进油气行业实现数字化和智能化转型。勘探知识图谱可以提供丰富的油气应用，例如：语义搜索、油藏类比、油气知识推荐，支撑油气勘探开发增储上产、降本增效。

第二章 知识图谱发展现状

一、知识图谱发展历程

（一）知识图谱类别

根据知识范围及应用范围的不同，可将知识图谱分为通用知识图谱和领域知识图谱。

1、定义

通用知识图谱：包含多领域的开放知识体系并面向通用应用场景，主要用来解决科普类、常识类问题。

领域知识图谱：包含特定领域范围的知识并面向一个或者多个领域应用场景，主要用来解决特定行业或细分领域的专业问题，通常也称为行业知识图谱、专业知识图谱、垂直知识图谱。

2、通用知识图谱和领域知识图谱对比

通用知识图谱和领域知识图谱在知识表示、知识建模、知识获取、知识融合、知识应用以及知识运维等环节存在较大差异，表2-1对知识图谱构建中各个环节差异进行了对比描述。

表2-1 通用知识图谱与领域知识图谱对比

知识环节	比较维度	通用知识图谱	领域知识图谱
知识表示	知识边界	宽	窄
	知识规模	相对较大	相对较小
	知识深度	相对较浅	相对较深
	知识粒度	相对较粗	相对较细

知识环节	比较维度	通用知识图谱	领域知识图谱
知识获取	质量要求	相对较低	相对较高
	专家参与	轻度	重度
	自动化程度	相对较高	相对较低
知识建模	构建方式	自底向上	自顶向下
	专家参与	轻度	重度
	知识粒度	轻度	相对较细
知识融合	专家参与	相对较低	重度
	自动化模型参与类型	相对较多	相对较低
	知识关联度	相对较低	相对较高
	质量要求	相对较低	精细
知识存储	知识粒度	粗	细
	存储规模	相对较大	相对较小
知识应用	知识粒度	相对较粗	相对较细
	推理链条	短	长
	推理的全面性	更概况	更精细全面
	应用复杂性	简单	复杂
	查询效率	相对较低	相对较高
知识运维	质量管控	相对较低	相对较高
	运维频率	相对稳定	视场景而定
	专业性	相对较低	相对较高
	知识反馈及修正机制	要求低	要求高
知识交换	共享能力	相对较高	相对较低

知识表示环节的差异主要体现在边界、规模、深度及粒度这四个维度，通用知识图谱的边界及规模明显大于领域知识图谱，而领域知识图谱的概念层级体系表现更深，对细粒度的知识表示需求更加强烈。在知识建模环节，相比通用知识图谱，领域知识图谱通常采用自顶向下的建模方式，需业务专家深度参与本体建模，本体层的知识体系定义也相对要求更加精细。在知识获取环节，领域知识图谱（如医疗、司法等）对知识质量有着较为严苛的要求，业务专家参与相对较高，较多人工干预也决定了其自动化构建程度相对较低。

在知识融合环节，领域知识图谱由于知识覆盖紧密具有更高的关联度，通常对融合质量也要求更高，而且需业务专家深度参与，也决定了算法模型的自动化参与程度较低。在知识存储环节，领域知识图谱的存储粒度更细，而通用知识图谱的存储规模相对更大。

在知识应用环节，领域知识图谱的应用对知识要求更加精细。而且，由于知识密集覆盖及计算复杂的特点，推理链条一般更长，领域应用往往会涉及复杂查询，但因其知识规模通常较小而具有较高的查询效率。相反，通用知识图谱的查询多为一到两步的邻居查询，应用相对简单；由于知识规模巨大而导致相对较低的查询效率。

在知识运维环节，通用知识图谱对知识质量管控要求相对较低，并需较高的运维频率。然而，领域知识图谱对运维人员的专业性要求相对更高，需面向不同细分场景的领域应用形成了良好的知识反馈与修正机制，运维频率则需根据场景而定。在知识交换环节，通用知识图谱通常可为各种领域知识图谱提供通用知识支持，要求具有相对较好的共用共享能力。

3、通用知识图谱和领域知识图谱关系

观点1：通用知识图谱对领域知识图谱建设具有支撑作用

许多领域知识图谱建立在通用知识图谱之上，通用知识图谱对领域

知识图谱建设的支撑作用显著。一方面，通过聚合通用知识图谱的描述模板，可形成领域知识图谱的基本领域图式(Schema)；另一方面，通用知识图谱可为领域知识图谱提供高质量的种子事实，这些种子事实可有助于领域知识的发现与挖掘，并可用作样本指导领域知识抽取模型的训练。

观点2：通用知识图谱与领域知识图谱可进行相互补充与完善

知识图谱预设的领域知识边界通常难以完全满足领域应用的知识需求。由于边界外事物与预设边界内事物间可能存在关联，导致沿着某一实体进行关联分析容易超出预先设定的知识边界。通用知识图谱可通过一定技术手段融合领域知识图谱，从而增加通用知识图谱的深度，更好的服务场景；领域知识图谱也可选择性的融入通用知识图谱提升知识的规模，提高应用效果。通用知识图谱与领域知识图谱两者相辅相成，相互促进，利用通用知识图谱的广度结合领域知识图谱的深度，可以形成更加完善的知识图谱。

（二）知识图谱构建技术路径

知识图谱构建技术路径主要包括：知识表示、知识建模、知识获取、知识融合、知识存储、知识计算、知识溯源、知识演化及质量评估等，其技术路径流程见图2-1。

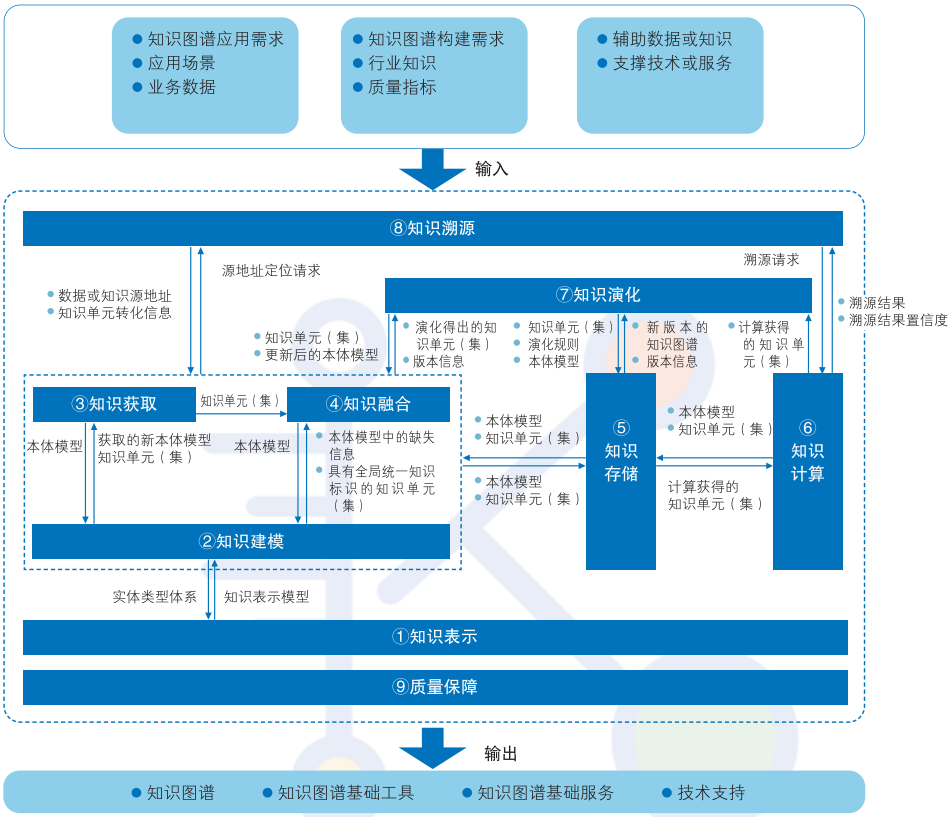

图2–1　知识图谱构建技术路径图

知识图谱的高质量构建既有赖于知识表示、知识建模、知识获取、知识融合、知识存储、知识计算等核心环节的协同工作，同时也依赖于知识溯源、知识演化、质量评估等支撑环节的保障。

其中，知识图谱构建所需的前端输入包括应用需求、数据、专家知识、行业知识、质量指标、支撑技术与服务，以及安全、监管、测评要求等内容；知识图谱构建所形成的知识图谱产品或服务将用于语义搜索、演化分析、知识问答、对话理解等通用知识图谱应用，以及智慧金融、智慧医疗、智能制造等领域知识图谱应用。

针对前端输入，应用需求主要用于明确所形成知识图谱产品或服务的

整体架构、应用方向、应用场景和验收考核指标等内容；数据包括模型基础训练的测试数据和业务数据等，主要用于支持知识表示学习、知识获取等环节算法模型的设计、训练测试以及后续知识图谱的构建环节；专家知识、行业知识主要用于支持知识建模、知识融合、知识计算等环节架构、算法和实现途径的设计、开发与验证；质量指标主要用于评估和控制知识图谱构建过程中各环节的质量高低，以满足应用需求；支撑技术与服务主要用于支持各环节实现过程中所需自然语言处理、机器学习、大数据等技术的融合与应用；安全、监管、测评要求等内容主要用于第三方管理或认证测评机构对知识图谱构建过程及最终输出产品或服务的质量监督等。

