第1章 知识图谱概述

陈华钧 浙江大学，漆桂林 东南大学，王昊奋 乐言科技，王鑫 天津大学

1.1 什么是知识图谱

知识图谱是一种用图模型来描述知识和建模世界万物之间的关联关系的技术方法[1]。知识图谱由节点和边组成。节点可以是实体，如一个人、一本书等，或是抽象的概念，如人工智能、知识图谱等。边可以是实体的属性，如姓名、书名，或是实体之间的关系，如朋友、配偶。知识图谱的早期理念来自Semantic Web[2,3]（语义网），其最初理想是把基于文本链接的万维网转化成基于实体链接的语义网。

1989年，Tim Berners-Lee 提出构建一个全球化的以“链接”为中心的信息系统（Linked Information System）。任何人都可以通过添加链接把自己的文档链入其中。他认为，相比基于树的层次化组织方式，以链接为中心和基于图的组织方式更加适合互联网这种开放的系统。这一思想逐步被人们实现，并演化发展成为今天的World Wide Web。

1994年，Tim Berners-Lee 又提出 Web 不应该仅仅只是网页之间的互相链接。实际上，网页中描述的都是现实世界中的实体和人脑中的概念。网页之间的链接实际包含语义，即这些实体或概念之间的关系；然而，机器却无法有效地从网页中识别出其中蕴含的语义。他于1998年提出了Semantic Web的概念[4]。Semantic Web仍然基于图和链接的组织方式，只是图中的节点代表的不只是网页，而是客观世界中的实体（如人、机构、地点等），而超链接也被增加了语义描述，具体标明实体之间的关系（如出生地是、创办人是等）。相对于传统的网页互联网，Semantic Web的本质是数据的互联网（Web of Data）或事物的互联网（Web of Things）。

在 Semantic Web 被提出之后，出现了一大批新兴的语义知识库。如作为谷歌知识图谱后端的Freebase[5]，作为IBM Waston后端的DBpedia[6]和Yago[7]，作为Amazon Alexa后端的True Knowledge，作为苹果Siri后端的Wolfram Alpha，以及开放的Semantic Web Schema——Schema.ORG[8]，目标成为世界最大开放知识库的Wikidata[9]等。尤其值得一提的是，2010年谷歌收购了早期语义网公司 MetaWeb，并以其开发的 Freebase 作为数据基础之一，于2012年正式推出了称为知识图谱的搜索引擎服务。随后，知识图谱逐步在语义搜索[10,11]、智能问答[12-14]、辅助语言理解[15,16]、辅助大数据分析[17-19]、增强机器学习的可解释性[20]、结合图卷积辅助图像分类[21,22]等多个领域发挥出越来越重要的作用。

如图1-1所示，知识图谱旨在从数据中识别、发现和推断事物与概念之间的复杂关系，是事物关系的可计算模型。知识图谱的构建涉及知识建模、关系抽取、图存储、关系

推理、实体融合等多方面的技术，而知识图谱的应用则涉及语义搜索、智能问答、语言理解、决策分析等多个领域。构建并利用好知识图谱需要系统性地利用包括知识表示（Knowledge Representation）、图数据库、自然语言处理、机器学习等多方面的技术。

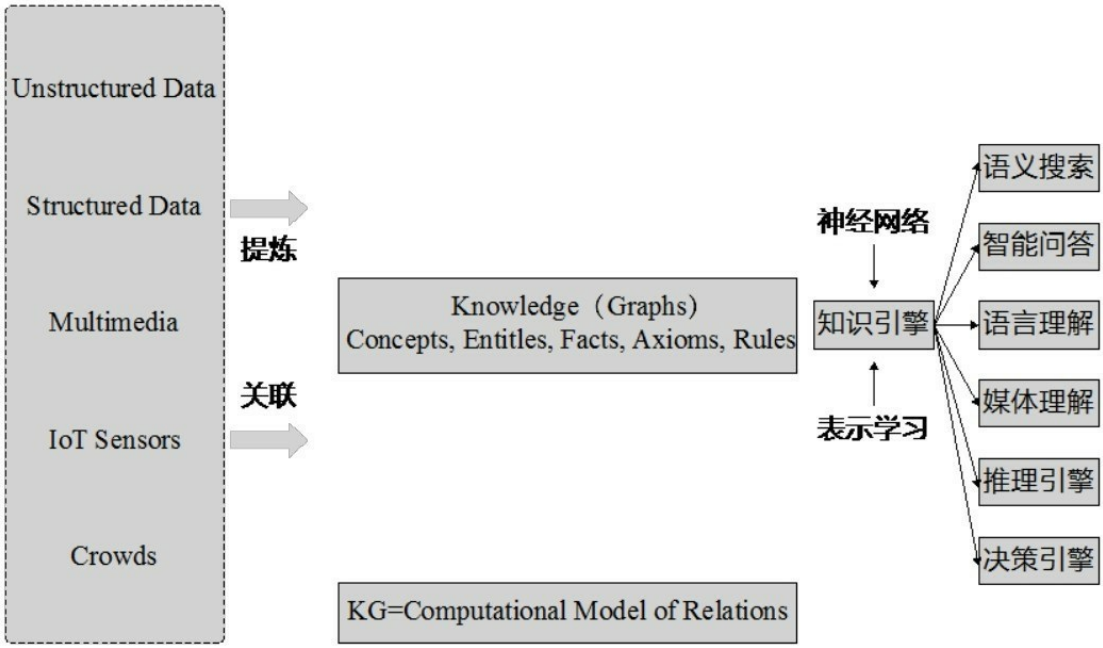

图1-1 知识图谱：事物关系的可计算模型

1.2 知识图谱的发展历史

知识图谱并非突然出现的新技术，而是历史上很多相关技术相互影响和继承发展的结果，包括语义网络、知识表示、本体论、Semantic Web、自然语言处理等，有着来自Web、人工智能和自然语言处理等多方面的技术基因。从早期的人工智能发展历史来看，Semantic Web是传统人工智能与Web融合发展的结果，是知识表示与推理在Web中的应用；RDF（Resource Description Framework，资源描述框架）、OWL（Web Ontology Language，网络本体语言）都是面向 Web 设计实现的标准化的知识表示语言；而知识图谱则可以看作是Semantic Web的一种简化后的商业实现，如图1-2所示。

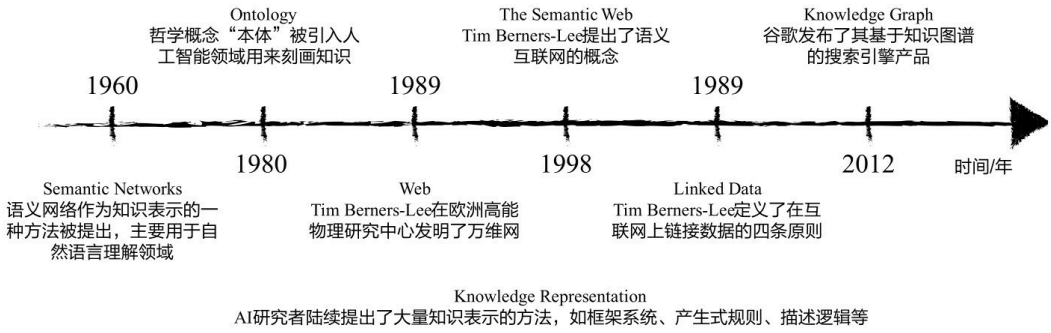

图1-2 从语义网络到知识图谱

在人工智能的早期发展流派中，符号派（Symbolism）侧重于模拟人的心智，研究怎样用计算机符号表示人脑中的知识并模拟心智的推理过程；连接派（Connectionism）侧重于模拟人脑的生理结构，即人工神经网络。符号派一直以来都处于人工智能研究的核心位置。近年来，随着数据的大量积累和计算能力的大幅提升，深度学习在视觉、听觉等感知处理中取得突破性进展，进而又在围棋等博弈类游戏、机器翻译等领域获得成功，使得人工神经网络和机器学习获得了人工智能研究的核心地位。深度学习在处理感知、识别和判断等方面表现突出，能帮助构建聪明的人工智能，但在模拟人的思考过程、处理常识知识和推理，以及理解人的语言方面仍然举步维艰。

哲学家柏拉图把知识（Knowledge）定义为“Justified True Belief”，即知识需要满足三个核心要素：合理性（Justified）、真实性（True）和被相信（Believed）。简而言之，知识是人类通过观察、学习和思考有关客观世界的各种现象而获得并总结出的所有事实（Fact）、概念（Concept）、规则（Rule）或原则（Principle）的集合。人类发明了各种

手段来描述、表示和传承知识，如自然语言、绘画、音乐、数学语言、物理模型、化学公式等。具有获取、表示和处理知识的能力是人类心智区别于其他物种心智的重要特征。人工智能的核心也是研究怎样用计算机易于处理的方式表示、学习和处理各种各样的知识。知识表示是现实世界的可计算模型（Computable Model of Reality）。从广义上讲，神经网络也是一种知识表示形式，如图1-3所示。

