计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：大规模企业级知识图谱实践综述
作者：王昊奋，丁军，胡芳槐，王鑫
DOI：10.19678/j.issn.1000-3428.0057869
网络首发日期：2020-04-17
引用格式：王昊奋，丁军，胡芳槐，王鑫．大规模企业级知识图谱实践综述．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.0057869

网络首发时间：2020-04-17 10:28:14
网络首发地址：http://kns.cnki.net/kcms/detail/31.1289.TP.20200417.0933.005.html

计 算 机 工 程
Computer Engineering

大规模企业级知识图谱实践综述

王昊奋[1]，丁军[2]，胡芳槐[2]，王鑫[3]
（1. 同济大学，设计创意学院，上海 200092；2. 海义知信息科技（南京）有限公司，南京 210008；3. 天津大学，智能与计算学部，天津 300354）

摘　要：近年来，知识图谱及其相关技术飞速发展，在工业界各种认知智能场景中得到广泛应用。在简述知识图谱相关研究的基础上，介绍知识图谱在工程应用中的关键技术。研究工业级知识图谱的典型应用场景与案例、具有代表性的工业级知识图谱平台以及知识图谱生命周期过程中的相关可用工具，分析企业级知识图谱平台的构建需求和面临的挑战，提出企业级知识图谱平台构建的方法及过程。针对平台化建设中遇到的问题给出相应的知识图谱中台解决方案，并对知识图谱未来的发展与挑战进行展望。

关键词：知识图谱；表示学习；知识抽取；知识存储；知识推理；企业级知识图谱平台；知识图谱中台

开放科学（资源服务）标志码（OSID）：

Survey on Large Scale Enterprise-level KG System Practices

Wang Haofen[1], Ding Jun[2], Hu Fanghuai[2], Wang Xin[3]
(1. College of Design and Innovation, Tongji University, Shanghai 200092, China; 2. Haiyizhi Info Technology (Nanjing) Co., Ltd, Nanjing 210008, China; 3. College of Intelligence and Computing, Tianjin University, Tianjin 300354 China)

【Abstract】In recent years, Knowledge Graph(KG) and its related technologies have developed rapidly and have been widely used in various cognitive intelligence scenarios in industry. The key technology of knowledge graph in engineering application is introduced on the basis of the brief description of knowledge graph related research. This paper studies the typical application scenarios of various industry knowledge graphs, the corresponding case studies supported by well-known knowledge graph platforms and some relevant available tools in each phase of the life cycle, analysis requirements and key challenges when developing an enterprise-level knowledge graph platform, and proposes the construction method and process of enterprise-level knowledge graph platform. In view of the problems encountered in the platform construction, this paper gives the corresponding solutions of knowledge graph middle platform construction, and prospects the future development and challenges of knowledge graph.

【Key words】knowledge graph; representation learning; knowledge extraction; knowledge storage; knowledge reasoning; enterprise-level knowledge graph platform; knowledge graph middle platform

DOI:10.19678/j.issn.1000-3428.0057869.

0　引言

知识是机器实现认知智能不可或缺的基础，而知识图谱则是用于表示、处理与运用知识的关键技术，从而让机器能够理解知识并基于其进行相应的推理计算。知识图谱以其强大的语义表达能力、存储能力和推理能力，为互联网时代的数据知识化组织和智能应用提供了有效的解决方案。知识图谱的构建及其应用一方面引起了学术界的密切关注，大量研究者对知识图谱相关的技术进行了深入地研究，包括知识获取、知识融合、知识计算、语义搜索和知识问答等；另一方面，大规模知识图谱在解决实际问题时效果出色，也得到了工业界的纷纷采纳，以微软 (Microsoft)、谷歌、脸谱 (FaceBook)、eBay 和 IBM 为代表的国际巨头和以 BAT 和华为小米等国内大型互联网企业在其产品和产业应用中均使用了知识图谱及其相关关键技术[1]。

本文一方面将对知识图谱的相关技术研究进行

基金项目：国家自然科学基金(61972275)

作者简介：王昊奋（1982-），男，特聘研究员、博士、计算机学会（CCF）会员（会员号：41530M），主要研究方向知识图谱与对话问答；丁军、胡芳槐，博士；王鑫，教授、博士。**E-mail：**carter.whfcarter@gmail.com

系统的综述，同时将着重讨论知识图谱在企业级应用场景中的工程实践，包括典型的工业级知识图谱应用场景、知识图谱工程落地的生命周期、企业级知识图谱平台的构建以及中台化演进等。现有的综述文章大多偏向于阐述知识图谱相关的技术研究，包括知识图谱的总体研究综述[2-3]以及面向特定子领域的研究如知识表示学习[4-5]、知识融合[6-7]、知识存储[8-10]、知识推理[11-13]、知识补全[14]等；文献[3,15-16]也系统地介绍了在特定的领域场景中使用相关技术进行知识图谱构建。但上述工作没有涉及本文的重点即工程化流程总结与知识图谱平台的建设。

1 知识图谱概述

1.1 知识图谱的定义与分类

知识图谱最早于 2012 年由 Google 正式提出[17]，其初衷是为了改善搜索，提升用户搜索体验。知识图谱至今没有统一的定义，一种比较普遍被接受的一种定义为“知识图谱本质上是一种语义网络 (semantic network)；网络中的结点代表实体 (entity)或者概念 (concept)，边代表实体/概念之间的各种语义关系。”一种更为宽泛的定义为“使用图 (graph)作为媒介来组织与利用大规模不同类型的数据，并表达明确的通用或领域知识。”

从覆盖的领域来看，知识图谱可以分为通用知识图谱和行业知识图谱；前者面向开放领域，而后者则面向特定的行业。通用知识图谱强调的是广度，即更多的实体，通常难以形成完整的全局性的本体规范。行业知识图谱主要用于辅助各种复杂的分析应用及决策支持场景，它需要考虑领域中的典型业务场景及参与人员的背景和交互方式，因而需要完备性和严格且丰富的模式定义，并保证对应的实例知识具有丰富的维度，即一定的深度。行业知识图谱当前已经在金融证券、生物医疗、图书情报、电商、农业、政务、运营商和传媒等行业中得到了不少成功的应用。企业级的知识图谱应用通常是基于行业知识图谱提供智能服务，可以是面向一个行业，也可以是多个行业的结合；因而企业级知识图谱平台将围绕行业知识图谱的管理进行建设。