根据《信息技术 人工智能 知识图谱技术框架》国家标准草案，知识图谱构建流程中各活动初步定义和任务组成描述如表2-2所述。

表2–2 知识图谱构建流程中各活动初步定义和任务组成描述

活动	序号	任务	备注
知识表示	1	定义知识表示需求	知识表示在人工智能的构建中具有关键作用，通过适当的方式表示人类知识，形成尽可能全面的知识表达，使机器能够通过学习这些知识，表现出类似于人类的行为。
	2	定义或确定拟遵循的规则、约束	
	3	定义或选择知识表示形式	
	4	定义和序列化知识表示元素	
	5	定义知识表示模型适用范围	
	6	定义知识表示模型质量评价体系	
	7	评估并明确知识表示模型的表达能力	

活动	序号	任务	备注
知识建模	8	确定知识的领域和范畴	知识建模是知识图谱构建的基础之一。高质量的知识模型能避免许多不必要、重复性的知识获取工作，有效提高知识图谱构建的效率，降低领域知识融合的成本。
	9	确定现有可复用本体模型	
	10	确定知识范畴内的关键术语	
	11	构建实体类别层级体系	
	12	定义实体类别的属性与关系	
	13	确定并创建本体模型及图式	
	14	评估本体模型质量	
知识获取	15	实体类型提取	获取的数据源根据数据组织结构的维度可分为结构化数据、半结构化数据（网页数据等）、非结构化数据（如纯文本、音频和视频数据等）。
	16	实体提取	
	17	关系提取	
	18	属性提取	
	19	事件提取	
	20	人工录入	
知识融合	21	本体对齐	本体对齐可包括实体类型对齐、关系对齐和属性对齐等。
	22	实体链接	
	23	实体对齐	
	24	知识一致性校验	
知识存储	25	完成数据库选型与数据库设计	知识存储方式及其质量直接影响到知识计算及知识演化的效率。
	26	执行存储操作： 1）完成知识单元的存储； 2）完成知识单元的查询； 3）完成知识单元的维护，如新增、删除、修改、更新等； 4）完成知识单元的可视化（可选）	
	27	完成存储管理	

活动	序号	任务	备注
知识计算	28	定义知识计算需求	知识计算可分为统计分析、推理计算等。知识的统计分析是对知识图谱蕴含知识结构及其特征的统计与归纳；知识的推理计算是从中已有的事实或关系推断隐性知识。
	29	设计计算所需的数据结构及算法模型	
	30	执行知识计算流程并评估计算性能	
	31	基于挖掘的隐性知识补全缺失的知识单元	
	32	通过接口等形式提供知识计算服务	
知识溯源	33	定义知识源地址的结构化描述	知识溯源活动的输出主要包括：知识源地址，即标识知识来源的唯一的资源访问地址；知识单元转化路径，即从源到目标知识的转变关系等。
	34	设计知识溯源方案	
	35	追溯知识来源，并完成知识定位	
	36	确定置信度计算维度并完成溯源结果计算	
	37	对原始数据和知识的更新进行版本管理	
知识演化	38	确认已有知识图谱中待变更的知识单元	知识演化的目标是提高知识质量，丰富知识语义信息，优化知识组织。通过知识计算得出的补全知识也可触发知识演化。
	39	设计或匹配演化的规则、算法或模型	
	40	更新知识单元	
	41	生成具有更新时间等信息的新版本知识图谱	
	42	完成知识图谱版本管理	

（三）国内外知识图谱发展对比

根据技术发展和产业化深度的差异，知识图谱主要经过了三个阶段：萌芽阶段、稳定阶段和繁荣发展阶段。通用知识图谱和领域知识图谱由于应用方向和要求不同，其发展路径存在一定差异。此外基于领域知识图谱演进情况，知识图谱的繁荣发展阶段可进一步细分为产业化探索阶段和产业化深度融合阶段，如图2-2所示。知识图谱相关专利发展趋势如图2-4所

示。

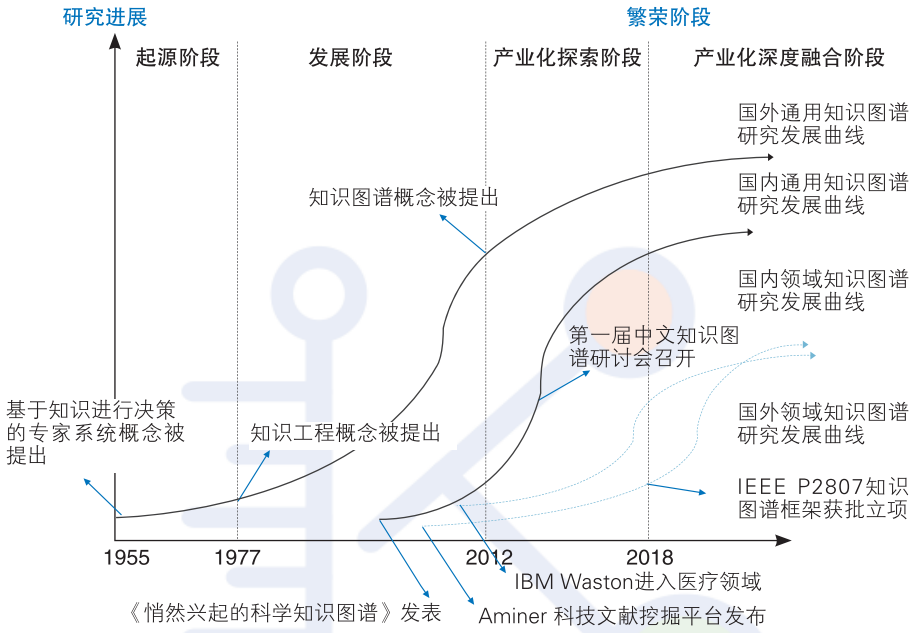

图 2–2 国内外知识图谱发展历程

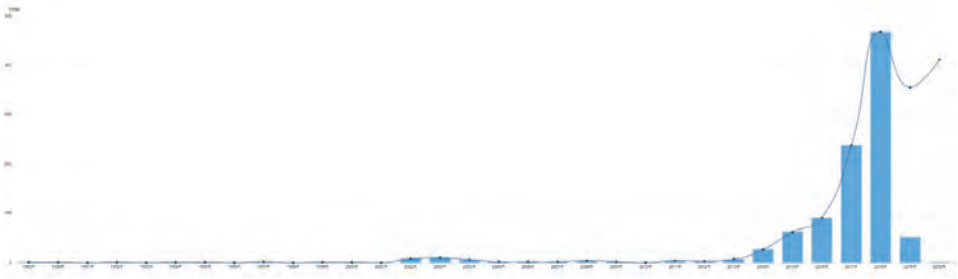

图 2–3 知识图谱相关专利发展趋势 [来源：Aminer]

国外知识图谱发展的具体情况如下：

1) **萌芽阶段**（1955-1977）：

Garfield于1955年开创了利用文献索引进行检索文献的思路；1965年，斯坦福大学的E.A.Feigenbaum提出专家系统的概念，基于知识进行

决策，使人工智能的研究从推理算法主导转变为知识主导；1968年，M.R.Quilian提出语义网络的知识表达模式，用相互连接的节点和边来表示知识，语义网络在人工智能领域被普遍应用。

2) **稳定阶段**（1977-2012）：

在第五届国际人工智能联合会议上，E.A.Feigenbaum提出了知识工程的概念，以知识为处理对象，研究如何用计算机表示知识，进行问题的求解；1989年，Tim Berners-Lee发明了万维网，并于1998年提出了语义网，将传统人工智能的发展与万维网结合，用RDF进行知识的表示和推理；2007年，Chris Bizer and Richard Cyganiak向W3C SWEO提交Linked Open Data Project项目申请，提出推动建设计算机能理解的语义数据网。

3) **繁荣发展阶段**（2012年-至今）：

2012年，谷歌提出“知识图谱”的概念，不同于传统专家系统和知识工程主要依靠手工获取知识的方式，知识图谱以RDF三元组和属性图表示知识，数据规模巨大，可应用机器学习、自然语言处理等技术进行自动化的知识图谱构建。2012年谷歌知识图谱项目发布时，其所拥有的实体已达5.7亿个，由语义关系连接起来的事实描述达18亿条。自知识图谱概念之后，学术界和产业界在构建各类结构化知识库方面积极投入，并以链接开放数据（Linked Open Data, LOD）项目的形式发布和共享了大量的RDF数据。截至2020年5月，LOD项目已包含了1260个数据集及相互之间的16187条连接，涉及到社会生产、生活的诸多领域。

根据当前知识图谱相关论文发表和引用情况来看，该领域论文较多的国家包括中国、美国、德国、意大利、英国、法国、印度等，如图2-4所示。

图 2-4 各国论文发表和引用情况[来源：Aminer]

国内知识图谱的研究晚于国外，但是其发展的趋势国内外基本一致，国内知识图谱演进的具体情况如下：

1) **萌芽阶段**（2005-2010）：

2005年，国内第一篇与知识图谱相关的文献——《悄然兴起的科学知识图谱》被发表，是我国知识图谱的开篇之作。该阶段主要是侧重于知识图谱的理论研究，陆汝钤院士、史忠植、董振东、王克宏等专家学者均为我国知识工程的发展做出了重要贡献。

2) **稳定阶段**（2010-2015）：

知识图谱开始被多个学科领域的专家学者关注，2012年，国内上线了首个关于搜索引擎的中文知识图谱“知立方”；2013年，第一届中文知识图谱研讨会在苏州大学召开，探讨了中文知识图谱的构建技术与策略等核心问题。这一时期，余菜花在中国低碳研究领域，程赛琰在电子政务领域，谢靖在文学学科领域，李伟平在体育领域，辛伟在军事心理学领域利用知识图谱对学科进行研究，并产生了系列研究成果。