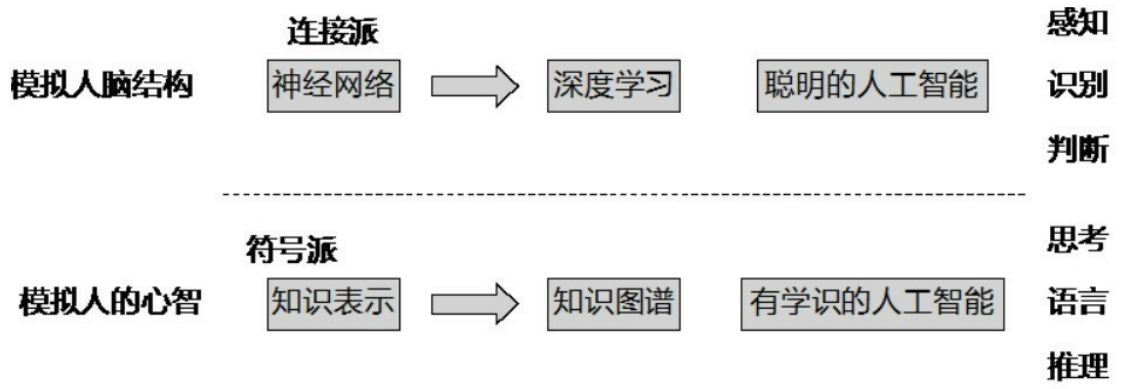

图1-3 知识图谱帮助构建有学识的人工智能

符号派关注的核心正是知识的表示和推理（KRR,Knowledge Representation and Reasoning）。早在1960年，认知科学家 Allan M.Collins 提出用语义网络（Semantic Network）研究人脑的语义记忆。例如，WordNet[23]是典型的语义网络，它定义了名词、动词、形容词和副词之间的语义关系。WordNet被广泛应用于语义消歧等自然语言处理领域。

1970年，随着专家系统的提出和商业化发展，知识库（Knowledge Base）构建和知识表示更加得到重视。专家系统的基本想法是：专家是基于大脑中的知识来进行决策的，因此人工智能的核心应该是用计算机符号表示这些知识，并通过推理机模仿人脑对知识进行处理。依据专家系统的观点，计算机系统应该由知识库和推理机两部分组成，而不是由函数等过程性代码组成。早期的专家系统最常用的知识表示方法包括基于框架的语言（Frame-based Languages）和产生式规则（Production Rules）等。框架语言主要用于描述客观世界的类别、个体、属性及关系等，较多地被应用于辅助自然语言理解。产生式规则主要用于描述类似于IF-THEN的逻辑结构，适合于刻画过程性知识。

知识图谱与传统专家系统时代的知识工程有着显著的不同。与传统专家系统时代主要依靠专家手工获取知识不同，现代知识图谱的显著特点是规模巨大，无法单一依靠人工和专家构建。如图1-4所示，传统的知识库，如Douglas Lenat从1984年开始创建的常识知识库 Cyc，仅包含700万条[1]的事实描述（Assertion）。Wordnet 主要依靠语言学专家定义名

词、动词、形容词和副词之间的语义关系，目前包含大约20万条的语义关系。由著名人工智能专家 Marvin Minsky于1999年起开始构建的 ConceptNet[24]常识知识库依靠了互联网众包、专家创建和游戏三种方法，但早期的 ConceptNet 规模在百万级别，最新的 ConceptNet 5.0也仅包含2800万个RDF三元组关系描述。谷歌和百度等现代知识图谱都已经包含超过千亿级别的三元组，阿里巴巴于2017年8月发布的仅包含核心商品数据的知识图谱也已经达到百亿级别。DBpedia已经包含约30亿个RDF三元组，多语种的大百科语义网络BabelNet包含19亿个RDF三元组[25],Yago3.0包含1.3亿个元组，Wikidata已经包含4265万条数据条目，元组数目也已经达到数十亿级别。截至目前，开放链接数据项目Linked Open Data[2]统计了其中有效的2973个数据集，总计包含大约1494亿个三元组。

现代知识图谱对知识规模的要求源于“知识完备性”难题。冯·诺依曼曾估计单个个体大脑的全量知识需要2.4×10^{20}个bits存储[26]。客观世界拥有不计其数的实体，人的主观世界还包含无法统计的概念，这些实体和概念之间又具有更多数量的复杂关系，导致大多数知识图谱都面临知识不完全的困境。在实际的领域应用场景中，知识不完全也是困扰大多数语义搜索、智能问答、知识辅助的决策分析系统的首要难题。

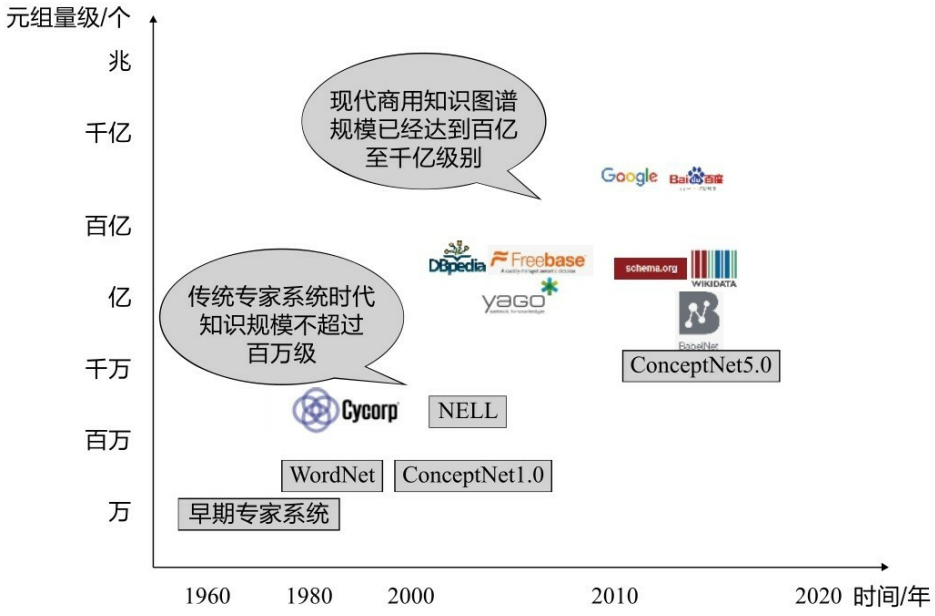

图1-4 现代知识图谱的规模化发展

1.3 知识图谱的价值

知识图谱最早的应用是提升搜索引擎的能力。随后，知识图谱在辅助智能问答、自然语言理解、大数据分析、推荐计算、物联网设备互联、可解释性人工智能等多个方面展现出丰富的应用价值。

1.辅助搜索

互联网的终极形态是万物的互联，而搜索的终极目标是对万物的直接搜索。传统搜索引擎依靠网页之间的超链接实现网页的搜索，而语义搜索是直接对事物进行搜索，如人物、机构、地点等。这些事物可能来自文本、图片、视频、音频、IoT 设备等各种信息资源。而知识图谱和语义技术提供了关于这些事物的分类、属性和关系的描述，使得搜索引擎可以直接对事物进行索引和搜索，如图1-5所示。

图1-5 知识图谱辅助搜索

2.辅助问答

人与机器通过自然语言进行问答与对话是人工智能实现的关键标志之一。除了辅助搜索，知识图谱也被广泛用于人机问答交互中。在产业界，IBM Watson 背后依托 DBpedia

和Yago等百科知识库和WordNet等语言学知识库实现深度知识问答。Amazon Alex主要依靠True Knowledge公司积累的知识图谱。度秘、Siri的进化版Viv、小爱机器人、天猫精灵背后都有海量知识图谱作为支撑。

伴随着机器人和 IoT 设备的智能化浪潮的掀起，基于知识图谱的问答对话在智能驾驶、智能家居和智能厨房等领域的应用层出不穷。典型的基于知识图谱的问答技术或方法包括：基于语义解析、基于图匹配、基于模板学习、基于表示学习和深度学习以及基于混合模型等。在这些方法中，知识图谱既被用来辅助实现语义解析，也被用来匹配问句实体，还被用来训练神经网络和排序模型等。知识图谱是实现人机交互问答必不可少的模块。

3.辅助大数据分析

知识图谱和语义技术也被用于辅助进行数据分析与决策。例如，大数据公司 Palantir 基于本体融合和集成多种来源的数据，通过知识图谱和语义技术增强数据之间的关联，使得用户可以用更加直观的图谱方式对数据进行关联挖掘与分析。

知识图谱在文本数据的处理和分析中也能发挥独特的作用。例如，知识图谱被广泛用来作为先验知识从文本中抽取实体和关系，如在远程监督中的应用。知识图谱也被用来辅助实现文本中的实体消歧（Entity Disambiguation）、指代消解和文本理解等。

近年来，描述性数据分析（Declarative Data Analysis）受到越来越多的重视。描述性数据分析是指依赖数据本身的语义描述实现数据分析的方法。不同计算性数据分析主要以建立各种数据分析模型，如深度神经网络，而描述性数据分析突出预先抽取数据的语义，建立数据之间的逻辑，并依靠逻辑推理的方法（如DataLog）来实现数据分析。

4.辅助语言理解

背景知识，特别是常识知识，被认为是实现深度语义理解（如阅读理解、人机问答等）必不可少的构件。一个典型的例子是Winograd Schema Challenge（WSC竞赛）。WSC由著名的人工智能专家 Hector Levesque 教授提出，2016年，在国际人工智能大会IJCAI上举办了第一届WSC竞赛。WSC主要关注那些必须要叠加背景知识才能理解句子语义的NLP任务。例如，在下面这个例子中，当描述it是big时，人很容易理解it指代trophy；而当it与small搭配时，也很容易识别出it指代suitcase。

The trophy would not fit in the brown suitcase because it was too big（small）.What was too big（small）?