1.2 知识图谱研究进展

随着知识图谱在各行业的应用落地，知识图谱技术的相关研究得到了大量研究者的关注。文献[2]从知识表示学习，知识获取与知识补全，时态知识图谱和知识图谱应用等方面进行了全面的综述。在此基础上，本文分别从以下几方面来介绍知识图谱研究进展：1）知识表示学习；2）知识获取与补全；3）知识融合；4）知识存储与图计算；5）知识推理； 6）基于知识图谱的问答；最后还将阐述事件图谱与事理图谱等图谱发展热点。

1.2.1 知识表示学习

知识表示学习是面向知识图谱中实体（或概念）和关系的表示学习。通过将实体或关系投影到低维稠密向量（嵌入表示），实现对实体和关系的语义信息的表示，高效地计算实体、关系及其之间的复杂语义关联。

知识学习方法可以分为基于翻译距离模型(translational distance models)的方法和基于语义匹配模型 (semantic matching models)的方法；前者代表模型有高斯嵌入[18]、TransE 及其扩展[19-21]；语义匹配代表模型有 RESCAL[22]及其扩展模型 DistMult[23]、ComplEx[24]和神经网络匹配模型[25]。另一个相关的研究领域是网络嵌入（Network Embedding）[26-28]，它侧重于考虑如何充分利用节点在网络中的复杂结构信息；包括保留网络结构与属性的方法如 SDNE 算法[29]、保留边信息的 LANE 方法[30]和融合节点文本属性的方法[31]。随着深度学习的发展，基于神经网络的语义匹配模型和图神经网络成为知识图谱表示的研究热点[32]。

1.2.2知识获取与知识补全

知识获取与知识补全是知识图谱构建过程中最重要的基础环节；前者从数据中获取新知识，主要包括实体识别和关系发现；而后者是对现有知识图谱进行扩充。

早期的知识获取方法主要为基于语言学模式的方法，最近的研究主要聚焦于基于深度学习的方法[33-34]，尤其是最新的使用 Transformer 模型的大规模预训练模型（如 BERT）在实体识别等任务上取得了更佳的性能[35]。同时，远程监督学习[36-38]广泛应用于语料难以获取的场景。

知识图谱补全[14,39]通过相应的推理和补全算法扩展现有的知识图谱，包括基于嵌入的排序补全算法、关系路径推理算法、基于深度强化学习的算法和基于规则的推理算法等。

1.2.3 知识融合

知识融合是指将多种来源的碎片化数据中获取的结构各异、语义多样和动态演化的知识，通过冲突检测和一致性检查，对知识进行正确性判断。知识融合按融合阶段分类包括知识评估和知识扩充[6]；而从人机协作角度来看，知识融合分为基于知识库的知识融合[40-42]、基于人工的知识融合以及基于知识库与人工协作相结合的知识融合[43]。

1.2.4 知识存储与图分析计算

大规模知识图谱的存储以三元组存储为核心，同时还包括其它类型知识的存储。三元组知识的存储主要有

RDF（Resource Description Framework，资源描述框架）存储和图数据库两种类型，前者以 RDF 图模型为基础，后者大多数采用属性图数据模型。由于图数据库成为当前使用的主流，本文更关注图数据库相关的工作，以及在存储上的图分析计算。

知识存储与图分析计算相关研究主要侧重于 RDF 图谱数据管理[8][9]、图数据查询[44]、图谱计算框架[45][46]等方面。文献[10]从知识图谱数据模型、知识图谱查询语言、知识图谱存储管理和知识图谱查询等四个方面对知识图谱数据管理相关研究进行了综述。在图计算框架方面，文献[45]进行了全面的综述。

1.2.5知识推理

推理是指基于已知的事实或知识推断得出未知的隐藏事实或知识的过程。面向知识图谱的知识推理[11]通常可以分为基于规则的推理[47-48]、基于知识表示学习的推理[5]，基于神经网络的推理[49-50]和混合推理[51]。基于规则的推理方法拥有较高的准确率，但难以扩展和平移；基于神经网络的推理具备更好的推理能力、学习能力和泛化能力，但神经网络结果不可预测和解析。因此，研究者提出混合推理以结合不同推理方法之间的优势[51]。基于神经-符号整合的推理[13]将符号系统的透明性和推理能力与人工神经网络的健壮性和学习能力结合在一起。

1.2.6基于知识图谱的问答

基于知识图谱的知识问答 (Knowledge Based Question Answer, KBQA)给定自然语言问题，通过对问题进行语义理解和解析，进而利用知识库进行查询、推理得出答案。

KBQA 主要方法有：基于语义解析的方法[52]、基于信息抽取的方法[53]和基于向量建模的方法[54]。随着深度学习的发展，知识表示学习和语义解析得益于神经网络的非线性表达能力对语义进行更好的建模，基于知识表示学习的 KBQA 和语义解析结合深度学习成为 KBQA 的主流方向。然而，多样化用户意图理解和语义的歧义性仍然是 KBQA 的主要挑战[53]。

1.2.7 事件图谱与事理图谱

事件知识图谱对于事件的建模具有明显的语义表达优势，有利于事件链知识推理。事件知识图谱相关的研究主要聚焦在事件抽取[55]、事件推理和事理图谱。

事件抽取的任务包括触发词检测、触发词事件分类、事件元素识别和事件元素角色识别。事件推理的相关工作主要包括事件因果关系推理、脚本事件推理、常识级别事件产生的意图和反应推理和周期性事件时间推理等。是一个事理逻辑知识库，描述事件之间的演化规律和模式，从结构上看它是一个有向有环图，节点代表事件，边代表事件之间的关系（顺承、因果等）。

1.3 知识图谱工程化

知识图谱的应用需要综合利用多方面的技术：知识图谱的构建涉及知识建模、实体识别、关系抽取、关系推理、实体融合等技术，而知识图谱的应用则涉及到语义搜索、智能问答、语言理解、决策分析等多个领域。总体而言，构建并应用知识图谱需要系统性的利用好包括知识表示、数据库、自然语言处理、机器学习等多个方面的技术。

规模化的知识图谱工程落地需要有完整的工程化流程作为指导。通常场景下其流程包括：首先确定知识表示模型，进行知识建模；然后进行数据收集，根据数据来源选择不同的知识获取方法，并对不同来源、不同方法获取的知识进行融合；接下来需要综合利用知识推理、知识挖掘等技术对所构建的知识图谱进行质量评估与补全，最后根据场景需求设计不同的知识应用场景，如语义搜索、问答交互、图谱可视化分析等。在经过大量的知识图谱研究与产业化落地实践后，逐步形成了行业知识图谱应用落地的全流程，称为行业知识图谱的全生命周期，包括知识建模、知识获取、知识融合、知识存储、知识计算与知识应用。