3) **繁荣发展阶段**（2015-至今）：

该阶段知识图谱的应用领域越发广泛，论文数量大幅增加，将知识

图谱研究推向高峰。同时，伴随着人工智能、大数据、深度学习的快速发展，知识图谱取得了更深层次的研究进展，推动我国知识图谱正处于一个飞跃式发展阶段。

目前，国内外多个研究机构建立了一些大规模的通用知识图谱。在国外，经典的通用知识图谱有DBpedia、YAGO、Freebase等；在国内，经典的通用知识图谱有百度知心、搜狗知立方、XLORE、zhishi.me等，国内外具有代表的通用知识图谱对比如表2-3所示。国内外具有代表性的领域知识图谱对比如表2-4所示：

表2–3 国内外部分通用知识图谱对比表

	通用知识图谱	数据源	实体数量	关系数量	描述	应用
国外	DBpedia	Wikipedia	458 万	30 亿	大规模跨语言的知识库，支持多达 125 种语言	语义标注，跨域共享服务
	Freebase	Wikipedia,NNDB, MusicBrainz	6800 万	10 亿	大规模开放结构数据集	
	WikiData	Wikipedia, 用户编辑	453.6 万		开放，协作，结构化，支持 281 种语言	语义搜索
	WordNet	专家人工构建	15 万	20 万	人工编辑，按词义组织	词义消歧，语义搜索
	YAGO	Wikipedi,WordNet GeoNames	1000 万	1.8 亿	大规模跨语言的语义知识库	
	Google Knowledge Vault	Wikipedia, Freebase	5 亿	180 亿	大规模知识库	语义搜索
	BabelNet	Wikipedia, WordNet	606 万	19 亿	多语言词典知识库，覆盖 1400 万同义词集合 ,7 亿个词义	多语词义消歧，计算语义相关性
	Microsoft Concept Graph	Web 网页	1255 万	8760 万	以概念层次体系结构为中心的知识图谱	

	通用知识图谱	数据源	实体数量	关系数量	描述	应用
国内	HowNet	专家人工构建	1.1 万		人工编辑，小规模、常用知识库	语义倾向计算
	THUOCL	主流网站的社会标签，搜索热词等	15.7 万		开放的中文词库	
	OpenKG	网页、百科、核心词汇	3000 万	10 亿	通过网络不断获取知识，存储其他开放知识库中有用知识	
	CN-DBpedia	中文百科类网站的纯文本网页	1686 万	2228 万	大规模通用领域百科知识图谱	自然语言问答
	XLore	中、英文维基百科，百度百科，互动百科	1.49 亿	51 万	中英文双语的百科知识图谱	
	知立方	百科类知识，Web 网页				语义搜索
	大词林	百度百科，同义词词林，Web 网页			包含同义、同类上下位关系的动态层级知识体系	

表2-4 国内外部分领域知识图谱对比表

	领域	知识图谱	研究机构	描述
国外	生物医学	Linked Life Data	保加利亚 Ontotext 公司与 LarKC 项目	超 100 亿三元组，包含基因知识库、蛋白质知识库、疾病的知识库
		UMLS	美国国家医学图书馆	一体化医学语言系统，可一体化检索病案记录、书目数据库、事实数据库以及专家系统中的电子式生物医学情报
		Bio2RDF	加拿大基因组、魁北克基因组	构建生命科学领域数据库的关联数据网络
	影视	Linked Movie Dataset		61.5 万三元组数据，描述关于演员、电影等知识
	社交	FOAF	Libby Miller and Dan Brickley	通过构建一个机器可读的本体，描述了人与人之间的关系
国内	医学	中文症状库	华东理工大学	包含症状实体和症状相关三元组数据集
		中医药学语义网络	中国中医科学院中医药信息研究所	以中医药学语言系统为骨架，将中医药领域现有的术语资源和数据库资源融合起来，构成大规模知识图谱
		中医药知识图谱		由中医医案知识图谱、中医特色诊疗技术知识图谱、中医美容知识图谱等多个中医中医美容知识图谱等多个中医相关知识图谱和中医药学语言系统组成
		疾病术语集	开放医疗与健康联盟	包含疾病实体及疾病相关同义词，术语集的数据来源于国内权威的临床医学术语网站

领域	知识图谱	研究机构	描述
科研	SciKG	清华大学	展现计算机科学领域的发展，实现计算机领域中专家与论文的搜索和推荐
影视	影视双语知识图谱	清华大学	融合 LinkedIMDB、百度百科、豆瓣等数据源
交通	基于 CNSchema 的城市知识图谱	浙江大学	涵盖上海市公交站点、地铁站点的静态数据、事件流动态数据

以Wikidata为例，其是由Wikimedia基金托管的协作编辑的多语言知识图谱，也被称为面向文档的数据库，主要存储表示各种主题、概念或物体的实体信息，并为每个实体分配一个唯一的识别码“QID”。截止2021年8月，Wikidata存储了共计九千四百多万条有效的结构化数据，其中每条信息的结构见图2-5。如果属性对应的值在该值中存在，即该值是一个实体，那么该值将会被分配一个QID。

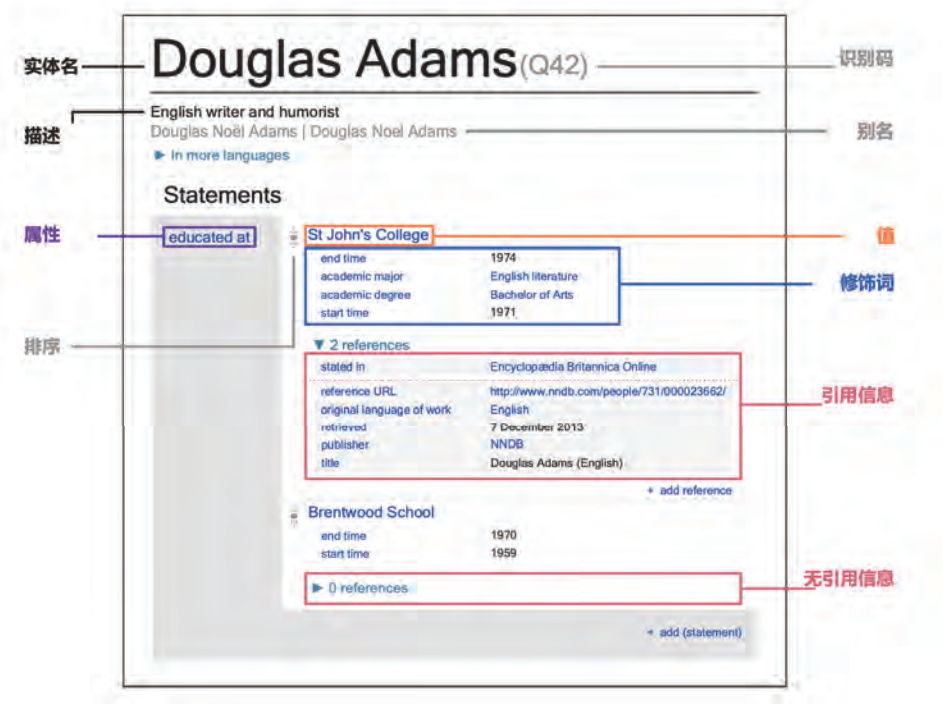

图2-5 Wikidata信息结构

（四）知识图谱中部分关键技术发展历程

知识图谱涉及知识抽取、知识表示、知识融合、知识推理等关键技术，其发展演化过程及研究现状如下：

（1）知识抽取

知识抽取是知识图谱构建的首要步骤，通过自动化或半自动化的知识抽取技术，从原始数据中获取实体、关系以及属性等可用的知识元素，为知识图谱的构建提供基础支撑。图2-6、图2-7分别展示了知识抽取技术演进历程以及该技术相关的专利数统计。

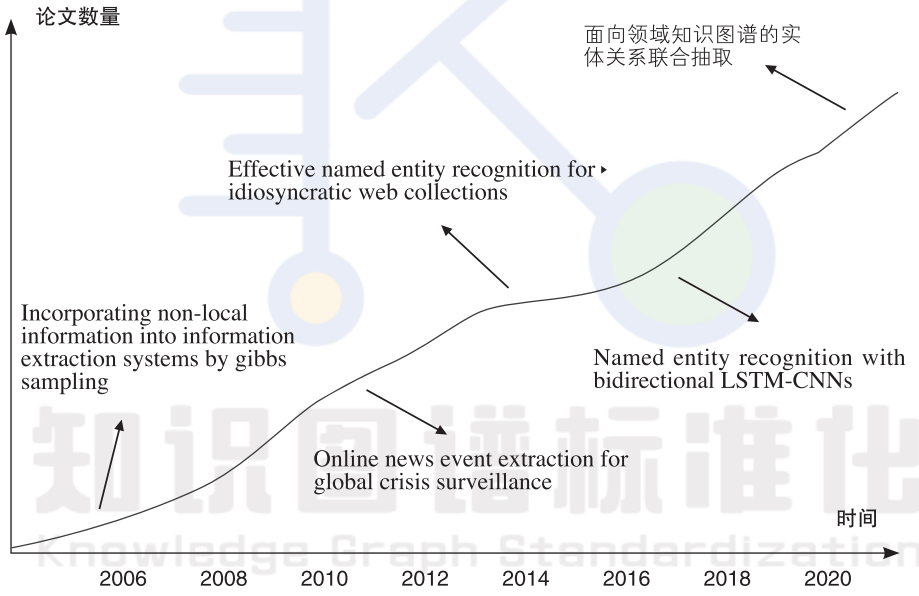

图2–6 知识抽取技术演进历程

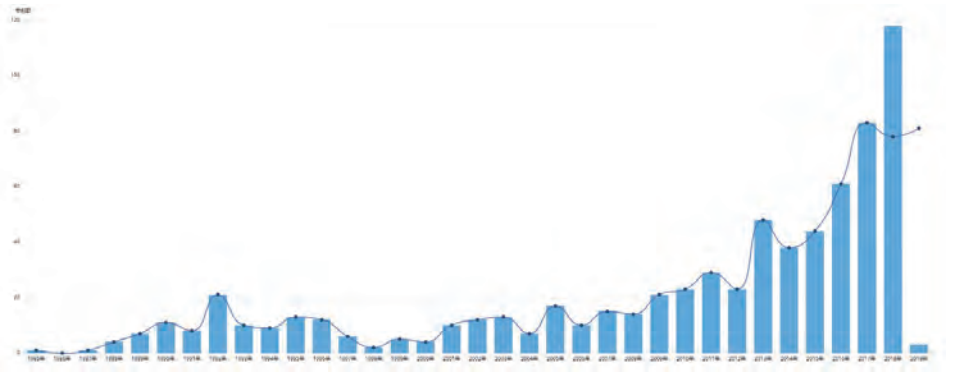

图2–7 知识抽取技术相关的专利数统计 [来源：Aminer]