Answer 0:the trophy Answer 1:the suitcase

这个看似非常容易的问题，机器却毫无办法。正如自然语言理解的先驱 Terry Winograd 所说的，当一个人听到一句话或看到一段句子的时候，会使用自己所有的知识和智能去理解。这不仅包括语法，也包括其拥有的词汇知识、上下文知识，更重要的是对

相关事物的理解。

5.辅助设备互联

人机对话的主要挑战是语义理解，即让机器理解人类语言的语义。另外一个问题是机器之间的对话，这也需要技术手段来表示和处理机器语言的语义。语义技术也可被用来辅助设备之间的语义互联。OneM2M　是2012年成立的全球最大的物联网国际标准化组织，其主要是为物联设备之间的互联提供“标准化黏合剂”。OneM2M　关注了语义技术在封装设备数据的语义，并基于语义技术实现设备之间的语义互操作的问题。此外，OneM2M还关注设备数据的语义和人类语言的语义怎样适配的问题。如图1-6所示，一个设备产生的原始数据在封装了语义描述之后，可以更加容易地与其他设备的数据进行融合、交换和互操作，并可以进一步链接进入知识图谱中，以便支持搜索、推理和分析等任务。

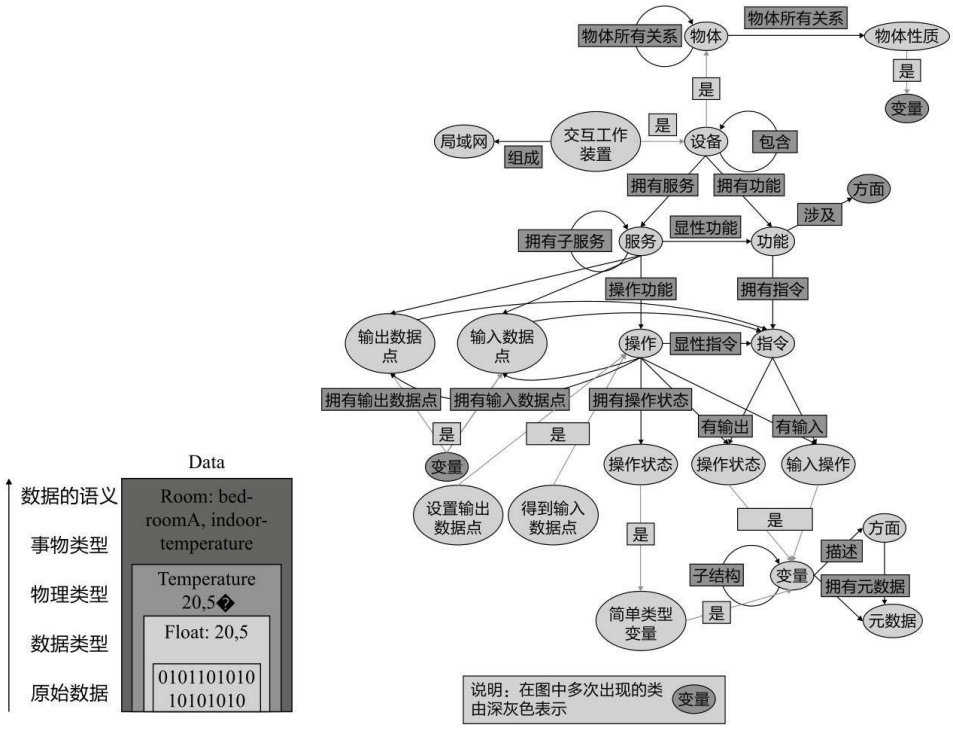

图1-6 设备语义的封装

1.4 国内外典型的知识图谱项目

从人工智能的概念被提出开始，构建大规模的知识库一直都是人工智能、自然语言理解等领域的核心任务之一。下面分别介绍早期的知识库项目、互联网时代的知识图谱、中文开放知识图谱和垂直领域知识图谱。

1.4.1 早期的知识库项目

Cyc 是持续时间最久、影响范围较广、争议也较多的知识库项目。Cyc 最初的目标是要建立人类最大的常识知识库。典型的常识知识如“Every tree is a plant”“Plants die eventually”等。Cyc 知识库主要由术语（Term）和断言（Assertion）组成。术语包含概念、关系和实体的定义。断言用来建立术语之间的关系，既包括事实（Fact）描述，也包含规则（Rule）描述。最新的 Cyc 知识库已经包含有50万条术语和700万条断言。Cyc的主要特点是基于形式化的知识表示方法刻画知识。形式化的优势是可以支持复杂的推理，但过于形式化也导致知识库的扩展性和应用的灵活性不够。

WordNet 是最著名的词典知识库，由普林斯顿大学认知科学实验室从1985年开始开发。WordNet 主要定义了名词、动词、形容词和副词之间的语义关系。例如，名词之间的上下位关系，如“猫科动物”是“猫”的上位词；动词之间的蕴涵关系，如“打鼾”蕴涵着“睡眠”等。

ConceptNet最早源于MIT媒体实验室的OMCS（Open Mind Common Sense）项目。与 Cyc 相比，ConceptNet 采用了非形式化、更加接近自然语言的描述，而不是像 Cyc 一样采用形式化的谓词逻辑。与链接数据和谷歌知识图谱相比，ConceptNet 比较侧重于词与词之间的关系。从这个角度来看，ConceptNet更加接近于WordNet，但是又比WordNet包含的关系类型多。

1.4.2 互联网时代的知识图谱

互联网的发展为知识工程提供了新的机遇。在一定程度上，互联网的出现帮助传统知识工程突破了在知识获取方面的瓶颈。从1998年Tim Berners Lee提出语义网至今，涌现出了大量以互联网资源为基础的新一代知识库。这类知识库的构建方法可以分为三类：互联

网众包、专家协作和互联网挖掘。

Freebase 是一个开放共享的、协同构建的大规模链接数据库。Freebase 是由硅谷创业公司MetaWeb于2005年启动的一个语义网项目。2010年，谷歌收购了Freebase，并作为其知识图谱数据来源之一。Freebase 主要采用社区成员协作方式构建，主要数据来源包括Wikipedia、世界名人数据库（NNDB）、开放音乐数据库（MusicBrainz）以及社区用户的贡献等。Freebase基于RDF三元组模型，底层采用图数据库进行存储。Freebase的一个特点是不对顶层本体做非常严格的控制，用户可以创建与编辑类和关系的定义。2016年，谷歌宣布将Freebase的数据和API服务都迁移至Wikidata，并正式关闭了Freebase。

DBpedia意指数据库版本的Wikipedia，是早期的语义网项目，是从Wikipedia抽取出来的链接数据集。DBpedia 采用了一个较为严格的本体，包含人、地点、音乐、电影、组织机构、物种、疾病等类定义。此外，DBpedia还与Freebase、OpenCYC、Bio2RDF等多个数据集建立了数据链接。DBpedia采用RDF语义数据模型，总共包含30亿个RDF三元组。

Schema.org是从2011年开始，由Bing、Google、Yahoo和Yandex等搜索引擎公司共同支持的语义网项目。Schema.org支持各个网站采用语义标签（Semantic Markup）的方式将语义化的链接数据嵌入到网页中。搜索引擎自动收集和归集这些数据，快速地从网页中抽取语义化的数据。Schema.org 提供了一个词语本体，用于描述这些语义标签。目前，这个词汇本体已经包含600多个类和900多个关系，覆盖范围包括个人、组织机构、地点、时间、医疗、商品等。谷歌于2015年推出的定制化知识图谱支持个人和企业在其网页中增加包括企业联系方法、个人社交信息等在内的语义标签，并通过这种方式快速汇集高质量的知识图谱数据。谷歌的一份统计数据显示，超过31%的网页和1200万家网站已经使用了Schema.org 发布语义化的链接数据。其他采用了部分 Schema.org 功能的还包括Cortana、Yandex、Pinterest、Siri 等。Schema.org 的本质是采用互联网众包的方式生成和收集高质量的知识图谱数据。

Wikidata 的目标是构建一个免费开放、多语言、任何人或机器都可以编辑修改的大规模链接知识库。Wikidata 由 Wikipedia 于2012年启动，早期得到微软联合创始人 Paul Allen、Gordon Betty Moore基金会以及谷歌的联合资助。Wikidata继承了Wikipedia的众包协作机制，但与 Wikipedia 不同的是，Wikidata 支持以三元组为基础的知识条目（Item）的自由编辑。一个三元组代表一个关于该条目的陈述（Statement）。例如，可以给“地球”的条目增加“<地球，地表面积是，五亿平方公里>”的三元组陈述。截至2018年，Wikidata已经包含超过5000万个知识条目。

BabelNet 是类似于 WordNet 的多语言词典知识库。BabelNet 的目标是解决 WordNet 在非英语语种中数据缺乏的问题。BabelNet采用的方法是将WordNet词典与Wikipedia集成。首先建立 WordNet 中的词与 Wikipedia 的页面标题的映射，然后利用 Wikipedia 中的

多语言链接，再辅以机器翻译技术，给 WordNet 增加多种语言的词汇。BabelNet3.7包含了271种语言、1400万个同义词组、36.4万个词语关系和3.8亿个从Wikipedia中抽取的链接关系，总计超过19亿个RDF三元组。BabelNet集成了WordNet在词语关系上的优势和Wikipedia在多语言语料方面的优势，成功构建了目前最大规模的多语言词典知识库。

NELL（Never-Ending Language Learner）是卡内基梅隆大学开发的知识库。NELL主要采用互联网挖掘的方法从Web中自动抽取三元组知识。NELL的基本理念是：给定一个初始的本体（少量类和关系的定义）和少量样本，让机器能够通过自学习的方式不断地从Web中学习和抽取新的知识。目前，NELL已经抽取了300多万条三元组知识。

Yago 是由德国马普研究所研制的链接数据库。Yago 主要集成了 Wikipedia、WordNet和GeoNames三个数据库的数据。Yago将WordNet的词汇定义与Wikipedia的分类体系进行了融合集成，使得 Yago 具有更加丰富的实体分类体系。Yago 还考虑了时间和空间知识，为很多知识条目增加了时间和空间维度的属性描述。目前，Yago包含1.2亿条三元组知识。Yago也是IBM Watson的后端知识库之一。

Microsoft ConceptGraph 是以概念层次体系为中心的知识图谱。与 Freebase 等知识图谱不同，ConceptGraph 以概念定义和概念之间的 IsA 关系为主。例如，给定一个概念“Microsoft”,ConceptGraph返回一组与“微软”有IsA关系概念组“Company”“Software Company”“Largest OS Vender”等，被称为概念化“Conceptualization”。ConceptGraph可以用于短文本理解和语义消歧。例如，给定一个短文本“the engineer is eating the apple”，可以利用ConceptGraph正确理解其中“apple”的含义是“吃的苹果”还是“苹果公司”。微软发布的第一个版本包含超过540万个概念、1255万个实体和8760万个关系。ConceptGraph主要通过从互联网和网络日志中挖掘数据进行构建。