1.4 行业知识图谱生命周期

知识建模的主要目标是为知识图谱定义本体，其主要挑战有：1）多类型数据的知识表示，2）是否能够自动或者半自动的生成模式层知识。知识建模通常采用两种方式：一种是自顶向下 (Top-Down)的、专家定义的方法；另一种则是自底向上(Bottom-Up)的、数据驱动的规约方法，从数据中通过自动映射、归纳等方法生成模式知识。

知识获取是指从不同来源、不同类型的数据中进行知识提取并存入知识图谱的过程。其主要挑战包括：1）如何从多源异构的数据中抽取知识；2）以及如何自动或半自动地从非结构化的数据中抽取；3）所获取知识的准确率；4）样本数据稀疏。

知识融合的目标是对从多源异构的数据中获取到的知识进行融合从而形成统一的、一致的知识放入到知识图谱中，通常分为模式层的知识融合和实例层的知识融合。

知识存储的目标是实现各类知识的存储，包括

基本实体知识、属性知识、关系知识、事件知识、时序知识和业务规则知识等。其主要挑战在于两个方面，一是实现对多种类型知识的存储，二是实现大规模知识图谱数据查询、推理、计算等的高速存取。

知识计算主要包括图挖掘计算和知识推理。图挖掘计算主要是指基于图论的相关算法实现对知识图谱数据的探索、挖掘与嵌入，其主要挑战在于大规模图算法的效率。知识推理的关键挑战包括：大数据量下的快速推理，以及对于增量知识和规则的快速加载。

知识图谱发展到当前，其应用场景非常多，最典型的应用为语义搜索、智能问答和可视化决策支持。对于语义搜索和智能问答，主要难点在于对用户的输入进行准确地意图理解；而对于可视化决策支持而言，一方面需要提供良好的用户交互方式实现用于与数据及算法的接口，同时还需要下层服务的有效性以及快速响应。

总体而言，行业知识图谱落地是一个系统性的工程问题，需要有上述生命周期的完整理论支撑；同时，还需要有相应的技术、算法、工具来支撑应用的落地；因而在工业级的应用场景中，通常会围绕生命周期构建相应的行业知识图谱平台，然后在平台的基础上进行应用的构建。

2 知识图谱应用与相关工具

本节先描述知识图谱相关的工业级应用；然后介绍知识图谱相关的系统平台，以及生命周期各环节应用的工具；本文所讨论的平台或工具以开源的产品为主，同时也包含一些在领域中具有较大影响力商业产品。

2.1 工业级知识图谱应用

以搜索为主要应用场景的有谷歌知识图谱、微软必应知识图谱、百度知识图谱和搜狗知识图谱等。

谷歌知识图谱是于2012年最早提出来的用于改善搜索的知识图谱，用户进行实体有关的查询时会发现结果中还包括了知识图谱提供的事实。目前它涵盖了广泛的主题，包括超过10亿个实体和700亿条事实。

微软必应知识图谱包含物理世界的知识，如人物、地点、事物、组织、位置等类型的实体，以及用户可能采取的行为。覆盖范围、正确性和时效性是该图谱质量和实用性的关键因素。

脸书拥有全球最大的社交图谱，该图谱以用户为中心，同时包括用户关心的其它信息如兴趣爱好、从事行业等信息。脸书的图谱主要用于增加用户对脸书产品的体验，包括内容搜索和兴趣推荐等。

阿里和易趣拥有大规模的商品知识图谱服务于他们的电商平台，他们实现了基于大规模知识图谱的快速搜索与推荐，从而提升了用户体验并提高了商品销售量。

2.2 工业级知识图谱平台

在工业级知识图谱应用快速增长的带动下，一些工业级的知识图谱平台也相应被推出。

2.2.1 Palantir

Palantir是用于知识图谱创建，管理，搜索，发现，挖掘，积累的可扩展的大数据分析平台。通过结合动态本体论思想和自身数据整合能力，形成以知识图谱为基础的知识管理体系，通过图挖掘、本体推理等算法引擎赋能知识图谱，为搜索和知识发现提供数据支撑，同时支持协同工作分析，而且整个分析过程通过可视化、交互式的方式进行。Palantir目前拥有两大产品线：Palantir Gotham和Palantir Metropolis，分别应用于国防安全与金融领域，形成了包括反欺诈、网络安全、国防安全、危机应对、保险分析、疾病控制、智能化决策等解决方案。Palantir通过整理、分析、利用不同来源的结构化和非结构化数据，创造一种人脑决策和计算机智能共生的大数据分析环境及工具系统，通过可视化技术形成“人机共生”的可视化大数据交互探索分析能力，促进人脑和大数据分析互补，提升客户的决策洞察力。

2.2.2 IBM Watson Discovery 知识图谱框架

IBM开发了Watson Discovery服务及其相关产品所使用的知识图谱框架，在外部许多行业中也进行了部署应用。IBM Watson知识图谱框架有两种典型的应用场景：一是直接使用结构化以及非结构化的数据来发现新的知识为下游产品提供服务；二是该框架允许用户以预先构建的知识图谱为基础来构建自己的知识图谱。

该知识图谱框架的特性有：1）使用了多态存储，支持多种索引、数据库结构、内存数据库和图存储，将数据分布到多个存储库中，每个存储库满足特定的应用需求和工作负载；2）保留原始“证据”， 这些元数据和其他相关信息通常在后续的知识应用非常重要；3）推迟实体消歧，因为在创建过程中消歧通常会损失实体的原有信息，这和知识发现的目标相冲突。

2.2.3 Oracle 知识图谱平台

Oracle 知识图谱平台基于其自身多年的存储经验，在具有明显优势的存储层上进行构建，上层通过 W3C 标准的 RDF 和 OWL 来组织和表示图谱，使用 SPARQL 来对数据统一查询服务。平台支持两种图的表示方式：属性图 (Property Graph)和 RDF 三元组；前者适合各种图计算如最短路径、权重排序和中心性 (Betweeness)等；后者适合进行知识的推理。Oracle 知识图谱平台的主要特性有：对数据存储与访问的支持性比较好，可以实现基于内存的并行图计算；提供了许多工具完成从各种大数据平台、关系数据库到知识图谱的映射与转换。