知识抽取主要包括：实体抽取、关系抽取和属性抽取等任务。早期的知识抽取采用基于规则的方法，根据领域专家定义的规则抽取知识。早期的知识抽取的可移植性差，当数据规模较大时，构建规则需要消耗大量的人力；基于特征工程的知识抽取通常与机器学习相结合，如隐马尔可夫、条件随机场、最大熵等。但是，以上方法需要通过特征工程构建合适的特征，特征工程依赖于大量的人力以及一些NLP工具。

当前知识抽取主要采用基于神经网络的方法，神经网络在特征抽取上具有显著的优异性，通常采用预训练语言模型如BERT等进行编码，在预训练语言模型上叠加RNN、CNN或采用注意力机制等实现抽取。实体关系联合抽取的方法通过建立统一的模型同时抽取实体、实体类型以及实体之间特定的关系类型，根据数据的不同特征，可以采取序列标注、指针网络等端到端的方法实现。

在知识抽取技术发展过程中，国外Miwa等首先将神经网络运用到关系抽取中，Katiyar等在关系抽取方面引入了注意力机制。国内，在医疗、农业和管理等领域都有众多基于上述技术的知识抽取案例。如：在实体命名识别方面，陈德鑫等将在线医疗文本中的实体抽取任务看作序列标注问题，通过将CNN模型和Bi-LSTM相结合来实现对医疗实体的抽取；赖英旭

等采用 CRF 技术对水稻育种专利文本进行了命名实体识别。在关系抽取方面，针对不同的领域综合采用多种深度学习机制进行研究。连明杰等公布了一项基于深度学习模型的关系抽取专利，其核心是采用 BERT模型与 BiLSTM 模型文本进行关系抽取，提取关系三元组；吴文涛等将实体和事件抽取看作一个统一的任务来处理，提出了一种混合神经网络模型，对两者之间的依赖关系进行抽取。

（2）知识表示

知识表示是对知识的一种描述，目的是将客观世界的各类知识表示成计算机容易识别理解的形式。传统知识表示方法包括谓词逻辑、产生式、语义网络、脚本、本体等。在知识表示的演进过程中，最主要的变换是由基于数理逻辑的知识表示过渡到基于向量空间学习到的分布式知识表示。图2-8、图2-9分别展示了知识表示技术演进历程及该技术相关的专利数统计。

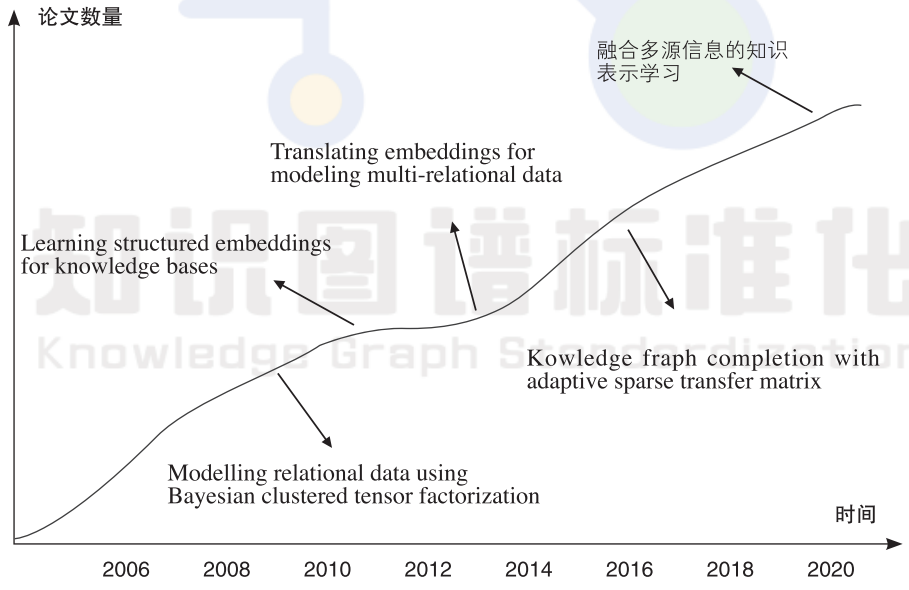

图2–8 知识表示技术演进历程

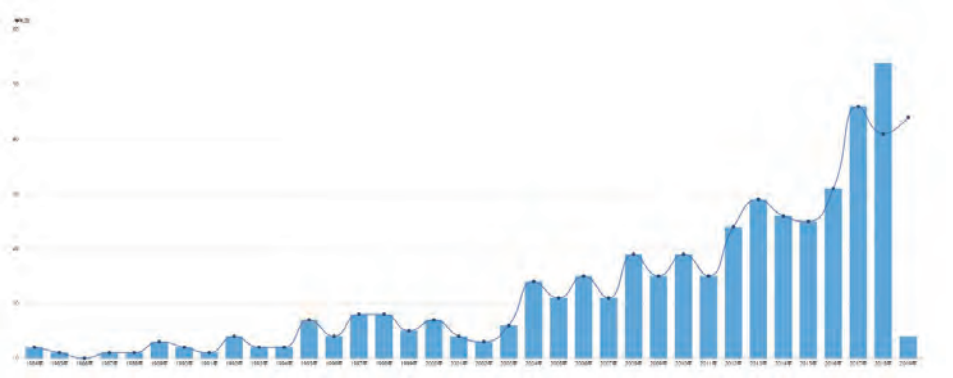

图2-9 知识表示技术相关的专利数统计 [来源：Aminer]

早期的知识表示方法主要有一阶逻辑、霍恩逻辑、产生式规则等。这类方法结构性强，可以准确表达知识，但灵活性差，难以扩展。随着语义网的提出，产生许多面向语义网知识表示的标准语言，具有代表性的包括XML、RDF、OWL等。这种基于符号逻辑的方法，能刻画显式、离散的知识，但是在计算效率、数据稀疏性等方面存在问题，且难以有效挖掘知识中隐藏的语义信息。近年来，随着深度学习的崛起，基于深度学习的特征表示在语音识别、图像分析和自然语言处理领域得到了广泛关注。这类方法通过将知识图谱中的实体和关系嵌入到低维连续的向量空间来完成语义计算，可以有效地挖掘隐藏知识，高效地计算实体间的复杂语义关系，是当前的主流方法。

在知识表示技术发展中，国外典型的知识表示模型有距离模型、单层神经网络模型、双线性模型、TransE模型等。TransE模型将知识库中实体之间的关系看成是从实体间的某种平移，并用向量表示；关系r可以看作是从头实体向量 h到尾实体向量t 的平移。对于知识库中的任意关系三元组<h, r, t >，Trans E都希望满足以下关系：。国内在后续的研究中，考虑到知识库中存在1-N、N-1、N-N的复杂关系，对TransE模型进行扩展，代表模型有TransH、TransR、TransD等，均取得了不错的效果。

（3）知识融合

知识融合是融合本体层和实例层的知识，包括不同知识库的同一实体、多个不同的知识图谱、多源异构的外部知识等，并实现对现有知识图谱的更新。图2-10、2-11分别展示了知识融合技术演进历程及该技术相关的专利数统计。

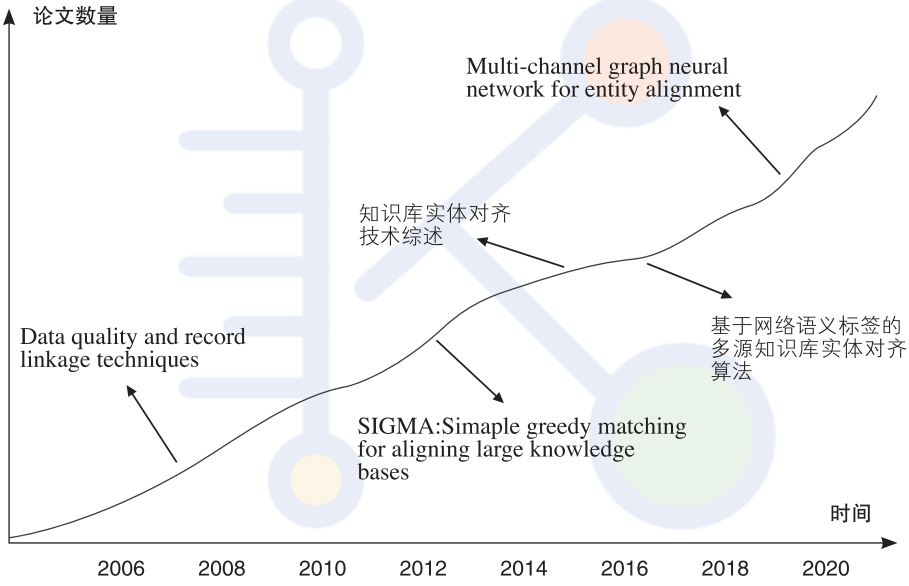

图2–10 知识融合技术演进历程

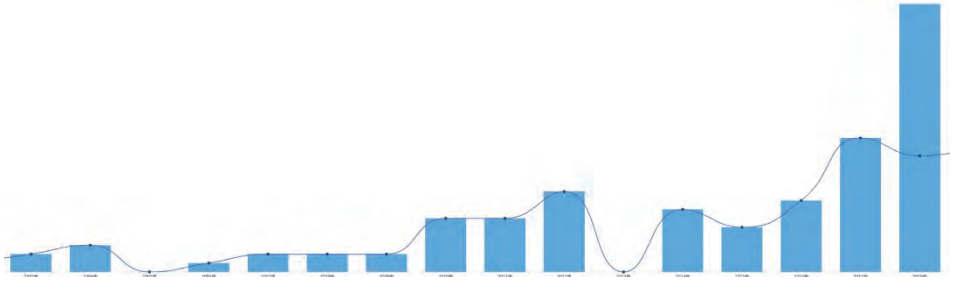

图2–11 知识融合技术相关的专利数统计 [来源：Aminer]