LOD（Linked Open Data）的初衷是为了实现Tim Berners-Lee在2006年发表的有关链接数据（Linked Data）作为语义网的一种实现的设想。LOD 遵循了 Tim 提出的进行数据链接的四个规则，即：使用URI标识万物；使用HTTP URI，以便用户可以（像访问网页一样）查看事物的描述；使用RDF和SPARQL标准；为事物添加与其他事物的URI链接，建立数据关联。LOD 已经有1143个链接数据集，其中社交媒体、政府、出版和生命科学四个领域的数据占比超过了90%。56%的数据集对外至少与一个数据集建立了链接。被链接最多的是 DBpedia 的数据。LOD 鼓励各个数据集使用公共的开放词汇和术语，但也允许使用各自的私有词汇和术语。在使用的术语中，有41%是公共的开放术语。

1.4.3 中文开放知识图谱

OpenKG 是一个面向中文域开放知识图谱的社区项目，主要目的是促进中文领域知识

图谱数据的开放与互联。OpenKG.CN　聚集了大量开放的中文知识图谱数据、工具及文献，如图1-7所示。典型的中文开放知识图谱数据包括百科类的Zhishi.me（狗尾草科技、东南大学）、CN-DBpedia（复旦大学）、XLore（清华大学）、Belief-Engine（中科院自动化所）、PKUPie（北京大学）、ZhOnto（狗尾草科技）等。OpenKG　对这些主要百科数据进行了链接计算和融合工作，并通过 OpenKG 提供开放的Dump或开放访问API，完成的链接数据集也向公众完全免费开放。此外，OpenKG 还对一些重要的知识图谱开源工具进行了收集和整理，包括知识建模工具 Protege、知识融合工具 Limes、知识问答工具YodaQA、知识抽取工具DeepDive等。

图1-7 OpenKG的主网站

知识图谱　Schema　定义了知识图谱的基本类、术语、属性和关系等本体层概念。cnSchema.ORG是OpenKG发起和完成的开放的知识图谱Schema标准。cnSchema的词汇集包括了上千种概念分类（classes）、数据类型（data　types）、属性（propertities）和关系（relations）等常用概念定义，以支持知识图谱数据的通用性、复用性和流动性。结合中文的特点，复用、连接并扩展了　Schema.org、Wikidata、Wikipedia　等已有的知识图谱 Schema　标准，为中文领域的开放知识图谱、聊天机器人、搜索引擎优化等提供可供参考和扩展的数据描述和接口定义标准。通过　cnSchema，开发者也可以快速对接上百万基于 Schema.org 定义的网站，以及 Bot 的知识图谱数据 API。cnSchema 主要解决如下三个问题：①Bots　是搜索引擎后新兴的人机接口，对话中的信息粒度缩小到短文本、实体和关系，要求文本与结构化数据的结合，要求更丰富的上下文处理机制等，这都需要　Schema 的支持；②知识图谱 Schema 缺乏对中文的支持；③知识图谱的构建成本高，容易重新发

明轮子，需要用合理的方法实现成本分摊。

OpenBase.AI 是 OpenKG 实现的类似于 Wikidata 的开放知识图谱众包平台。与 WikiData 不同，OpenBase 主要以中文为中心，更加突出机器学习与众包的协同，将自动化的知识抽取、挖掘、更新、融合与群智协作的知识编辑、众包审核和专家验收等结合起来。此外，OpenBase 还支持将图谱转化为 Bots，允许用户选择算法、模型、图谱数据等定制生成Bots，即时体验新增知识图谱的作用。

1.4.4 垂直领域知识图谱

领域知识图谱是相对于 DBPedia、Yago、Wikidata、百度和谷歌等搜索引擎在使用的知识图谱等通用知识图谱而言的，它是面向特定领域的知识图谱，如电商、金融、医疗等。相比较而言，领域知识图谱的知识来源更多、规模化扩展要求更迅速、知识结构更加复杂、知识质量要求更高、知识的应用形式也更加广泛。如表1-1所示，从多个方面对通用知识图谱和领域知识图谱进行了比较分析。下面以电商、医疗、金融领域知识图谱为例，介绍领域知识图谱的主要特点及技术难点。

表1-1 通用知识图谱与领域知识图谱的比较

分类 / 比较项目	通用知识图谱	领域知识图谱
知识来源及规模化	以互联网开放数据，如 Wikipedia 或社区众包为主要来源，逐步扩大规模	以领域或企业内部的数据为主要来源，通常要求快速扩大规模
对知识表示的要求	主要以三元组事实型知识为主	知识结构更加复杂，通常包含较为复杂的本体工程和规则型知识
对知识质量的要求	较多地采用面向开放域的 Web 抽取，对知识抽取质量有一定容忍度	知识抽取的质量要求更高，较多地依靠从企业内部的结构化、非结构化数据进行联合抽取，并依靠人工进行审核校验，保障质量
对知识融合的要求	融合主要起到提升质量的作用	融合多源的领域数据是扩大构建规模的有效手段
知识的应用形式	主要以搜索和问答为主要应用形式，对推理要求较低	应用形式更加全面，除搜索问答外，通常还包括决策分析、业务管理等，并对推理的要求更高，并有较强的可解释性要求
举例	DBpedia、Yago、百度、谷歌等	电商、医疗、金融、农业、安全等

1.电商领域知识图谱

以阿里巴巴电商知识图谱为例[27]，最新发布的知识图谱规模已达到百亿级别。其知识图谱数据主要以阿里已有的结构化商品数据为基础，并与行业合作伙伴数据、政府工商管理数据、外部开放数据进行融合扩展。在知识表示方面，除简单的三元组外，还包含层

次结构更加复杂的电商本体和面向业务管控的大量规则型知识。在知识的质量方面，对知识的覆盖面和准确性都有较高的要求。在应用形式方面，广泛支持商品搜索、商品导购、天猫精灵等产品的智能问答、平台的治理和管控、销售趋势的预测分析等多个应用场景。电商知识也具有较高的动态性特征，例如交易型知识和与销售趋势有关的知识都具有较强的时效性和时间性。

2.医疗领域知识图谱

医疗领域构建有大量的规模巨大的领域知识库。例如，仅 Linked Life Data 项目包含的RDF三元组规模就达到102亿个[3]，包含从基因、蛋白质、疾病、化学、神经科学、药物等多个领域的知识。再例如国内构建的中医药知识图谱[28]，通常需要融合各类基础医学、文献、医院临床等多种来源的数据，规模也达到20多亿个三元组。医学领域的知识结构更加复杂[29-31]，如医学语义网络 UMLS 包含大量复杂的语义关系，GeneOnto[29]则包含复杂的类层次结构。在知识质量方面，特别涉及临床辅助决策的知识库通常要求完全避免错误知识。

3.金融领域知识图谱

金融领域比较典型的例子如 Kensho 采用知识图谱辅助投资顾问和投资研究，国内以恒生电子为代表的金融科技机构以及不少银行、证券机构等也都在开展金融领域的知识图谱构建工作。金融知识图谱构建主要来源于机构已有的结构化数据和公开的公报、研报及新闻的联合抽取等。在知识表示方面，金融概念也具有较高的复杂性和层次性，并较多地依赖规则型知识进行投资因素的关联分析。在应用形式方面，则主要以金融问答和投顾投研类决策分析型应用为主。金融知识图谱的一个显著特点是高度动态性，且需要考虑知识的时效性，对金融知识的时间维度进行建模。

由上面的例子可以看出，如图1-8所示，领域知识图谱具有规模巨大、知识结构更加复杂、来源更加多样、知识更加异构、具有高度的动态性和时效性、更深层次的推理需求等特点。

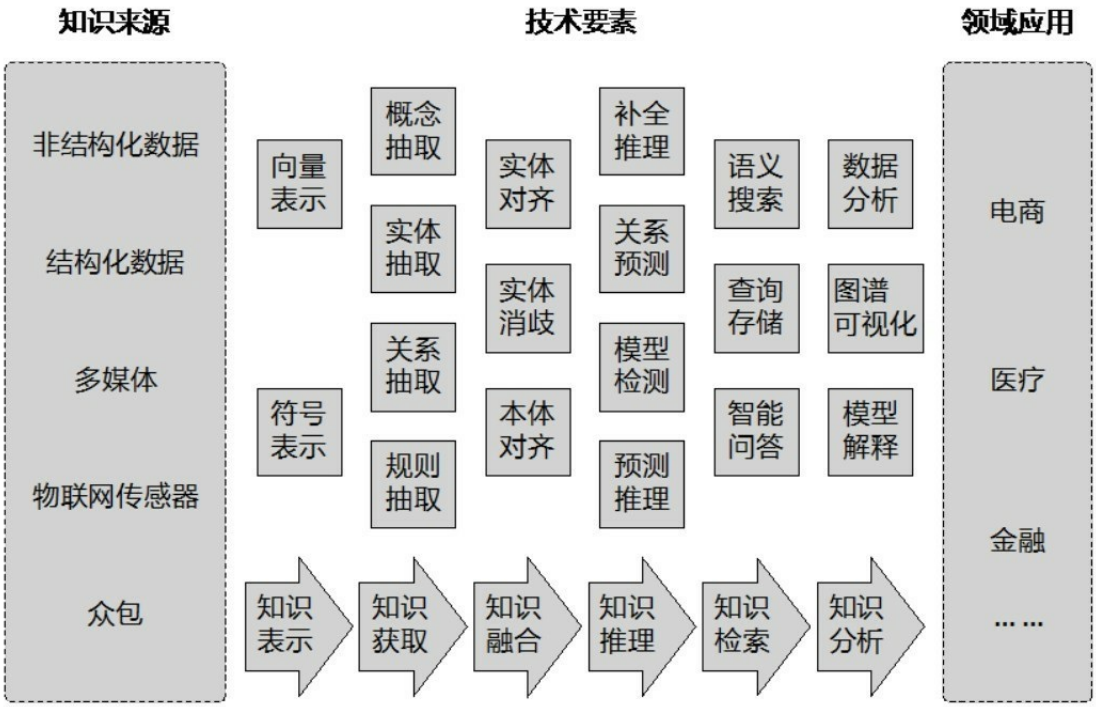

图1-8 规模化的知识图谱系统工程

1.5 知识图谱的技术流程

知识图谱用于表达更加规范的高质量数据。一方面，知识图谱采用更加规范而标准的概念模型、本体术语和语法格式来建模和描述数据；另一方面，知识图谱通过语义链接增强数据之间的关联。这种表达规范、关联性强的数据在改进搜索、问答体验、辅助决策分析和支持推理等多个方面都能发挥重要的作用。