2.2.4 Metaphactory

Metaphactory 提供了一套从知识存储、知识管理到知识查询与应用开发的端到端的知识图谱平台解决方案；知识图谱存储可以兼容使用常见的三元组存储，如 Blazegraph、Stardog、Amazon Neptune、GraphDB 和 Virtuoso 等；数据交互使用了标准的 SPARQL 作为交互协议，从而规避了存储使用不同数据库带来的影响，实现不同数据源、不同格式的知识场景进行混合查询；同时提供了搜索、可视化和知识编辑管理的 UI 接口，并为 Tabular 等 BI 工具提供了数据接口。但 Metaphactory 主要还是针对结构化数据进行查询和管理，并没有提供对非结构化数据处理的能力。

2.2.5 Stardog

Stardog 是一个企业级知识图谱平台，通过把数据转换成知识，使用知识图谱进行组织，对外提供查询、检索和分析等服务。主要特点：1）把关系数据库映射成虚拟图；2）支持 OWL2 的推理；3）支持 Gremlin。但 Stardog 仅包含对结构化数据（RDBMS、Excel 等）的处理，没有针对非结构化数据的知识抽取，也没有包含知识融合功能。

2.2.6 其它知识图谱平台

上述这些平台都是商业的平台，通常提供试用的版本可以供非商业用途学习和研究；而开源知识图谱项目的典型代表为 LOD2。LOD2 的主要目标是构建结构化链接数据的企业级管理工具和方法，提供一个搜索、浏览和生成链接数据的平台。它侧重于链接数据的生命周期管理，其它类型的数据需要首先转换成链接数据。

在中国，以百度（百度 AI 开放平台）、腾讯（腾讯知识图谱，Tencent Knowledge Graph, TKG)）、阿里巴巴（藏经阁）、华为（华为知识图谱云）等为代表的国内互联网公司也在积极构建知识图谱，并且针对垂直领域构建知识图谱平台，促进知识图谱的发展和工业落地。

2.3 知识图谱生命周期相关工具

除了前述所提到的知识图谱平台以外，还有许多与知识图谱生命周期中特定环节相关的工具。这些工具通常不像完整的平台一样完成一站式的服务，但是它们也为知识图谱的应用构建提供了便利，可以在构建完整的企业级知识图谱平台时进行集成使用。本节接下来介绍生命周期各环节的相关工具；知识计算将分为知识推理和图挖掘分析两部分介绍。

2.3.1 知识建模工具

Protégé是一个本体编辑器，基于 RDF(S)，OWL 等语义网规范，提供 PC 图形化界面和在线 Web 版本-- WebProtégé，通常适用于原型构建场景。NeOn Toolkit 是一个适用于本体工程生命周期的工具，它以 Eclipse 插件的方式为用户提供服务。

这些本体编辑工具不足点包括：基本只提供单人编辑，而协同编辑时需要通过文件共享来实现；对大数据量支持不佳；不支持复杂事件及时态的建模；基本依赖手工编辑，难以实现与知识图谱（半）自动化构建过程的交互。

2.3.2 知识获取工具

知识获取包括从结构化数据、半结构化数据和非结构化数据中获取知识。

从结构化数据中获取知识的目标通常是把关系数据库中的数据转换成 RDF 形式的知识，W3C 为此制定了从关系数据库映射到RDF数据集的标准语言 R2RML。典型的开源工具有 D2R MAP 和 D2RQ[56]。D2RQ 是一个将关系数据库转换为虚拟的 RDF 数据库的平台，主要包含 D2R Server[57]、D2RQ Engine 和 D2RQ Mapping Language 三个组件。这些工具把数据直接转换成 RDF，难以与知识建模结果结合与映射，也难以同其它类型的知识进行融合；对于大规模海量数据映射以及新数据的增量映射支持困难。

从半结构化数据中获取知识通常是指使用包装器的方法从网页数据中获取知识，代表性工具有：Lixtio[58]提供了一种用户可视化配置的方式进行半自动化生成网页包装器的工具；WIE 是一个通过网页自动分析从而辅助生成包装器的工具，适用于抽取目标数据中的表格信息。这些工具基本是针对早期的静态 HTML 页面开发的，已经难以适用于当前前端动态页面技术，需要在它们的基础上进行动态页面支持扩展。

DeepDive 与 Snorkel 提供了一套面向特定关系的、基于远程监督学习的抽取框架；使用现有知识库和规则定义来自动生成语料，框架自动完成模型的训练过程，并使用机器学习算法来减少各种形式的噪音和不确定性；用户可以使用简单的规则来影响（反馈）学习过程以提升结果的质量。DeepKE 是浙江大学基于深度学习方法的开源中文关系抽取工具，使用了包括卷积神经网络、循环神经网络、注意力机制网络、图卷积神经网络、胶囊神经网络以及使用语言预训练模型等在内的多种深度学习算法；该工具同样仅用于关系的抽取。上述工具主要针对关系的抽取，未提供针对概念、实体、事件等的抽取功能。

2.3.3 知识融合工具

知识融合的目标是对于来源不同、抽取方法不同、结构不同的知识进行合并形成统一的知识。DBpedia MappingTool 是一个用于把从 Wikipedia 中抽取的信息通过映射融入到 DBpedia 中的工具，以可视化的方式让用户进行 DBpedia 中本体（类、实体、数据类型等）和信息模块的映射。Knowledge Vault[59]是谷歌推出的一个互联网规模的知识库，它融合了海量的从互联网中基于先验知识库抽取的信息，并通过监督学习的方法来对这些知识进行融合。这些融合工具通常是针对特定的场景设计的，通用性和可配置程度通常比较低，难以实现复杂多变场景下的知识整合。

2.3.4 知识图谱存储工具

知识图谱中最主要的数据结构为基于图的结构，图结构数据的存储主要有两种方式:RDF 存储 和图数据库 (Graph Database)。在工业级的场景下，一般从如下几个维度衡量知识图谱存储的性能：支持的数据规模、是否支持数据分布存储、知识建模管理能力、查询语言表达丰富性、ACID 支持以及是否有开源产品等。常见图数据库的对比如表 1 所示。

表 1 常见图数据库对比 (Comparison of graph databases)

	Neo4J	JanusGraph (Titan)	TigerGraph	PlantGraph
支持数据规模	亿级	百亿级+	千亿级+	千亿级+
数据分布存储	否	是	是	是
知识建模管理能力	无	无	无	有
查询语言	Cypher	Gremlin	GSQL	Cypher Gremlin
ACID 支持	是	否	是	是
是否开源	否，有社区版	是	否	否