知识融合主要包括本体对齐、实体对齐和实体消歧等。以实体对齐和实体消歧为例：

（1）实体对齐在于发现不同知识图谱中表示相同语义的实体。早期采用机器学习的方法，根据不同实体的相似度得分来进行分类，代表的模型有马尔可夫逻辑网络、支持向量机等，通过比较特征向量进行实体对齐，或者考虑实体的相似度使相似实体聚类对齐。基于神经网络的方法将不同的知识图谱嵌入到低维向量空间，计算不同实体嵌入之间的相似度来进行实体对齐，是目前实体对齐方法的研究重点。

（2）实体消歧是消除文本中实体指称的歧义，即解决一词多义的问题。近年来，词嵌入在自然语言处理领域应用广泛，目前的主流方法是利用分布式低维向量中的语义特征消除文本中实体指称的歧义，通过计算各个单词间的语义相似度提高实体消歧的效果，引入注意力的方法增强实体邻近上下文语义信息等。

在知识融合技术发展过程中，国外Newcombe等将基于属性相似度评分来判断实体是否匹配的问题转化为一个分类问题，建立了该问题的概率模型；Cochinwala等使用分类回归树、线性分析判别等方法完成了实体辨析。国内在后续的研究中逐步改善，例如，王会勇等提出了一种基于联合知识表示学习的多模态实体对齐方法，可以较好地实现多模态实体对齐，为知识图谱构建和补全提供了新的思路；李攀成等针对现有基于知识嵌入方法进行实体对齐导致精度较低的问题，提出了在对齐模型上附加类型约束匹配来加强选择条件，同时运用迭代实体对齐方法克服表述数据少的问题，提高了实体对齐的精确率；苏佳林等提出了自适应属性选择的实体对齐方法，该方法采用实体的语义和结构信息来训练基于两个图谱联合表示学习的实体对齐模型，再根据数据集特征生成最优的结果，提高了对齐的效果。

（4）知识推理

知识推理作为知识计算的重要组成部分，是针对知识图谱中已有事实或者关系不完备的情况，推断出未知的或隐藏的语义知识，从而扩展和丰富知识网络。图2-12、图2-13分别展示了知识推理技术演进历程及该技术相关的专利数统计。

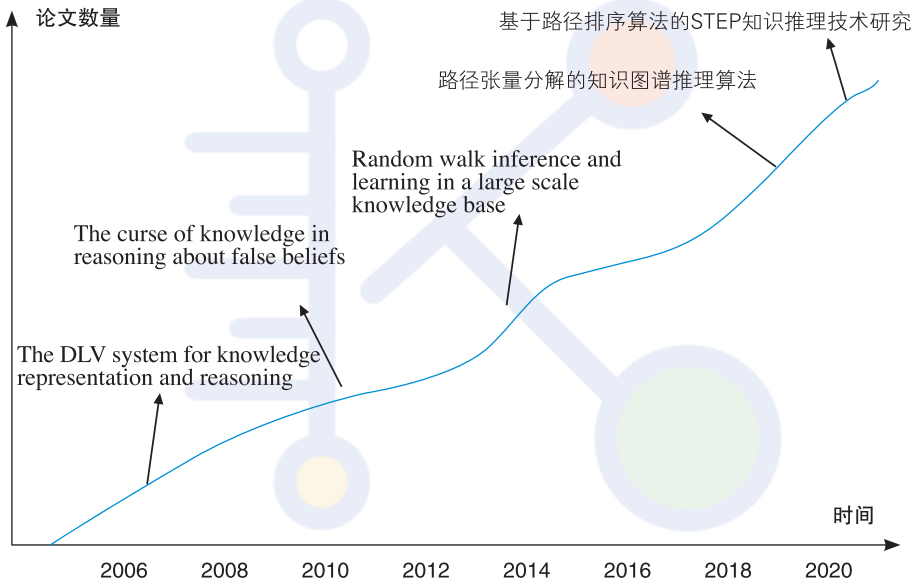

图2–12 知识推理技术演进历程

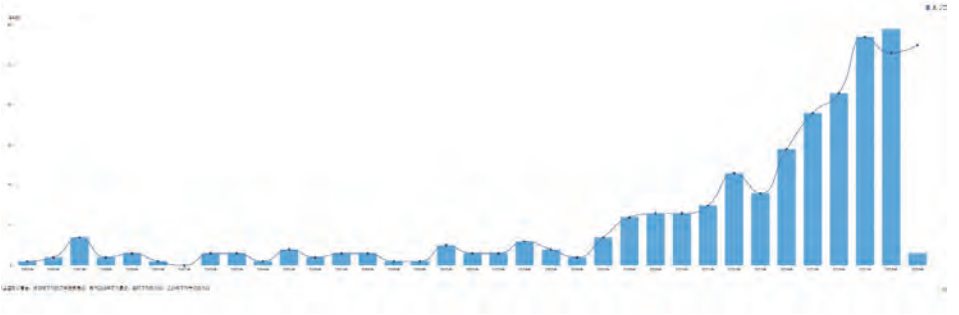

图2–13 知识推理技术相关的专利数统计 [来源：Aminer]

知识推理主要包括基于逻辑规则的知识推理、基于嵌入表示的知识推理、基于神经网络的知识推理：

（1）基于逻辑规则的知识推理，通过在知识图谱上运用简单的规则及特征，推理得到新的事实。这类方法可以很好地利用知识的符号性，准确性高且推理得到的结果具有可解释性，但是依赖于定义的规则，计算复杂度高。具有代表性的方法包括：马尔可夫逻辑网络，结合专家定义的逻辑规则与概率图模型构建网络，并在构建好的网络上执行推理；关联规则挖掘算法，从知识图谱中自动挖掘出置信度高的规则，并利用这些规则在知识图谱上推理得到新的知识；基于图结构的路径排序算法，将知识图谱中链接目标实体对之间的路径作为特征，为每一类关系训练一个逻辑回归模型，从而完成知识推理。

（2）基于嵌入表示的知识推理，将知识图谱嵌入到低维连续的向量空间，使得原来难以发现的关联关系变得显而易见。该类方法主要包括：张量分解模型，将整个知识图谱看作一个大的张量，通过张量分解技术分解为很多个小的张量，将高维的知识图谱进行降维处理；距离模型，将知识图谱中的每个关系看作从主体向量到客体向量的平移变换，以此推断出可能的关系；语义匹配模型，设计基于相似度的目标函数，该类模型认为存在的三元组应该具有较高的相似度，通过相似度来推断关系。

（3）基于神经网络的知识推理，充分利用神经网络学习到知识图谱的结构特征和语义特征，实现对图谱缺失关系的有效预测。具有代表性的方法包括：基于图卷积网络的知识推理，图卷积网络能够提取到图的拓扑结构特征，根据特征推理出缺失的关系信息。在实际应用中，基于神经网络的推理方法能有效提取图谱特征，计算效率高，但是输出结果的可解释性不强，所以可以将基于神经网络的推理方法和基于逻辑规则的推理结合以增加知识推理结果的可解释性。

在知识推理技术的发展中，有很多经典的方法。国外有路径排序算

法，通过将两实体间的路径作为特征以判断它们之间可能存在的关系；Socher提出的基于神经张量网络的图谱推理方法，通过对输入层的词向量求平均值提高推理精度。国内有很多后续的研究，如张美玉等将路径排序算法应用到产品设计的推理领域，取得了很好的效果；吴运兵等提出基于路径张量分解的知识图谱推理算法,利用路径排列算法获得知识图谱中各实体对间的关系路径，然后对实体对间的关系路径进行张量分解,并在优化更新过程中采用交替最小二乘法。

相较于国外，国内对知识图谱的研究主要集中在中文领域，且国内对知识图谱的研究起步较晚，许多研究都是从英文领域衍生而来。另外，相较于英文，中文具有更多的复杂性：首先，中文的词汇边界模糊，缺少英文文本中空格这样明确的分隔符，也没有明显的词性变换等特征，容易造成边界歧义；其次，中文中一词多义现象普遍，不同语境下的同一词表达的意思并不一样，且一种语义存在多种不同的表达，句式灵活多变。而且，随着互联网的快速发展，许多词将具有新的不同意义，进一步加强了难度，在设计模型时仍要着重考虑中文的特殊性。本节提及的部分文献名称详见文末参考文献。