知识图谱方法论涉及知识表示、知识获取、知识处理和知识利用多个方面。一般流程为：首先确定知识表示模型，然后根据数据来源选择不同的知识获取手段导入知识，接着综合利用知识推理、知识融合、知识挖掘等技术对构建的知识图谱进行质量提升，最后根据场景需求设计不同的知识访问与呈现方法，如语义搜索、问答交互、图谱可视化分析等。下面简要概述这些技术流程的核心技术要素。

1.知识来源

可以从多种来源获取知识图谱数据，包括文本、结构化数据库、多媒体数据、传感器数据和人工众包等。每一种数据源的知识化都需要综合各种不同的技术手段。例如，对于文本数据源，需要综合实体识别、实体链接、关系抽取、事件抽取等各种自然语言处理技术，实现从文本中抽取知识。

结构化数据库如各种关系数据库，也是最常用的数据来源之一。已有的结构化数据库通常不能直接作为知识图谱使用，而需要将结构化数据定义到本体模型之间的语义映射，再通过编写语义翻译工具实现结构化数据到知识图谱的转化。此外，还需要综合采用实体消歧、数据融合、知识链接等技术，提升数据的规范化水平，增强数据之间的关联。

语义技术也被用来对传感器产生的数据进行语义化。这包括对物联设备进行抽象，定义符合语义标准的数据接口；对传感数据进行语义封装和对传感数据增加上下文语义描述等。

人工众包是获取高质量知识图谱的重要手段。例如，Wikidata和Schema.org都是较为典型的知识众包技术手段。此外，还可以开发针对文本、图像等多种媒体数据的语义标注工具，辅助人工进行知识获取。

2.知识表示与**Schema**工程

知识表示是指用计算机符号描述和表示人脑中的知识，以支持机器模拟人的心智进行推理的方法与技术。知识表示决定了图谱构建的产出目标，即知识图谱的语义描述框架（Description Framework）、Schema 与本体（Ontology）、知识交换语法（Syntax）、实

体命名及ID体系。

基本描述框架定义知识图谱的基本数据模型（Data Model）和逻辑结构（Structure），如国际万维网联盟（World Wide Web Consortium,W3C）的RDF。Schema与本体定义知识图谱的类集、属性集、关系集和词汇集。交换语法定义知识实际存在的物理格式，如Turtle、JSON等。实体命名及ID体系定义实体的命名原则及唯一标识规范等。

按知识类型的不同，知识图谱包括词（Vocabulary）、实体（Entity）、关系（Relation）、事件（Event）、术语体系（Taxonomy）、规则（Rule）等。词一级的知识以词为中心，并定义词与词之间的关系，如 WordNet、ConceptNet 等。实体一级的知识以实体为中心，并定义实体之间的关系、描述实体的术语体系等。事件是一种复合的实体。

W3C 的 RDF 把三元组（Triple）作为基本的数据模型，其基本的逻辑结构包含主语（Subject）、谓词（Predicate）、宾语（Object）三个部分。虽然不同知识库的描述框架的表述有所不同，但本质上都包含实体、实体的属性和实体之间的关系几个要素。

3.知识抽取

知识抽取按任务可以分为概念抽取、实体识别、关系抽取、事件抽取和规则抽取等。传统专家系统时代的知识主要依靠专家手工录入，难以扩大规模。现代知识图谱的构建通常大多依靠已有的结构化数据资源进行转化，形成基础数据集，再依靠自动化知识抽取和知识图谱补全技术，从多种数据来源进一步扩展知识图谱，并通过人工众包进一步提升知识图谱的质量。

结构化和文本数据是目前最主要的知识来源。从结构化数据库中获取知识一般使用现有的 D2R 工具[32]，如 Triplify、D2RServer、OpenLink、SparqlMap、Ontop 等。从文本中获取知识主要包括实体识别和关系抽取。以关系抽取为例，典型的关系抽取方法可以分为基于特征模板的方法[33-35]、基于核函数的监督学习方法[36-44]、基于远程监督的方法[45,46]和基于深度学习的监督或远程监督方法，如简单 CNN、MP-CNN、MWK-CNN、PCNN、PCNN+ Att 和 MIMLCNN 等[47-52]。远程监督的思想是，利用一个大型的语义数据库自动获取关系类型标签。这些标签可能是含有噪声的，但是大量的训练数据在一定程度上可以抵消这些噪声。另外，一些工作通过多任务学习等方法将实体和关系做联合抽取[46,53]。最新的一些研究则利用强化学习减少人工标注并自动降低噪声[54]。

4.知识融合

在构建知识图谱时，可以从第三方知识库产品或已有结构化数据中获取知识输入。例如，关联开放数据项目（Linked Open Data）会定期发布其经过积累和整理的语义知识数据，其中既包括前文介绍过的通用知识库 DBpedia 和 Yago，也包括面向特定领域的知识库产品，如 MusicBrainz 和 DrugBank 等。当多个知识图谱进行融合，或者将外部关系数

据库合并到本体知识库时，需要处理两个层面的问题：通过模式层的融合，将新得到的本体融入已有的本体库中，以及新旧本体的融合；数据层的融合，包括实体的指称、属性、关系以及所属类别等，主要的问题是如何避免实例以及关系的冲突问题，造成不必要的冗余。

数据层的融合是指实体和关系（包括属性）元组的融合，主要是实体匹配或者对齐，由于知识库中有些实体含义相同但是具有不同的标识符，因此需要对这些实体进行合并处理[55,56]。此外，还需要对新增实体和关系进行验证和评估，以确保知识图谱的内容一致性和准确性，通常采用的方法是在评估过程中为新加入的知识赋予可信度值，据此进行知识的过滤和融合。实体对齐的任务包括实体消歧和共指消解，即判断知识库中的同名实体是否代表不同的含义以及知识库中是否存在其他命名实体表示相同的含义。实体消歧专门用于解决同名实体产生歧义的问题，通常采用聚类法，其关键问题是如何定义实体对象与指称项之间的相似度，常用方法有空间向量模型（词袋模型）[57]、语义模型[58]、社会网络模型[59]、百科知识模型[60]和增量证据模型[61]。一些最新的工作利用知识图谱嵌入方法进行实体对齐，并引入人机协作方式提升实体对齐的质量[62,63]。

本体是针对特定领域中 Schema 定义、概念模型和公理定义而言的，目的是弥合词汇异构性和语义歧义的间隙，使沟通达成共识。这种共识往往通过一个反复的过程达到，每次迭代都是一次共识的修改。因此，本体对齐通常带来的是共识模式的演化和变化，本体对齐的主要问题之一也可以转化为怎样管理这种演化和变化[64]。常见的本体演化管理框架有KAON[65]、Conto-diff[66]、OntoView等。

5.知识图谱补全与推理

常用的知识图谱补全方法包括：基于本体推理的补全方法，如基于描述逻辑的推理[67-69]，以及相关的推理机实现，如 RDFox、Pellet、RACER、HermiT、TrOWL 等。这类推理主要针对TBox，即概念层进行推理，也可以用来对实体级的关系进行补全。

另外一类的知识补全算法实现基于图结构和关系路径特征的方法，如基于随机游走获取路径特征的 PRA 算法[70]、基于子图结构的 SFE 算法[71]、基于层次化随机游走模型的 PRA 算法[72]。这类算法的共同特点是通过两个实体节点之间的路径，以及节点周围图的结构提取特征，并通过随机游走等算法降低特征抽取的复杂度，然后叠加线性的学习模型进行关系的预测。此类算法依赖于图结构和路径的丰富程度。

更为常见的补全实现是基于表示学习和知识图谱嵌入的链接预测[73-80]，简单的如前面介绍最基本的翻译模型、组合模型和神经元模型等。这类简单的嵌入模型一般只能实现单步的推理。对于更为复杂的模型，如向量空间中引入随机游走模型的方法，在同一个向量空间中将路径与实体和关系一起表示出来再进行补全的模型[81,82]。

文本信息也被用来辅助实现知识图谱的补全[83-88]。例如，Jointly(w)、Jointly(z)、DKRL、TEKE、SSP 等方法将文本中的实体和结构化图谱中的实体对齐，然后利用双方的语义信息辅助实现关系预测或抽取。这类模型一般包含三个部分：三元组解码器、文本解码器和联合解码器。三元组解码器将知识图谱中的实体和关系转化为低维向量；文本解码器则要从文本语料库中学习实体（词）的向量表示；联合解码器的目的是要保证实体、关系和词的嵌入向量位于相同的空间中，并且集成实体向量和词向量。

6.知识检索与知识分析

基于知识图谱的知识检索的实现形式主要包括语义检索和智能问答。传统搜索引擎依靠网页之间的超链接实现网页的搜索，而语义搜索直接对事物进行搜索，如人物、机构、地点等。这些事物可能来自文本、图片、视频、音频、IoT 设备等各种信息资源。而知识图谱和语义技术提供了关于这些事物的分类、属性和关系的描述，使得搜索引擎可以直接对事物进行索引和搜索。