Neo4J 是第一代图数据库的最知名代表，它使用了原生图存储结构；它不使用 schema（即 schema free），是一种自由的图数据管理方式；同时它还支持 ACID 事务的处理，并提供了 Cypher 查询语言。Neo4J 在企业级数据管理中主要碰到的问题有：1）不使用 schema 会难以从整体组织和理解图谱数据；2）并未实现真正意义上的数据分布式存储，因此在大规模的数据场景下会遇到性能瓶颈。

JanusGraph 是在 Titan 的基础上发展起来的第二代图数据库的代表。它设计的原理是在现有的成熟存储（如 NoSQL）上实现对图的存储逻辑；底层存储的分布式能力使其天然具备分布式能力。但此类数据库最大的问题是会遇到图连接查询的性能瓶颈，尤其是在大规模图数据的多步查询的场景下；另一方面，这种架构也不能有效地支持离线分析，需要使用外部的分析引擎，但这种结合难以做到数据快速加载与更新。

在数据量大规模增长与实时查询分析要求不断提升的背景下，基于原生、并行图设计的图数据库逐渐成为发展方向，也被称为第三代图数据库。其中的代表产品为商业数据库 TigerGraph 与 PlantGraph，它能够有效地支持 OLTP 和 OLAP 等多种应用场景，能够解决大规模图数据场景下的多步连接问题。目前第三代图数据库还只在一些的拥有大数据量与高性能要求的商业场景下得到使用，尚未有开源的产品出现。

2.3.5 知识推理工具

知识推理主要分为基于逻辑的推理与基于统计的推理，逻辑推理又主要包括本体推理和规则推理。

RDFox[60]是一个本体知识推理工具，支持共享内存并行 OWL 2 RL 推理。RDFox 支持 Java、Python 多语言 APIs 访问，并且 RDFox 还支持一种简单的脚本语言与系统的命令行交互。RDFox 完全基于内存，对硬件的要求较高，大超大规模的数据场景下会难以使用。 Drools 是一个使用 Java 语言开发的基于 RETE 算法（一种前向推理算法）的业务规则推理引擎；它使用“If--Then”形态的句式和事实的定义，使引擎的使用非常直观；它还支持将 Java 代码直接嵌入到规则文件中。Link Prediction Tool 是一个在大规模网络中自动发现缺失的链接的工具，主要用于社交网络中的链接预测。SNAP (Stanford

Network Analysis Platform)是斯坦福大学研发的一个通用高性能大规模网络分析与操作平台，能够高效地实现大规模网络中的链接预测。

2.3.6 图挖掘分析工具

前述提到的大多数图数据均只支持 OLTP 模式的图查询功能以及一些简单的图算法，对于大规模的图挖掘分析支持较少。因此，基于图数据库实现图挖掘分析的模式是集成第三方的图挖掘分析工具如 Spark GraphX、GraphLab 和 Giraph 等。最常用的是 Spark GraphX，它是在实时计算引擎 Spark 上为图计算设计与实现的一套计算框架，方便用户通过统一的模式进行图算法编程；不过由于它是基于通用的计算框架来实现图计算，性能较图分析的专用系统要低。

Plato 是腾讯开源的一个支持十亿级别节点的超大规模图计算框架；基于其自适应图计算引擎，它能够根据不同类型的图算法，提供自适应计算模式、共享内存计算模式和流水线计算模式等多种计算模式。但它是一个重量级的图计算框架，其集成成本相对较高；开发者需要基于其独特的底层 API 编程，因此定制化开发成本较高。

Euler 是阿里开源的大规模分布式图表示学习框架，内置 DeepWalk、Node2Vec 等业界常见的图嵌入算法。

2.3.7 语义搜索与智能问答工具

知识链接是支持语义搜索的重要方法，知识的实体链接工具有 Wikipedia Miner 和 DBPedia Spotlight。这些早期的工具通常是以开放的知识图谱（Wikipedia、DBpedia 等）为知识链接的目标知识库，使用字符串匹配、向量相似度等算法进行计算；当前，基于深度学习、知识图谱表示学习的方法已经成为知识链接的最新发展方向。

智能问答方向知名的开源工具有 ActiveQA 和 gAnswer 等。ActiveQA 是谷歌开源的一款使用强化学习来训练 AI 智能体进行问答的研究项目，在强化学习框架的推动下，智能体逐步学会提出更具针对性的具体问题并理解、问答问题，从而得到所寻求的结果。gAnswer 是一个基于知识图谱的自然语言问答系统，能够将自然语言问题转化成包含语义信息的查询图，并将查询图转化成标准的 SPARQL 查询，将这些查询在图数据库中执行，最终得到用户的答案。

这些问答工具适用于特定的场景（如 gAnswer 用于 KBQA），而在复杂企业级的场景中通常需要支持上述所有类型的问答任务。

3 企业级知识图谱平台

本节首先介绍企业级知识图谱平台的构建需求与挑战，然后以金融行业知识图谱的构建与应用为例详细描述该知识图谱平台完整的构建过程。

3.1 企业级知识图谱平台构建需求

从确定待采集的原始数据到最终的应用开发，企业级的知识图谱应用落地需要对数据背后的知识进行建模、抽取、融合、校验、补全、分析计算等一系列加工处理；这些过程的每一步都需要专业的图谱知识和技能才能完成。如果没有平台或者工具进行支撑，图谱的应用构建将是一项门槛极高甚至无法完成的工作。因此，企业级图谱的应用普及亟需一个功能强大的知识图谱平台。该平台需要覆盖行业知识图谱生命周期的所有环节；同时须满足企业级应用的各种功能性与非功能性需求：

1）知识建模：除基本的本体编辑功能外，还必须具备表示多类型知识的能力，尤其是对动态事件知识、多媒体数据和业务过程数据等的知识表示；同时，企业知识图谱的建模通常需要支持多人在线协同工作以及知识的多版本管理；最后，需要集成如下文 2）中所述的各种知识抽取能力，其旨在从数据中自动发现知识，避免纯手工构建大规模图谱带来的工作量大、效率低下并易出错等问题。

2）知识获取：需要提供分别从结构化数据、半结构化数据和非结构化数据中获取知识的工具；以本体数据模式 (data schema)为基础支持大规模、增量数据的 D2R 映射，实现动态网页的包装器配置与归纳学习，提供从文本中抽取实体、关系、属性和事件等多维度知识的方法；同时，需要降低从非结构化数据中获取知识的成本（数据标注规模和标注代价），提供弱监督或自监督学习的能力；最后，需要保障所获取知识的质量，尤其是从非结构化数据抽取知识的难度最大。