二、 知识图谱现有开源内容

表2-4 国内外知识图谱开源内容

类型	开源组织	区域	开源内容	描述
标准	W3C	国外	RDF	即资源描述框架，是知识图谱的一种数据模型，主要用于表示 Web 上的资源及其关联。
			RDFS	即 RDF 模型，是用于描述 RDF 知识图谱上本体信息的基本元素。
			OWL	即 Web 本体语言，是用于描述 RDF 知识图谱上本体信息的语言，包括类层次结构和关系属性定义。
			SKOS	即简单知识组织系统，在 RDF 知识图谱上表达受控结构化词表的描述，包括叙词表、分类法、标题表系统。
			SPARQL	RDF 知识图谱的查询语言。包括：RDF 知识图谱的更新语言、联邦查询、服务与协议、蕴含范畴。
			SHACL	RDF 知识图谱的约束语言。用于定义 RDF 知识图谱需要满足的拓扑与值范围约束。
	ISO	国外	Dublin Core	即 Dublin 核心元数据元素集，是 ISO 15836 定义的用于描述资源的 15 个核心元素（即属性）。
			Schema.org	是用于给 Web 上资源标注模式元数据信息的公共分类词汇表。
知识库	莱比锡大学	国外	DBpedia	从维基百科中抽取结构化信息构建的知识图谱。
	维基媒体基金会	国外	Wikidata	用于为维基百科等提供数据服务的多语言知识图谱。
	马普研究所	国外	YAGO	多语言知识图谱，将维基百科分类体系与 WordNet 词汇定义融合。
	Luminoso Technologies	国外	ConceptNet	多语言常识知识图谱，旨在帮助计算机理解人们使用的单词的含义。
	OpenKG	国内	OpenBase	中文开放域知识图谱

类型	开源组织	区域	开源内容	描述
建模工具	斯坦福大学医学院生物信息研究中心	国外	Protégé	本体编辑和本体开发工具。
知识存储	Apache	国外	Jena	语义 Web 领域主要的开源框架和 RDF 三元组库，较好地遵循 W3C 标准。
	北京大学计算机科学技术研究所数据管理实验室	国内	gStore	面向 RDF 知识图谱的开源图数据库系统，遵循 Apache 开源协议。gStor 原生基于图数据模型，在存储 RDF 数据时维持并根据其图结构构建了基于二进制位图索引的新型索引结构——VS 树。
	Neo Technology	国外	Neo4j 社区版	高性能的 NoSQL 图形数据库，以结点和结点之间的关系为基本存储内容。
	ArangoDB	国外	ArangoDB	多模型数据库，兼有键、图和文档数据模型，提供了涵盖三种数据模型的统一的数据库查询语言。
实体抽取	斯坦福大学	国外	DeepDive	三元组抽取工具。
	开源工作者	国内	jieba 分词	一个 Python 中文分词组件，可以对中文文本进行分词、词性标注、关键词抽取等功能，并且支持自定义词典。
	哈尔滨工业大学	国内	LTP	提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等工作。
关系抽取	斯坦福大学	国外	DeepDive	实体间的关系抽取工具。
	华盛顿大学	国外	Reverb	开放三元组抽取工具，可以从英文句子中抽取形如(augument1, relation,argument2)的三元组。它不需要提前指定关系。
	华盛顿大学	国外	OLLIE	知识库三元组抽取组件，支持基于语法依赖树的关系抽取，对于长线依赖效果更好。
	浙江大学	国内	DeepKE	基于深度学习的开源中文知识图谱抽取工具。

类型	开源组织	区域	开源内容	描述
知识融合	开源工作者	国外	Falcon-AO	自动的本体匹配系统，已经成为 RDF(S) 和 OWL 所表达的 Web 本体相匹配的一种实用和流行的选择。
	开源工作者	国内	Sematch	用于知识图谱的语义相似性的开发、评价和应用的集成框架。
知识推理	Apache	国外	Jena	除了提供三元组的存储，RDF、RDFS、OWL 数据的接口，还提供规则引擎，用于知识的推理。
	JBoss	国外	Drools	具有一个易于访问企业策略、易于调整以及易于管理的开源业务规则引擎，符合业内标准，具有速度快、效率高的特点。
	牛津大学	国外	RDFox	一个高度可扩展的内存 RDF 三元组存储，支持共享内存并行 OWL 2 RL 推理。
	开源工作者	国内	SparkSRE	一个基于分布式内存计算框架 Spark 的语义推理引擎实现方案。

三、知识图谱标准化现状

（一）标准化现状

《国家标准化发展纲要》指出：标准是经济活动和社会发展的技术支撑，是国家基础性制度的重要方面。标准化在推进国家治理体系和治理能力现代化中发挥着基础性、引领性作用。新时代推动高质量发展、全面建设社会主义现代化国家，迫切需要进一步加强标准化工作。知识图谱的标准化对于提升知识图谱构建效率、推动数据在多领域复用、发挥知识图谱分析和技术价值有重要意义。

近年来资源描述框架RDF(Resource Description Framework)、资源描述框架模式RDFs（Resource Description Framework Schema）、网络本体语言

OWL（Web Ontology Language）等知识表示和知识建模相关标准，为知识图谱的规模化构建和应用提供了重要的支撑作用，而且随着知识图谱在各领域的深化应用，知识图谱技术框架、测试评估、能力成熟度模型、知识建模、知识融合、知识交换、知识计算及领域知识图谱构建与应用要求等方面的标准化需求日益攀升。此外，由于知识图谱应用系统日益增多，知识要素在行业内、集团内、企业间的安全交换与可靠流通需求也逐步显现，同样有待相关标准和配套工具来引领和支撑。

目前，知识图谱相关标准化需求已获得了国际标准化组织/国际电工委员会的第一联合技术委员会（ISO/IEC JTC 1）、电气电子工程师学会（IEEE）、国家人工智能标准化总体组、中国电子工业标准化技术协会等国内外标准化组织或协会的关注。在ISO/IEC JTC 1/SC 42（人工智能分技术委员会）、IEEE知识图谱标准化工作组、知识图谱国家标准编制工作组推动下制定了多项知识图谱领域相关的国际标准、国家标准、团体标准，形成了知识工程顶层标准、知识图谱顶层标准、知识图谱共性基础标准、知识图谱细分领域标准及配套白皮书、案例集等协同推进的局面。其制定的标准如图2-14所示。

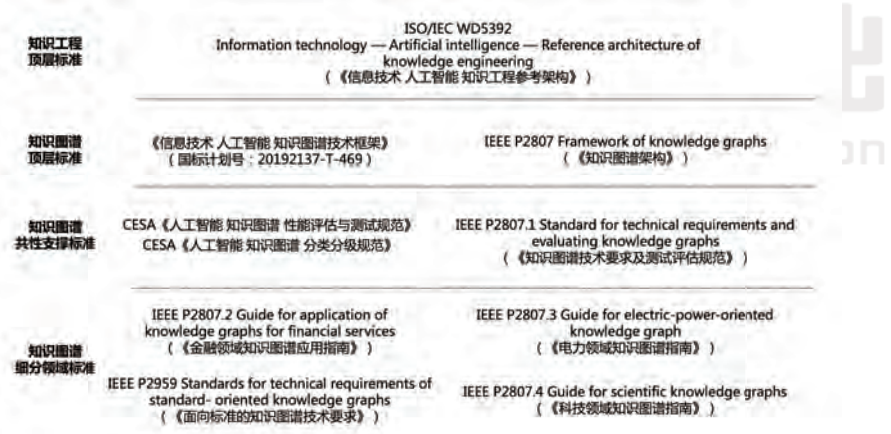

图2-14 知识图谱领域现有标准

此外，W3C、NIST、ISO/IEC JTC 1/SC 32（数据管理与交换分技术委员会）等标准化组织围绕知识图谱领域知识表示、知识获取、知识建模等关键技术标准也进行了研制，并相继发布了RDF、RDFs、OWL、本体模型等方面的一系列标准。当前，根据知识图谱相关在研及已发布标准所针对的标准化需求及所处环节，其分布情况如图2-15所示。

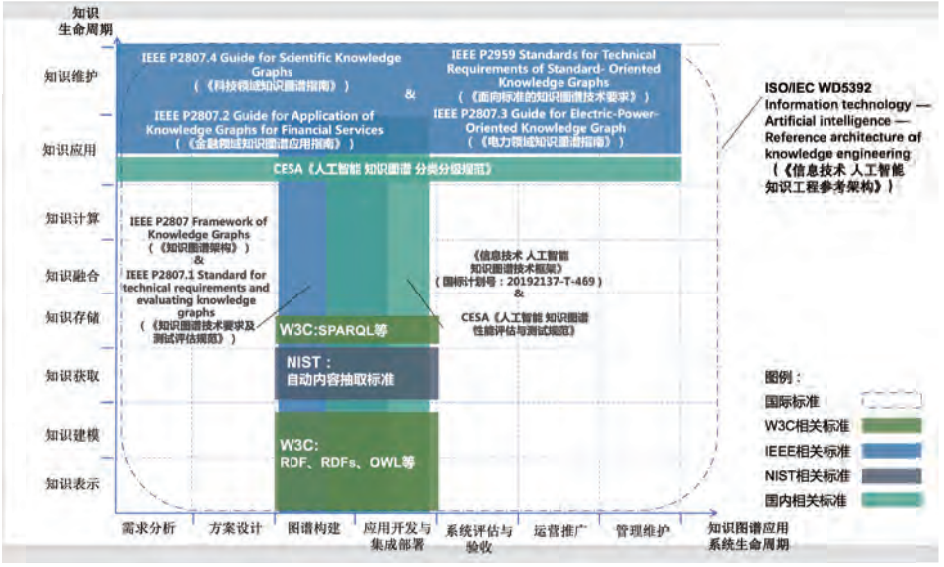

图2–15 知识图谱领域标准及相关标准关系图

1、国际标准化

（1）ISO/IEC JTC 1

ISO/IEC JTC 1，全称：International Organization for Standardization/ International Electro technical Commission JTC 1,中文名称：国际标准化组织/国际电工委员会 联合技术委员会1。ISO/IEC JTC 1是信息技术领域的国际标准化委员会，已经在人工智能领域进行了二十多年的标准化研制工作，主要集中在人工智能词汇、人机交互、计算机图像处理、云计算、大数据等人工智能关键技术领域。