知识图谱和语义技术也被用来辅助做数据分析与决策。例如，大数据公司 Plantir 基于本体融合和集成多种来源的数据，通过知识图谱和语义技术增强数据之间的关联，使得用户可以用更加直观的图谱方式对数据进行关联挖掘与分析。近年来，描述性数据分析（Declarative Data Analysis）越来越受到重视[89]。描述性数据分析是指依赖数据本身的语义描述实现数据分析的方法。不同于计算性数据分析主要以建立各种数据分析模型，如深度神经网络，描述性数据分析突出预先抽取数据的语义，建立数据之间的逻辑，并依靠逻辑推理的方法（如DataLog）实现数据分析[90]。

1.6 知识图谱的相关技术

知识图谱是交叉领域，涉及的相关领域包括人工智能、数据库、自然语言处理、机器学习、分布式系统等。下面分别从数据库系统、智能问答、机器推理、推荐系统、区块链与去中心化等角度介绍知识图谱有关的相关技术进展。

1.6.1 知识图谱与数据库系统

随着知识图谱规模的日益增长，知识图谱数据管理问题愈加突出。近年来，知识图谱和数据库领域均认识到大规模知识图谱数据管理任务的紧迫性。由于传统关系数据库无法有效适应知识图谱的图数据模型，知识图谱领域形成了 RDF 数据的三元组库（Triple Store），数据库领域开发了管理属性图的图数据库（Graph Database）。

知识图谱的主要数据模型有RDF图（RDF graph）和属性图（Property Graph）两种；知识图谱查询语言可分为声明式（Declarative）和导航式（Navigational）两类。

RDF 三元组库主要是由 Semantic Web 领域推动开发的数据库管理系统，其数据模型RDF图和查询语言SPARQL均遵守W3C标准。查询语言SPARQL从语法上借鉴了SQL语言，属于声明式查询语言。最新的 SPARQL 1.1版本[91]为有效查询 RDF 三元组集合设计了三元组模式（Triple Pattern）、基本图模式（Basic Graph Pattern）、属性路径（Property Path）等多种查询机制。

图数据库是数据库领域为更好地存储和管理图模型数据而开发的数据库管理系统，其数据模型采用属性图，其上的声明式查询语言有：Cypher[92]、PGQL[93]和 G-Core[94]。Cypher 是开源图数据库 Neo4j 中实现的图查询语言。PGQL 是 Oracle 公司开发的图查询语言。G-Core是由LDBC（Linked Data Benchmarks Council）组织设计的图查询语言。考虑到关系数据库采用统一的查询语言SQL，目前学术和工业界关于开发统一图数据库语言的呼声越来越高。

目前，基于三元组库和图数据库能够提供的知识图谱数据存储方案可分为三类：

（1）基于关系的存储方案。包括三元组表、水平表、属性表、垂直划分、六重索引和DB2RDF等。

三元组表是将知识图谱中的每条三元组存储为一行具有三列的记录（主语，谓语，宾语）。三元组表存储方案虽然简单明了，但三元组表的行数与知识图谱的边数一样，其问

题是将知识图谱查询翻译为SQL后会产生大量三元组表的自连接操作，影响效率。

水平表存储方案的每行记录存储知识图谱中一个主语的所有谓语和宾语，相当于知识图谱的邻接表。但其缺点在于所需列数目过多，表中产生大量空值，无法存储多值宾语等。

属性表存储方案将同一类主语分配到一个表中，是对水平表存储方案的细化。属性表解决了三元组表的自连接问题和水平表的列数目过多问题。但对于真实大规模知识图谱，属性表的问题包括：所需属性表过多，复杂查询的多表连接效率，空值问题和多值宾语问题。

垂直划分存储方案为知识图谱中的每种谓语建立一张两列的表（主语，宾语），表中存放由该谓语连接的主语和宾语，支持“主语-主语”作为连接条件的查询操作的快速执行。垂直划分有效解决了空值问题和多值宾语问题；但其仍有缺点，包括：大规模知识图谱的谓语表数目过多、复杂查询表连接过多、更新维护代价大等。

六重索引存储方案是将三元组全部6种排列对应地建立为6张表。六重索引通过“空间交换时间”策略有效缓解了三元组表的自连接问题，但需要更多的存储空间开销和索引更新维护代价。

DB2RDF 存储方案[95]是一种较新的基于关系的知识图谱存储方案，是以往存储方案的一种权衡优化。三元组表的灵活性体现在“行维度”上，无论多少行三元组数据，表模式只有3列固定不变；DB2RDF 方案将这种灵活性推广到了“列维度”，列名称不再和谓语绑定，将同一主语的所有谓语和宾语动态分配到某列。

（2）面向 RDF 的三元组库。主要的 RDF 三元组库包括：商业系统 Virtuoso、AllegroGraph、GraphDB和BlazeGraph，开源系统Jena、RDF-3X和gStore[96]。

RDF4J目前是Eclipse基金会旗下的开源孵化项目，其功能包括RDF数据的解析、存储、推理和查询等。RDF4J 本身提供内存和磁盘两种 RDF 存储机制，支持全部的 SPARQL 1.1查询和更新语言，可以使用与访问本地 RDF 库相同的 API 访问远程 RDF 库，支持所有主流的 RDF 数据格式，包括 RDF/XML、Turtle、N-Triples、N-Quads、JSON-LD、TriG和TriX。RDF4J框架的主要特点是其模块化的软件架构设计。

RDF-3X 是德国马克斯·普朗克计算机科学研究所开发的三元组数据库，其特点是为RDF优化设计的物理存储方案和查询处理方法，是实现六重索引的典型系统。

gStore 是由北京大学、加拿大滑铁卢大学和香港科技大学联合研究项目开发的基于图的 RDF三元组数据库。gStore 的底层存储使用 RDF图对应的标签图（Signature Graph）并建立“VS树”索引结构以加速查找。gStore系统提出建立“VS树”索引，其基本思想实际上是为标签图G*建立不同详细程度的摘要图（Summary Graph）；利用“VS树”索引提供的摘要图，gStore系统提出可以大幅削减SPARQL查询的搜索空间，以加快查询速度。

Virtuoso 是由 OpenLink 公司开发的商业混合数据库产品，支持关系数据、对象-关系数据、RDF 数据、XML 数据和文本数据的统一管理。因为 Virtuoso 可以较为完善地支持W3C的Linked Data系列协议，包括DBpedia在内的很多开放RDF知识图谱选择其作为后台存储系统。

AllegroGraph 是美国 Franz 公司开发的 RDF 三元组数据库。AllegroGraph 遵循对W3C语义 Web 相关标准的严格支持，包括 RDF、RDFS、OWL 和 SPARQL 等。AllegroGraph 对语义推理功能具有较为完善的支持。AllegroGraph 除了三元组数据库的基本功能，还支持动态物化的 RDFS++推理机、OWL2 RL 推理机、Prolog 规则推理系统、时空推理机制、社会网络分析库、可视化RDF图浏览器等。

GraphDB是由保加利亚的Ontotext软件公司开发的RDF三元组数据库。GraphDB实现了 RDF4J 框架的 SAIL 层，与 RDF4J API 无缝对接，也就是说，可以使用 RDF4J 的RDF模型、解析器和查询引擎直接访问GraphDB。GraphDB的特色是良好支持RDF推理功能，其使用内置的基于规则的“前向链”（Forward-Chaining）推理机，由显式知识经过推理得到导出知识，对这些导出知识进行优化存储；导出知识会在知识库更新后相应地同步更新。

Blazegraph 是一个基于 RDF 三元组库的图数据库管理系统，在用户接口层同时支持RDF三元组和属性图模型，既实现了SPARQL语言，也实现了Blueprints标准及Gremlin语言。通过分布式动态分片 B+树和服务总线技术，Blazegraph 支持真正意义上的集群分布式存储和查询处理。正是缘于此，Blazegraph 在与 Neo4j 和 Titan 的竞争中脱颖而出，被Wikidata选为查询服务的后台图数据库系统。

Stardog是由美国Stardog Union公司开发的RDF三元组数据库，支持RDF图数据模型、SPARQL 查询语言、属性图模型、Gremlin 图遍历语言、OWL2标准、用户自定义的推理与数据分析规则、虚拟图、地理空间查询以及多用编程语言与网络接口支持。Stardog虽然发布较晚，但其对 OWL2推理机制具有良好的支持，同时具备全文搜索、GraphQL查询、路径查询、融合机器学习任务等功能，能够支持多种不同编程语言和 Web 访问接口，使得Stardog成了一个知识图谱数据存储和查询平台。

（3）原生图数据库。Neo4j是用Java实现的开源图数据库。可以说Neo4j是目前流行程度最高的图数据库产品。Neo4j 的不足之处在于其社区版是单机系统，虽然 Neo4j 企业版支持高可用性（High availability）集群，但其与分布式图存储系统的最大区别在于每个节点上存储图数据库的完整副本（类似于关系数据库镜像的副本集群），而不是将图数据划分为子图进行分布式存储，而并非真正意义上的分布式数据库系统。如果图数据超过一定规模，系统性能就会因为磁盘、内存等限制而大幅降低。

JanusGraph 是在原有 Titan 系统基础上继续开发的开源分布式图数据库，目前是Linux

基金会旗下的一个开源项目。JanusGraph 的存储后端与查询引擎是分离的，由于其可使用分布式Bigtable存储库Cassandra或HBase作为存储后端，因此JanusGraph自然就成了分布式图数据库。JanusGraph 的主要缺点是分布式查询功能仅限于基于 Cassandra 或HBase 提供的分布式读写实现的简单导航查询，对于很多稍复杂的查询类型，目前还不支持真正意义上的分布式查询处理，例如子图匹配查询、正则路径查询等。

OrientDB最初是由OrientDB公司开发的多模型数据库管理系统。OrientDB虽然支持图、文档、键值、对象、关系等多种数据模型，但其底层实现主要面向图和文档数据存储管理的需求设计。其存储层中数据记录之间的关联并不像关系数据库那样通过主外键的引用，而是通过记录之前直接的物理指针。