3）知识融合：提供用户基于业务配置融合规则与自动算法相结合的知识融合功能；提供本体映射、实体对齐和属性融合等能力；自动进行冲突检测并能够依据（预先设定的）策略进行解决。

4）知识存储：首先需要实现多类型知识的存储；其次需要支持大规模图谱存储及其之上的高效查询，在企业级的应用场景中，图谱通常包含百亿甚至千亿级别的知识（以三元组形式表示）；具备复杂知识模式管理的功能，用于支持知识建模工具的高效交互；提供 SPARQL、Cypher、Gremlin 和 GQL

等多种常见图查询语言。

5）知识计算：需要具有大规模知识图谱推理与图挖掘的能力。具体而言，能够高效地加载大规模图谱数据并进行推理计算；支持多种图挖掘算法并能实现并行挖掘分析；考虑图谱的演化或新知识的持续加入并实现高效的增量计算与推理。

6）知识应用：提供多种知识可视化视图及交互方式并与后台的存储、计算能力相结合，为用户提供快速的知识应用服务；基于知识图谱提供语义搜索能力；并提供能够支持诸如问答对检索、交互式分析和阅读理解等多种场景的综合问答能力。

3.2 企业级知识图谱平台构建面临的困难

1) 多类型知识的表示、获取与存储。首先面临的问题是如何实现企业级应用场景中多类型数据的统一知识表示，数据类型的复杂性和多样性使得传统的三元组表示方法难以胜任；其次，如何从这些数据中高效获取知识是另一个难点，需要采用不同的方法甚至是多方法的集成来实现大规模知识的获取；最后，如何统一存储这些知识从而能够同时支持上层各种任务与服务也非常困难。

2) 大规模知识图谱的性能。企业级知识图谱的规模通常在百亿、千亿甚至更高的级别。如何实现大规模知识的可扩展存储、并支持其上的高效查询、并行计算与推理服务面临着巨大的挑战。

3) 图谱数据的统一消费利用。如何无缝集成可视化、语义搜索和问答分析等多种交互方式，在不增加用户额外学习成本和使用门槛的情况下提供统一的知识图谱消费体验是一项综合人工智能和人机交互等多学科知识的技术难题。

另外，知识的演化与时效性也是一个难以回避的难题。随着外部世界的变化和企业业务的变迁与升级，业务数据及相应的知识也不断扩展与变更，支持知识图谱中知识的时态表示，及时检测知识的时效性，并根据图谱的演化支持自适应知识推理与计算同样是挑战。

3.3 企业级知识图谱平台构建

构建知识图谱平台有三种可能的方式：1）在现在的开源知识图谱平台上进行扩展；2）把行业知识图谱生命周期中每个环节对应的工具进行集成形成完整的平台；3）从零开始构建。整体而言，第一种方法通常难以执行，因为这些开源的知识图谱平台从设计、可扩展性等方面均难以进行深度二次开发；而从零开始构建则成本过高；因此，最佳实践方法应对行业知识图谱生命周期对应的工具进行综合利用，在此基础上进行满足上述需求的全流程全局设计，并且对缺乏工具的环节进行针对性开发，对需要改进的工具进行完善，从而整合形成完整实用的企业级知识图谱平台。

3.3.1 知识建模

企业级的知识建模工具首先需要有多类型知识表示的能力，实现概念、实体、属性、关系、事件、业务规则以及多媒体数据对应的语义内容的统一表示。最佳的实现方法是把 W3C 推荐的标准知识表示模型（RDF 和 OWL）与其它的知识表示框架相结合，这些框架包括产生式规则和文件对象等。RDF 和 OWL 能够良好地以三元组的形式表示概念、实体、属性和关系等知识；事件可以视作一个特殊的概念，例如可以把“投资事件”定义成一个概念，并给它定义属性（金额、时间等）和关系（投资方、融资方）；业务规则的一种有效表示方法为产生式规则，例如“IF 企业.估值 > 1 亿美元 THEN 企业是准独角兽”；使用文件对象来表示多媒体形态的数据（如视频、图片或文档等），然后使用链接标引的技术手段使其与领域图谱中的相关知识进行关联，形成多模态知识图谱。

其次，为实现协同知识编辑，企业级建模工具以在线 Web 的形式实现多用户登录与权限管理、并发控制、编辑过程主动提示与自动补全等功能；并依托平台存储能力使得面向大规模知识图谱的可扩展建模成为可能。

最后，平台通过以下方法实现半自动化建模能力：1）基于 E-R 图模式解析的方法实现从结构化数据中自动发现模式；2）基于“统计+规则”的方法从现有知识中自动规约概念与属性的算法，在发现过程中通常需要进行人工干预、确认，通过人机交互的方式得到最终的图谱模式层知识。

3.3.2 知识获取

平台需要包括对不同类型数据进行知识获取的工具。具体而言，涵盖面向结构化数据的 D2R 工具，面向半结构化数据的包装器配置与生成工具，以及面向非结构化数据的自动抽取工具；同时需要额外支持对事件等复杂类型知识的抽取。

D2R 映射工具的一种可行实现方法为以 R2RML 映射语言为基础，开发在线 Web 形式的所见即所得的交互式配置交互页面，并把源数据与知识图谱的模式（定义的概念与属性）进行映射；同时还需要提供设置融合合并的规则配置、以及增量数据的判断依据（例如更新时间）等。

包装器的配置同样需要提供所见即所得的配置方式或配置文件的配置方式，提供基于源码字符串、正则表达式、XPath 等进行知识元素的位置确定方式。基本的步骤如下：1）获取源码，通过集成 selenium 等引擎实现动态页面加载成 HTML；2）预处理，去除相关的噪声如 CSS、JS 代码等；3）字段配置，基于定义的模式层知识配置每个字段解析数据的方法，包括前置规则、后置规则、正则表达式等；4）后处理，进行结果的过滤与转换。同时，平台依据第一节中描述的模板学习方法实现相应的包装器自动学习算法，用户可基于学习的模板进行配置，减少人工工作量。