在知识图谱相关术语方面，ISO/IEC JTC 1词汇组在已发布ISO/IEC 2382-28：1995《信息技术 词汇 第28部分：人工智能基本概念与专家系统》、ISO/IEC 2382-34:1997《信息技术 词汇 第31部分：人工智能机器学习》等标准中对知识、知识库、知识获取、知识工程、知识表示、认知建模等知识图谱相关专业术语进行了定义和说明，并在2015年最新研制发布的ISO/IEC 2382：2015《信息技术 词汇》标准中进行了更新。

在知识工程标准化工作方面，2017年10月，ISO/IEC JTC1第32届全会上批准并成立了SC 42人工智能分技术委员会，主要进行基础标准、计算方法、可信赖性和社会关注等方面开展国际标准化工作。2020年8月23日由我国提出的《信息技术 人工智能 知识工程参考架构》（Information technology - Artificial intelligence - Reference architecture of knowledge engineering）国际标准提案在ISO/IEC JTC 1/SC 42正式获批立项（项目号：ISO/IEC WD 5392）。该提案作为知识工程领域首个国际标准项目，规定了知识工程参考架构，明确了知识工程重要术语和概念，描述了知识工程中的角色、活动、构建层级、组件及其关系，由中国电子技术标准化研究院专家担任编辑。

此外，ISO/IEC JTC 1/SC 42/WG 5于2020年7月成立了本体、知识工程/表示临时咨询组，对该领域标准化需求进行进一步研究与梳理。ISO/IEC CD 22989.2《信息技术 人工智能 概念与术语》、ISO/IEC TR 24372《信息技术 人工智能 人工智能系统计算方法概览》等在研标准对知识表示、知识服务、知识获取与应用等相关内容也进行了研究和体现。

（2）IEEE

IEEE，全称：Institute of Electrical and Electronics Engineers，中文名称：电气与电子工程师协会，总部位于美国纽约，是一个国际性的电子技术与信息科学工程师的协会，也是全球最大的非营利性专业技术学会。IEEE标准协会隶属于IEEE，标准制定内容涵盖信息技术、通信、电力和

能源等多个领域,包括IEEE 802®有线与无线的网络通信系列标准、IEEE 7000TM人工智能伦理系列标准等。

中国电子技术标准化研究院联合国内多家企事业单位向IEEE标准协会提报的标准提案《知识图谱框架》（Framework of Knowledge Graph，项目编号：P2807）于2019年3月20日正式获批立项，并同步获批成立了IEEE知识图谱标准化工作组，主要开展知识图谱框架、关键技术、性能指标、典型应用等领域方向的标准研制工作。中国电子技术标准化研究院物联网研究中心李瑞琪担任工作组主席，清华大学人工智能研究院知识智能研究中心主任李涓子担任工作组副主席。

当前，在工作组的推动下已相继立项了6项IEEE标准，覆盖知识图谱测试评估规范及金融领域、电力领域、标准制修订领域科技信息领域等细分领域知识图谱构建技术要求，并有医疗领域知识图谱、知识图谱等多项潜在标准化需求正在论证中，初步形成了跨领域和细分领域标准协同推进的研制路线。在研标准名称及范围如表2-5所示。

表2–5 IEEE知识图谱相关在研标准

标准项目号	标准名称	范围
IEEE P2807	Framework of Knowledge Graphs 《知识图谱架构》	拟规范知识图谱框架、关键技术、性能指标、典型应用及相关领域、所需的数字基础设施等
IEEE P2807.1	Standard for Technical Requirements and Evaluating Knowledge Graphs 《知识图谱技术要求及测试评估规范》	拟规范知识图谱技术要求、技术指标、评估准则、测试用例等
IEEE P2807.2	Guide for Application of Knowledge Graphs for Financial Services 《金融服务领域知识图谱应用指南》	拟规范金融服务领域知识图谱技术框架、工作流程、实施指南和应用场景等

标准项目号	标准名称	范围
IEEE P2807.3	Guide for Electric-Power-Oriented Knowledge Graph《面向电力行业的知识图谱指南》	拟规范从已发布标准中提取知识并构建知识图谱的数据要求、构建流程及应用场景等
IEEE P2807.4	Guide for Scientific Knowledge Graphs《科技知识图谱指南》	拟规定科技知识图谱的数据模式、构建过程及共享应用等内容。
IEEE P2959	Standards for Technical Requirements of Standard-Oriented Knowledge Graphs《面向标准文本的知识图谱技术要求》	拟规范从已发布标准中提取知识并构建知识图谱的数据要求、构建流程及应用场景等。

（3）W3C

W3C，全称：World Wide Web Consortium，中文名称：万维网联盟，是万维网主要的国际标准化组织机构，同时也是万维网领域最具有权威性和影响力的国际中立性技术标准化组织。W3C标准化组织1994年建立，主要宗旨是通过促进通用协议的发展并确保相关标准具有通用性，对web关键技术进行标准化工作。

W3C于2018年7月推动成立了人工智能知识表示社区小组，旨在探讨人工智能领域知识的概念化和规范的要求、最佳做法和实施选项等。在知识图谱领域，W3C相关标准化工作主要集中在语义网知识描述体系方面，研制与发布的XML、RDF、SPARQL、RDF Schema、OWL等系列标准，形成了一系列知识表示、知识建模、知识存储关键技术相关标准，语义网知识描述技术栈如图2-16所示。

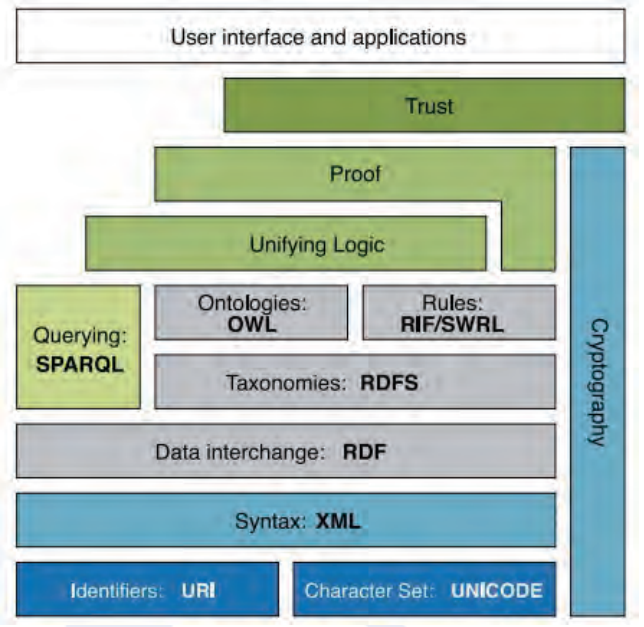

图2–16 语义网知识描述技术栈

在知识表示方面，W3C理事会推荐了XML、RDF、RDFS、OWL四项主要技术标准，其中，XML是一种元数据语法标准，也是一种标记语言，用于传输和存储数据，是语义网基础层。RDF系列标准包括RDF Primer、RDF Test Cases、RDF Concept、RDF Syntax。RDF是一种元数据语义描述标准，它被设计为一种描述信息的通用方法，可以被计算机应用程序读取并理解，现实中任何实体都可以表示成RDF模型中的资源。同时W3C理事会提议的SPARQL Requirements与SPARQL Language标准成为检索和操作基于RDF存储的知识图谱。

在知识建模方面，W3C理事会推荐了RDFS（RDF Schema）与OWL系列标准。其中RDFS是RDF的扩展，规范了用于描述RDF资源的属性、类的词汇表以及属性和类在语义上的层次结构。OWL是一种语义网本体语言，用于构建领域相关的本体，主要技术标准包括OWL Overview、

OWL Guide、OWL Reference、OWL Syntax、OWL Test Cases、OWL Use Cases、以及Parsing OWL in RDF。OWL是在RDFS基础上丰富了类和属性的词汇，例如类不相交性、基数约束、类的布尔组合等，主要增加了类、属性之间关系的定义或约束。

（4）NIST

NIST，全称：National Institute of Standards and Technology，中文名称：美国国家标准技术研究院，直属美国商务部，主要从事物理、生物和工程方面的基础和应用研究。在MUC-7之后，MUC由美国国家标准技术研究院组织的自动内容抽取(Automatic Content Extraction Evaluation, ACE)评测取代，ACE评测标准从1999年开始筹划，2000年正式启动，其中关系识别和检测任务定义了较为详细的关系类别体系，用于两个实体间的语义关系抽取。ACE-2008包括了7大类和18个子类的实体关系，从2004年开始，事件抽取成为ACE评测的主要任务。

此外，国际电信联盟（International Telecommunications Union, ITU）2016年开始进行人工智能相关标准化研究。但前尚未发布知识图谱相关标准以及研制计划。