Cayley 是由谷歌公司工程师开发的一款轻量级开源图数据库，于2014年6月在GitHub上发布。Cayley 的开发受到了 Freebase 知识图谱和谷歌知识图谱背后的图数据存储的影响，目标是成为开发者管理 Linked Data 和图模型数据（语义 Web、社会网络等）的有效工具之一。

总体来讲，基于关系的存储系统继承了关系数据库的优势，成熟度较高，在硬件性能和存储容量满足的前提下，通常能够适应千万到十亿三元组规模的管理。官方测评显示，关系数据库Oracle 12c配上空间和图数据扩展组件（Spatial and Graph）可以管理的三元组数量高达1.08万亿条[97]。对于一般在百万到上亿三元组的管理，使用稍高配置的单机系统和主流RDF三元组数据库（如Jena、RDF4J、Virtuoso等）完全可以胜任。如果需要管理几亿到十几亿以上大规模的 RDF 三元组，则可尝试部署具备分布式存储与查询能力的数据库系统（如商业版的 GraphDB 和 BlazeGraph、开源的 JanusGraph 等）。近年来，以Neo4j为代表的图数据库系统发展迅猛，使用图数据库管理RDF三元组也是一种很好的选择；但目前大部分图数据库还不能直接支持 RDF 三元组存储，对于这种情况，可采用数据转换方式，先将RDF预处理为图数据库支持的数据格式（如属性图模型），再进行后续管理操作。

目前，还没有一种数据库系统被公认为是具有主导地位的知识图谱数据库。但可以预见，随着三元组库和图数据库的相互融合发展，知识图谱的存储和数据管理手段将愈加丰富和强大。

1.6.2 知识图谱与智能问答

基于知识图谱的问答（Knowledge-based Question Answering,KBQA，下称“知识问答”）是智能问答系统的核心功能，是一种人机交互的自然方式。知识问答依托一个大型知识库（知识图谱、结构化数据库等），将用户的自然语言问题转化成结构化查询语句

（如SPARQL、SQL等），直接从知识库中导出用户所需的答案。

近几年，知识问答聚焦于解决事实型问答，问题的答案是一个实义词或实义短语。如“中国的首都是哪个城市？北京”或“菠菜是什么颜色的？绿色”。事实型问题按问题类型可分为单知识点问题（Single-hop Questions）和多知识点问题（Multi-hop Questions）；按问题的领域可分为垂直领域问题和通用领域问题。相对于通用领域或开放领域，垂直领域下的知识图谱规模更小、精度更高，知识问答的质量更容易提升。

知识问答技术的成熟与落地不仅能提高人们检索信息的精度和效率，还能提升用户的产品体验。无论依托的知识库的规模如何，用户总能像“跟人打交道一样”使用自然语言向机器提问并得到反馈，便利性与实用性共存。

攻克知识问答的关键在于理解并解析用户提出的自然语言问句。这涉及自然语言处理、信息检索和推理（Reasoning）等多个领域的不同技术。相关研究工作在近五年来受到越来越多国内外学者的关注，研究方法主要可分为三大类：基于语义解析（Semantic Parsing）的方法、基于信息检索（Information Retrieval）的方法和基于概率模型（Probabilistic Models）的方法。

大部分先进的知识问答方法是基于语义解析的，目的是将自然语言问句解析成结构化查询语句，进而在知识库上执行查询得到答案。通常，自然语言问句经过语义解析后，所得的语义结构能解释答案的产生。在实际工程应用中，这一点优势不仅能帮助用户理解答案的产生，还能在产生错误答案时帮助开发者定位错误的可能来源。

微软在利用语义解析技术解决单知识点问答（Single-hop Question Answering）中有突出贡献。2014年，叶等人[98]指出，解决单知识点问答的关键在于将原任务分解为两个子任务——话题词识别和关系检测。如回答“姚明的妻子是谁？”，可先通过计算语义相似性将问句解析成形如“（姚明，妻子，?）”的查询。其中，话题词是“姚明”，问题中包含的关系为“妻子”（或“配偶”），再在知识库中执行查询，得到答案。2015年，叶等人[99]强调，直接从大型知识库中寻找与问句含义匹配的关系是比较困难的。论文中首先采用实体链接（Entity Linking）定位话题词，再从与话题词相关的关系子集中寻找与问句含义匹配的关系，从而将问句解析成一个结构化的查询。2016年，叶等人[100]继承了斯坦福自然语言处理组开源的 WebQuestions 数据集，并在此基础上标注了问题的语义解析结果（SPARQL查询），贡献了WebQuestionsSP数据集。

在基于语义解析的方法训练过程中，问答模型隐式地学习了标注数据中蕴涵的语法解析规律。这使得模型能具有更好的可解释性。但是，数据标注需要花费大量的人力和财力，是不切实际的。而基于信息检索的方法回避了这个问题。基于信息检索的知识问答大致可分为两步：①通过粗粒度信息检索，在知识库中直接筛选出候选答案；②根据问句中抽取出的特征，对候选答案进行排序。这就要求模型对问句的语义有充分的理解。而在自

然语言中，词语同义替换等语言现象提升了理解问题的难度。[102]

为了实现有效的信息检索式知识问答，学者们聚焦于如何让机器理解用户的问题，以及掌握问题与知识库间的匹配规律。可行的方法包括：

●集成额外的文本信息[101]，如Wikipedia或搜索引擎结果；

●提出更多、更复杂的网络结构，如多列卷积神经网络[102]（Multi-Column Convolutional Neural Networks,MCCNN）、深度残差双向长短时记忆网络[6]（Deep Residual Bidirectional Long Short-term Memory Network）和注意力最大池化层[103]（Attentive Max Pooling Layer）；

●联合训练[104]实体链接和关系检测两个模块。

除上述两大流派外，有部分学者将知识问答问题看作是一个条件概率问题[105,106]，即是要求给定问句 Q 时，答案为 α 的概率 P（A=α|Q），进而引入概率分解[9]或变分推理[107]的技巧，将目标概率分而治之。

大部分已有的知识问答解决方案都停留在回答单知识点事实型问题上。在这类问题中，基于语义解析的方法和基于信息检索的方法并不呈完全割裂、对立的关系[1]。二者几乎都把知识问答看作是话题词识别和关系检测两个子任务串行。在一些论文中，学者们声称单知识点问答已接近人类水平[108]。

未来，学者们必然将更多的精力投入解决复杂的多知识点事实型问答上。这类问题涉及的自然语言现象更丰富，如关系词的词汇组合性[109]（Sub-Lexical Compositionality）、多关系词间语序等。另外一种思路是：研究如何将多知识点问题转化为单知识点问题。因此，先进的单知识点问答模型直接被复用。

除此之外，在理解问题、回答问题的过程中，模型应具备更强的推理能力和更好的可解释性。更强的推理能力能满足用户的复杂提问需求。更好的可解释性使用户在“知其然”的同时“知其所以然”。

1.6.3 知识图谱与机器推理

推理是指基于已知的事实或知识推断得出未知的事实或知识的过程。典型的推理包括演绎推理（Deductive Reasoning）、归纳推理（Inductive Reasoning）、溯因推理（Abductive Reasoning）、类比推理（Analogical Reasoning）等。在知识图谱中，推理主要用于对知识图谱进行补全（Knowledge Base Completion,KBC）和知识图谱质量的校验。

知识图谱中的知识可分为概念层和实体层。知识图谱推理的任务是根据知识图谱中已有的知识推理出新的知识或识别出错误的知识。其中，概念层的推理主要包括概念之间的包含关系推理，实体层的推理主要包括链接预测与冲突检测，实体层与概念层之间的推理主要包括实例检测。推理的方法主要包含基于规则的推理、基于分布式表示学习的推理、基于神经网络的推理以及混合推理。

1.基于规则的推理

基于规则的推理通过定义或学习知识中存在的规则进行推理。根据规则的真值类型，可分为硬逻辑规则和软逻辑规则。硬逻辑规则中的每条规则的真值都为1，即绝对正确，人工编写的规则多为硬逻辑规则。软逻辑规则即每条规则的真值为区间在0到1之间的概率，规则挖掘系统的结果多为软逻辑规则，其学习过程一般是基于规则中结论与条件的共现特征，典型方法有 AMIE[110]等。软逻辑规则可通过真值重写转化为硬逻辑规则。硬逻辑规则可写成知识图谱本体中的SWRL规则，然后通过如Pellet、Hermit等本体推理机进行推理。规则推理在大型知识图谱上的效率受限于它的离散性，Cohen 提出了一个可微的规则推理机TensorLog[111]。

基于规则的推理方法最主要的优点是在通常情况下规则比较接近人思考问题时的推理过程，其推理结论可解释，所以对人比较友好。在知识图谱中已经沉淀的规则具有较好的演绎能力。

2.基于分布式表示学习的推理

分布式表示学习的核心是将知识图谱映射到连续的向量空间中，并为知识图谱中的元素学习分布式表示为低维稠密的向量或矩阵。分布式表示学习通过各元素的分布式表示之间的计算完成隐式的推理。多数表示学习方法以单步关系即单个三元组为输入和学习目标，不同的分布式表示学习方法对三元组的建模基于不同的空间假设。例如，以TransE[112]为代表的Trans系列模型基于的是关系向量表示在空间中的平移不变性，故将关系向量看作是头实体向量到尾实体向量的翻译并采用向量加法模拟；以 DistMult[113]为代表的线性转换模型将关系表示为矩阵，头实体的向量可经过关系矩阵的线性变换转换为尾实体；以 RESCAL[114]为代表的模型将知识图谱表示为高维稀疏的三维张量，通过张量分解得到实体和关系的表示。考虑到知识图谱中的多步推理的表示学习方法有PTransE[115]和CVSM[116]。

3.基于神经网络的推理

基于神经网络的推理通过神经网络的设计模拟知识图谱推理，其中 NTN[117]用一个双线性张量层判断头实体和尾实体的关系，ConvE[118]等在实体和关系的表示向量排布出的二维矩阵上采用卷机神经网络进行链接预测，R-GCN[119]通过图卷积网络捕捉实体的相邻