对于非结构化数据的抽取，最佳实践方法是首先集成现有开源的工具，如前文所述的 Snorkel、DeepKE 等。其次，提供基于规则的抽取方法，其实现的过程与包装器配置基本相同；基于规则的方法可以快速获得准确率较高的知识，一方面作为抽取结果，另一方面可作为机器学习模型训练的语料。然后，对于需要定制训练抽取模型的数据，提供第三方模型集成的能力以及在线训练模型的平台，集成第三方模型通过微服务的注册来实现；在线训练平台的后端通常通过集成现有的深度学习框架如 TensorFlow、PyTorch 等实现，用户在线标注或上传指定格式的语料后，后端启动模型的训练。

对于企业级的复杂数据，为保证抽取知识的质量同时降低对人工标注语料的依赖，可以使用如下的多策略最佳实践方法：利用不同数据源之间的信息冗余，使用较易抽取的知识（结构化数据库中的）来辅助抽取困难的信息（文本信息抽取）。整体架构如图 1 所示；其中展示的围绕企业信息的抽取，首先优先从工商企业库中通过 D2R 配置的方式抽取得到准确率高的企业基本知识，然后从专利网站中通过包装器配置实现专利数据解析形成企业的专利信息，最后基于这些已经抽取的知识以及通过规则的方法从文本中得到的知识，自动生成文本信息抽取模型训练所需的语料，实现远程监督学习。

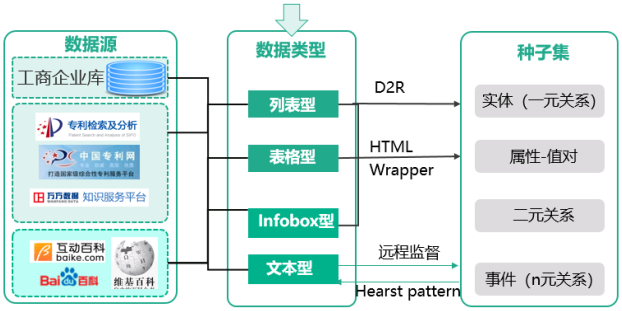

图 1 多策略信息抽取方法 (Multi-strategy information extraction method)

3.3.3 知识融合

对于模式层的知识融合，通常采用人工融合的方法，因此平台需要提供交互配置界面进行融合编辑。对于实例层的融合，首先同样提供人工融合编辑的功能，用于对（半）自动融合算法结果进行修正；其次，需要提供给用户配置界面实现基于业务规则的融合，规则通常包括基于名称、属性、置信度等的相似程度的方法；第三，提供自动的融合算法，算法的依据通常为待融合知识的相似度，基本计算方法包括基于字符串匹配、基于向量空间模型、语义距离计算和图嵌入向量相似度等。

平台还需要提供冲突检测与自动解决功能。冲突检测可基于知识推理工具来实现；而冲突自动解决通常依据置信度来实现（通常选取置信度高的）。

3.3.4 知识存储

企业级知识图谱中的多类型数据和应用的多样性决定了知识图谱的存储必然是一种混合存储的模型。一种最佳实践的方法则以存储三元组数据的图数据库为核心，使用关系数据库、NoSQL、文件存储等方式存放记录型、文档型、文件等数据从而实现对多场景应用交互的需求。

表 2 公开的图数据库性能评测报告 (A public evaluation report of graph database performance)

	Redis Graph	Tiger Graph	Neo4J	Janus Graph
1 步	0.8	24.1	205	395
2 步	503	460	18340	27400
3 步	9301	6730	298000	4324000
6 步	78730	63000	N/A	N/A

在存储性能方面，传统的图数据库（如 Neo4J 和 JanusGraph 等）通常难以实现对大规模（百亿到万亿级别）知识的高效存取和查询，一个公开的性能评测报告如表 2 所示[1]，其中第一列表示从选定节点出发进行广度遍历的步数，时间单位为毫秒，数据集为公开的 14.68 亿关系的 twitter 数据集，N/A 代表测试超时；表中说明在十亿级别的三元组中，Neo4J 与 JanusGraph 已然无法满足深度查询的要求。原生并行图是当前实现大规模知识图谱数据实时存取的最佳解决方案，其基本的思想是使用原生的图存储结构，数据存放在文件系统或计算机主存中，

[1] https://redislabs.com/blog/new-redisgraph-1-0-achieves-600x-faster-performance-graph-databases/

同时通过图分割实现数据的分布式存储并提供图分割场景下的相关图算法实现。这种工业级的图数据库的实现复杂度通常非常高，因此企业级的应用场景中也可以考虑部署商用的图数据库（如TigerGraph、PlantGraph 等）；若自行研发实现，则需要从底层的原生图存储开始设计，然后实现数据的分割存储以及分布式并行计算，这通常需要投入大量的研发成本。

3.3.5 知识计算

企业级知识图谱平台中需要包括图挖掘计算、知识推理等功能。

图挖掘计算方面，首先实现常见的图算法，包括图遍历、路径发现、关联分析、社区发现、连通子图等；通常的方法是基于一些开源的工具实现，如 python-graph、JGraphT 等。其次，实现图挖掘分析引擎，代价较低的方法是集成现有的开源分布式图分析框架(如 2.3.6 节提到的 Spark GraphX、Giraph 等），这种方法适用于对实时性要求不是特别高的场景；而在实时性要求较高的场景中，则需要基于原生并行图存储之上单独开发相应的图分析引擎，需要考虑分布式协同计算、图分割等复杂技术实现。

知识推理主要实现方法为集成现有的成熟工具 RDFox 实现本体知识推理，以及集成 Drools 并进行一定的扩展实现业务规则推理。当数据规模超过这些工具能够承载的能力时，一方面可以提供相应的筛选方法从而只对关心的知识进行推理，另一种方法则是基于这些工具进行扩展从而实现分布式推理的能力。

3.3.6 知识应用

企业级知识图谱平台中需要提供知识可视化、语义检索、智能问答等算法和基础工具的支持。知识可视化通常采用基于现有的开源工具（如 D3.js、ECharts 等）进行扩展开发，提供多种可视化视图如星形图、树状图、点阵图等，以及钻取、放大缩小等交互方式。

语义检索主要解决传统的关键词检索中遇到的两个难题：即自然语言表达的多样性和自然语言的歧义性。这两个问题可以通过使用基于知识图谱的实体链接和意图理解有效地进行解决。同时，语义检索还为用户展现类似于实体搜索所提供的丰富的知识切面，让用户更便捷地获取和理解结果。