2、国内标准化

（1）国家标准

在知识图谱相关国家标准的制定方面，2019年7月8日，国家标准化管理委员会下达2019年第二批国家标准制修订计划（国标委发[2019]22号），其中由中国电子技术标准化研究院提出的《信息技术 人工智能 知识图谱技术框架》标准（计划号：20192137-T-469）获得立项，并由全国信息技术标准化技术委员会）归口。本标准拟就知识图谱技术框架、利益相关方、关键技术要求、性能指标、典型应用及相关领域、数字基础设施、使能技术等内容进行研究，以厘清知识图谱核心标准化需求，提升我

国知识图谱标准化工作水平，并促进知识图谱在各行业的推广应用。

此外，全国信息技术标准化技术委员会在相关国际标准研制的基础之上制定了《信息技术 词汇 第28部分:人工智能基本概念与专家系统》、《信息技术 词汇 第31部分:人工智能机器学习》、《信息技术 大数据 术语》三项项基础国家标准，其中给出了知识工程、知识表示、知识获取、本体等部分知识图谱相关术语。

（2）团体标准

在知识图谱相关团体标准方面，由中国电子技术标准化研究院向中国电子工业标准化技术协会提出的《人工智能 知识图谱 性能评估与测试规范》(项目号：CESA-2020-2-020)、《人工智能 知识图谱 分类分级规范》(项目号：CESA-2020-2-019)两项团体标准于2020年6月正式获批立项，其标准化范围如下：

- 《人工智能 知识图谱 分类分级规范》针对当前知识图谱供应商、集成商和服务商能力良莠不齐、分类不清晰和评价方法缺失等标准化需求，拟规定知识图谱相关系统供应商的分类分级模型、能力框架、能力评价方法、评估指标等内容。
- 《人工智能 知识图谱 性能评估与测试规范》针对当前知识图谱性能指标及测试方法不明确、构建过程中各环节性能与质量评估不规范等标准化需求，拟规定知识图谱质量评估要求、知识图谱性能指标、测试框架、测试需求模型及度量准则等内容。

（3）标准体系研究

中国电子技术标准化研究院联合中电科大数据研究院有限公司、东软集团股份有限公司、联想(北京)有限公司、南华大学、星环信息科技（上海）有限公司、上海思贤信息技术股份有限公司、成都数联铭品科技有限公司、阿里巴巴网络技术有限公司等21家知识图谱领域相关开发商、系统

集成商、用户企业、科研院所、高校联合编写并发布了《知识图谱标准化白皮书》（2019年版）。其中，从哲学层面、政策层面、产业层面、行业层面、技术层面、工具层面、支撑技术等多个层面对知识图谱的实际需求、关键技术、面临的问题与挑战、标准化需求、展望与建议等进行了梳理，并提出了知识图谱技术架构和标准体系框架等。

知识图谱标准体系结构和框架图如图2-17和图2-18所示。

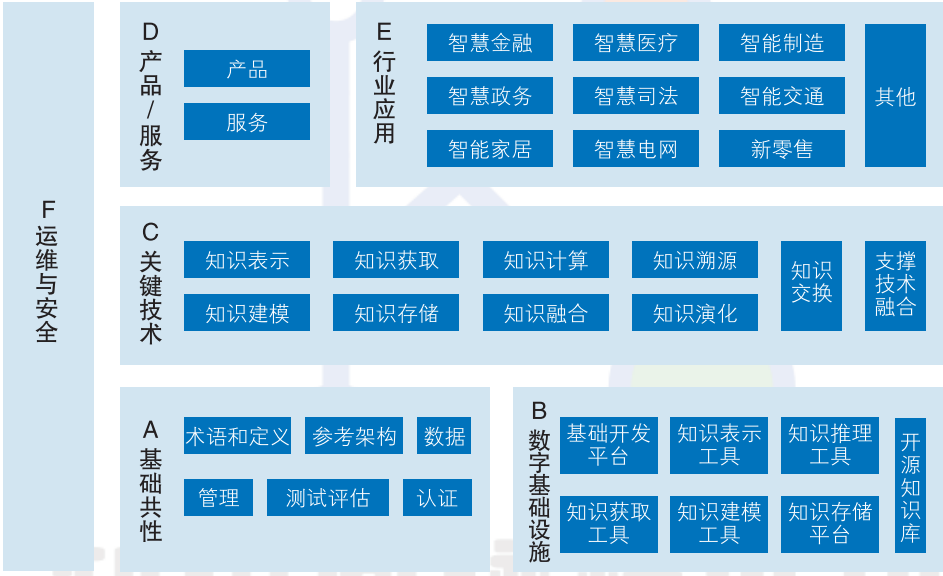

图2–17 知识图谱标准体系结构图

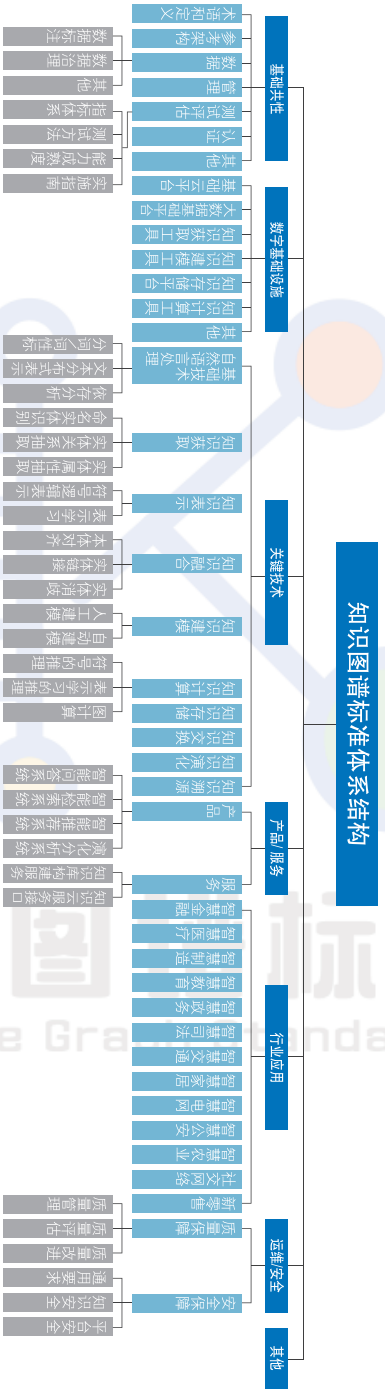

图2–18 知识图谱标准体系框架图

（二）标准化挑战

1、知识图谱质量评估与测试相关标准缺失

知识图谱质量的保障不仅涉及知识图谱构建过程知识表示、知识建模、知识获取、知识存储、知识融合、知识计算等各环节的质量评估，也涉及知识图谱应用系统各模块功能和性能的测试。因此，有赖于从知识图谱的内容和系统两个层面构建较为完备的质量评估体系和质控指标，并结合当前企业实践情况给出指标通过准则，进而为知识图谱应用系统策划、开发与部署过程提供指导和参考。

2、本体模型构建与联动更新相关标准缺失

本体模型及其Schema构建过程涉及对领域知识的高度抽象化建模，无法简单固化或设定，需领域专家的深度参与。而且，由于知识图谱应用系统部署实施后，随着时间推移、领域研究深度广度的拓展及业务模式的变化，本体模型也可能需要不断演进，以保障其准确性和适用性。因此，有待规范化的本体模型描述格式及联动模式，以保障其应用和更新的可持续性。

3、跨域知识交换与融合相关标准缺失

随着知识图谱应用系统在各领域、各企业的逐步建设和完善，目前已出现了一批优秀的成果。然而，由于建设初期相关系统着重于聚焦企业内部需要，顶层本体模型的构建流程和表达方式差异大且知识表示形式多样，导致建成后各系统间知识交换、知识图谱集成与融合困难，加深了集团内企业/部门间的信息壁垒，阻碍了行业内知识的流通。与数据交换相比，知识交换中不仅涉及知识本身还涉及配套的概念、语义等，因此有赖于规范化的知识交换与融合协议，对本体模型、知识表示、知识访问、交换模式等多个方面进行统一。

4、知识图谱中知识查询格式与语言相关标准缺失

当前，知识图谱无统一的查询语言，各厂商多根据自身需要进行设计和选择，存在差异大。这导致不同厂商无法对同一知识图谱进行直接操作，进而增加了用户企业在后期维护和升级知识图谱应用系统过程中的投入成本，也阻碍了通用知识检索或计算工具的研制与开发，有待相关标准进行支撑。

5、知识图谱服务方能力评估相关标准缺失

由于知识图谱应用系统在建成后将逐渐成为企业内部的重要知识服务基础设施，在构建过程中不仅需大量企业内、行业内专家的介入与支持，还需与企业内必要业务系统进行集成调试，并可能涉及多源异构数据的清洗等问题。这对知识图谱服务方在项目管理、系统集成、知识图谱构建、数据安全保障与数据治理等能力均提出了相应要求。因此，有赖于对其能力进行合理评估与分析以保障最终知识图谱应用系统的可靠交付。

除上述问题以外，细分行业中还面临专业术语集或术语库匮乏，知识图谱应用系统与业务系统集成，典型知识服务部署、实施与评估等方面的标准化需求与挑战。

四、知识图谱产品认证现状

围绕知识图谱构建与应用相关软件产品或服务平台测评与认证需求，中国电子技术标准化研究院联合北京赛西认证有限责任公司、联想、华为、百度、腾讯云、蚂蚁金服、百分点、网智天元、华宇等企事业单位依托上述知识图谱相关国家标准和团体标准等联合研制了《知识图谱构建平台认证技术规范》、《知识图谱应用平台认证技术规范》、《知识图谱构建平台认证实施规则》和《知识图谱应用平台认证实施规则》。其中，给出了测评与认证指标体系及配套检测项、功能点和合格要求，覆盖了知识