实体信息，IRN[120]采用记忆矩阵以及以递归神经网络为结构的控制单元模拟多步推理的过程。基于神经网络的知识图谱推理表达能力强，在链接预测等任务上取得了不错的效果。网络结构的设计多样，能够满足不同的推理需求。

4.混合推理

混合推理一般结合了规则、表示学习和神经网络。例如，NeuralLP[121]是一种可微的知识图谱推理方法，融合了关系的表示学习、规则学习以及循环神经网络，由 LSTM 生成多步推理中的隐变量，并通过隐变量生成在多步推理过程中对每种关系的注意力。DeepPath[122]和 MINERVA[123]用强化学习方法学习知识图谱多步推理过程中的路径选择策略。RUGE[124]将已有的推理规则输入知识图谱表示学习过程中，约束和影响表示学习结果并取得更好的推理效果。文献[125]使用了对抗生成网络（GAN）提升知识图谱表示学习过程中的负样本生成效率。混合推理能够结合规则推理、表示学习推理以及神经网络推理的能力并实现优势互补，能够同时提升推理结果的精确性和可解释性。

基于规则的知识图谱推理研究主要分为两部分：一是自动规则挖掘系统，二是基于规则的推理系统。目前，二者的主要发展趋势是提升规则挖掘的效率和准确度，用神经网络结构的设计代替在知识图谱上的离散搜索和随机游走是比较值得关注的方向。

基于表示学习的知识图谱推理研究的主要研究趋势是，一方面提高表示学习结果对知识图谱中含有的语义信息的捕捉能力，目前的研究多集中在链接预测任务上，其他推理任务有待跟进研究；另一方面是利用分布式表示作为桥梁，将知识图谱与文本、图像等异质信息结合，实现信息互补以及更多样化的综合推理。

基于神经网络的知识表示推理的主要发展趋势是设计更加有效和有意义的神经网络结构，来实现更加高效且精确的推理，通过对神经网络中间结果的解析实现对推理结果的部分解释是比较值得关注的方向。

1.6.4 知识图谱与推荐系统

随着互联网技术的飞速发展，各种信息在互联网上汇集，信息呈指数级增长，人们面临着信息过载的问题，推荐系统的提出是解决这一问题的有力途径。但是，推荐系统在启动阶段往往效果不佳，存在冷启动问题，而且用户历史记录数据往往较为稀疏，使得推荐算法的性能很难让用户满意。知识图谱作为先验知识，可以为推荐算法提供语义特征，引入它们可以有效地缓解数据稀疏问题，提高模型的性能。

基于知识图谱的推荐模型大部分是以现有的推荐模型为基础的，如基于协同过滤和基于内容的推荐模型，将知识图谱中关于商品、用户等实体的结构化知识加入推荐模型中，

通过引入额外的知识改善早期推荐模型中数据稀疏的问题。文献[126]提出了利用 DBpedia 知识图谱中的层次类别信息应用于推荐任务中，他们通过传播激活算法在知识图谱中寻找推荐实体。文献[127]通过计算知识图谱中蕴涵的语义距离建立音乐推荐模型。下面分别介绍三类利用知识图谱的推荐模型，分别为：基于知识图谱中元路径的推荐模型、基于概率逻辑程序的推荐模型、基于知识图谱表示学习技术的推荐模型。

考虑到知识图谱是一个表示不同实体之间关系的图，研究人员利用图上路径的连通信息计算物品之间的相似度[128]。研究人员通过元路径的概念利用图的信息[129,130]，元路径是图中不同类型实体和关系构成的路径。文献[131]利用元路径在图上传播用户偏好，并结合传统的协同过滤模型，最终实现了个性化的推荐模型。其具体方法如下：首先，沿着不同元路径利用路径相似度计算用户对不同物品的偏好，最终学得在元路径P下的偏好矩阵。针对每条元路径学得偏好矩阵，通过潜在因子模型对每个偏好矩阵进行分解，最终可获得每条路径下用户和物品的潜在因子矩阵，最终通过对每条路径下推荐结果的求和获得最终的全局推荐模型。其工作有效地利用了知识图谱中不同类型实体间路径的语义信息传递用户的偏好，但是路径需要人工选择。

文献[132]提出了基于概率逻辑程序的推荐模型，文献作者[133]将推荐问题形式化为逻辑程序，该逻辑程序对目标用户按查询得分高低输出推荐物品的结果，最终寻找到目标用户的推荐物品。文献作者提出了三种不同的推荐方法，分别为 EntitySim、TypeSim 和 GraphLF，性能超过了以前的最佳方法[134]。这三种方法都是基于通用目的的概率逻辑系统ProPPR。其中，EntitySim 方法只使用图上的连接信息；TypeSim 方法使用了实体的type信息，GraphLF 提出了一个结合概率逻辑程序和用户物品潜在因子模型的方法。他们的基本思路类似于文献[134]的工作，通过规则在知识图谱中传递用户的偏好，解决了路径人工选择的问题。但是，他们将推荐的流程分为寻找用户偏好实体和通过偏好实体寻找物品两个步骤，导致无法有效地利用物品与物品之间的关系和用户与用户之间的关系。例如，在电影推荐的例子中，电影《谍影重重4》是《谍影重重3》的续集，但是《谍影重重4》更换了主演，而如果通过他们方法中的规则，用户无法通过《谍影重重3》的主演马特·达蒙寻找到《谍影重重4》。

通过知识图谱表示学习技术，可以获得知识图谱中实体和关系的低维稠密向量，其可以在低维的向量空间中计算实体间的关联性，与传统的基于符号逻辑在图上查询和推理的方法相比，大大降低了计算的复杂度。文献[134]提出使用知识图谱表示学习技术提取知识图谱中的特征，以该特征向量使用 K 近邻的方法寻找用户最相近的物品，但是该模型与推荐模型结合较为松散，仅使用知识图谱表示学习作为特征提取的一种方法。

文献[135]在王灏等人[136]工作的基础上进行扩展，通过表示学习的方法将知识图谱中

的信息加入推荐模型中，提出了协同知识图谱表示学习的推荐模型（Collaborative Knowledge Base Embedding Recommender System），他们方法的具体思路如下：首先，通过知识表示学习获得知识图谱中和推荐物品相关的结构化信息，通过去噪编码器网络从物品相关的文本中学习编码层的文本表示向量，并通过和文本建模相似的去噪编码器网络从图像中学习视觉表示向量，并将这些表示向量引入物品的潜在因子向量中，结合矩阵分解算法完成推荐。该工作通过贝叶斯理论的角度解释并联合了不同算法的优化目标。但是，在推荐领域的知识图谱中，实体之间的关系非常稠密，且关系类型较少。以 TransE 为代表的模型不适合处理一对多、多对多的关系，尽管 TransR 针对该问题进行了一定的改进，但当应对相同类型关系的一对多、多对一和多对多关系时，算法实际退化为TransE。因此，本书在协同过滤算法上引入一类新的知识图谱表示学习的技术[137,138]提取知识图谱中的结构化信息，最终提出了一个基于知识图谱表示学习的协同过滤推荐系统。

1.6.5 区块链与去中心化的知识图谱

语义网的早期理念实际上包含三个方面：知识的互联、去中心化的架构和知识的可信。知识图谱在一定程度上实现了“知识互联”的理念，然而在去中心化的架构和知识可信两个方面都仍然没有出现较好的解决方案。

对于去中心化，相比起现有的多为集中存储的知识图谱，语义网强调知识以分散的方式互联和相互链接，知识的发布者拥有完整的控制权。近年来，国内外已经有研究机构和企业开始探索通过区块链技术实现去中心化的知识互联。这包括去中心化的实体 ID 管理、基于分布式账本的术语及实体命名管理、基于分布式账本的知识溯源、知识签名和权限管理等。

知识的可信与鉴真也是当前很多知识图谱项目面临的挑战和问题。由于很多知识图谱数据来源广泛，且知识的可信度量需要作用到实体和事实级别，怎样有效地对知识图谱中的海量事实进行管理、追踪和鉴真，也成为区块链技术在知识图谱领域的一个重要应用方向。

此外，将知识图谱引入智能合约（Smart Contract）中，可以帮助解决目前智能合约内生知识不足的问题。例如，PCHAIN[139]引入知识图谱 Oracle 机制，解决传统智能合约数据不闭环的问题。

1.7 本章小结

知识图谱本身可以看作是一种新型的信息系统基础设施。从数据维度上看，知识图谱要求用更加规范的语义提升企业数据的质量，用链接数据的思想提升企业数据之间的关联度，终极目标是将非结构、无显示关联的粗糙数据逐步提炼为结构化、高度关联的高质量知识。每个企业都应该将知识图谱作为一种面向数据的信息系统基础设施进行持续性建设。

从技术维度上看，知识图谱的构建涉及知识表示、关系抽取、图数据存储、数据融合、推理补全等多方面的技术，而知识图谱的利用涉及语义搜索、知识问答、自动推理、知识驱动的语言及视觉理解、描述性数据分析等多个方面。要构建并利用好知识图谱，也要求系统性地综合利用来自知识表示、自然语言处理、机器学习、图数据库、多媒体处理等多个相关领域的技术，而非单个领域的单一技术。因此，知识图谱的构建和利用都应注重系统思维是未来的一种发展趋势。

互联网促成了大数据的集聚，大数据进而促进了人工智能算法的进步。新数据和新算法为规模化知识图谱构建提供了新的技术基础和发展条件，使得知识图谱构建的来源、方法和技术手段都发生了极大的变化。知识图谱作为知识的一种形式，已经在语义搜索、智能问答、数据分析、自然语言理解、视觉理解、物联网设备互联等多个方面发挥出越来越大的价值。AI　　浪潮愈演愈烈，而作为底层支撑的知识图谱赛道也从鲜有问津到缓慢升温，虽然还谈不上拥挤，但作为通往未来的必经之路，注定会走上风口。