企业级的智能问答需要支持 IRQA（information retrieval question answering，基于信息检索的问答）、KBQA 和 MRCQA（machine reading comprehension question answering，基于机器阅读理解问答）等多种问答模式。不同的问答技术擅长回答的问题场景不同，单一地采用一种范式具有一定的局限性，需要将三种问答技术进行融合，构建多策略问答引擎，最大限度地覆盖用户问题，更好地满足企业应用的需求。多策略问答实现的基本过程如下：首先，根据问题与资源的不同，多策略问答引擎会根据语义理解的结果在 IRQA、KBQA、MRCQA 中选择一种或多种并行执行，对于单一问答技术，也会使用多种实现策略并行执行来完成候选答案的生成，同时为每一组候选答案收集相应的证据并进行置信度打分；把收集到的证据与置信度作为特征送入到下一阶段，在此阶段中，会根据上一阶段的结果对候选答案集进行重新排序，最终选择得分最高的答案生成最终回答。

3.4 产业化应用

工业级的知识图谱应用在金融证券、军工情报、图情分析、生物医疗、电商、农业等行业均得到了有效地利用。以金融证券领域为例，知识图谱在金融情报分析检索、反欺诈分析、金融智能化等场景已经有诸多成功应用落地。

以前述的金融创投场景为例，基于知识图谱平台的知识图谱应用过程如下：1）首先进行应用场景分析，依据分析结果进行知识图谱的建模，有公司、人物等概念，人物的属性、投资关系等；2）选择数据源，包括企业的基本信息、工商数据、专利数据、网络上的公开新闻数据等，对数据进行接入并预处理，利用平台的知识获取工具进行知识的抽取；3）依据平台的自动融合功能以及基于业务规则的融合配置，实现各种知识的融合并存储到平台中，例如人物可通过配置身份证号相同进行合并；4）应用开发，依据应用的场景进行算法选择或定制开发、模型训练和业务系统定制化二次开发。

4 企业级知识图谱中台建设

基于知识图谱平台的应用落地范式虽然流程清晰，但是仍然会碰到一些问题：

- 从知识图谱的建设到应用周期过长；
- 图谱构建过程难度较高，需要专业技能；
- 跨项目、跨领域迁移成本高；
- 数据、知识、模型、算法等可复用程度低；
- 应用构建复杂，需要技术人员深度开发。

对于上述问题，引入当前热门的中台相关技术可以有效地解决。中台是指在一些系统被共用的中间件的集合，通过使用中台可以抽象出可复用的各

种能力（数据、知识、模型、算法、功能模块等），以达到缩短应用构建周期、快速响应业务需求的目的，同时降低跨领域的迁移成本。

企业级知识图谱中台是在知识图谱平台的基础上引入中台相关的理念和技术，对平台进行重构升级的结果。形成的知识图谱中台整体架构如图 2 所示,包括数据接入层、知识图谱平台层 (KGBox)、中台层（自下而上依次为组件微服务化、预构建与应用编排）和应用层。总体而言是在原有知识图谱平台（简化成 KGBox）的基础上进行上述三个过程的升级重构，从而更加灵活地支撑上层的应用场景。

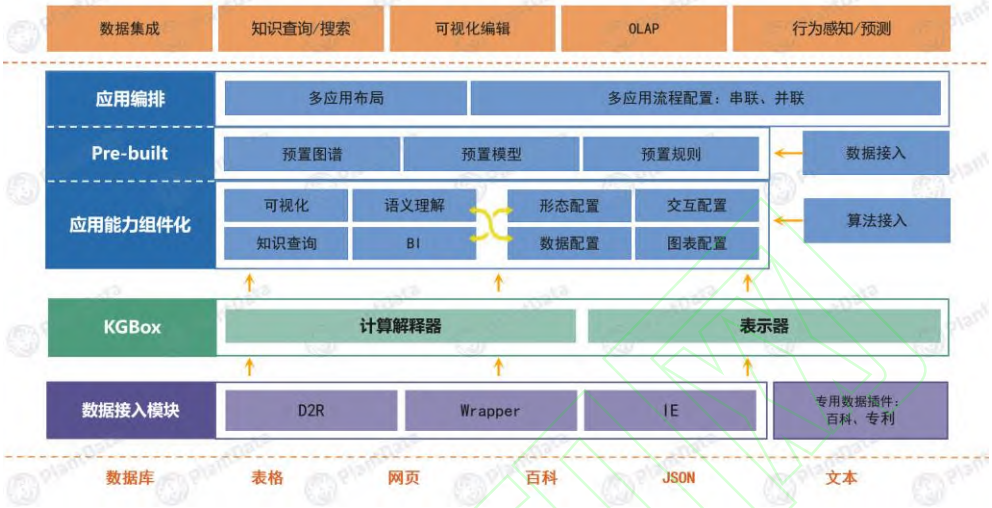

图 2 知识图谱中台基本架构 (The framework of knowledge graph middle-platform)

4.1 组件微服务化

组件微服务化的过程是指对知识图谱平台的各个功能进行抽象与细粒度的拆分，一方面降低单个组件的开发难度，更重要的是能够在不同的应用场景中快速对这些细粒度的进行重新组织从而达到利用的目标。进行抽象的服务包括知识图谱全生命周期的全部服务，包括构建相关的组件及应用相关的组件 ：知识图谱构建组件包括知识建模组件、知识获取组件、知识融合组件与知识存储组件，覆盖行业知识图谱全生命周期中的知识图谱构建阶段；知识应用组件则包括统一检索、智能问答、智能推荐、图挖掘分析、事件分析、交互式 BI、知识服务等组件，包括了知识图谱最典型的应用场景。

通过使用统一的微服务架构实现服务的统一治理、独立运行，实现中台的高可用、可扩展，通过使用容器化相关技术实现服务的快速发布与扩展.。

4.2 预构建

预构建的理念来自于迁移学习；迁移学习和领域适应指的是在一种环境中学到的知识被用在另一个领域中来提高它的泛化性能；即反预训练的模型重新用在另一个任务中。典型的应用案例包括图像识别领域和自然语言处理领域，前者的代表有 VGG 模型、Inception 模型和 ResNet 模型；在自然语言处理领域的应用从早期的词向量模型 (word2vec)开始，到近两年热门的 BERT、XLNet 等。

在知识图谱中台中，预构建的使用分为几个层面，一是直接把预训练的语言模型应用于知识图谱构建过程的知识抽取环节，在数据量非常多的行业中也可以训练专用的领域语言模型。同时，预构建的思想还可以用于知识建模的本体、知识库、模型和算法等；在特定的领域应用场景或项目中所定义的本体、获取的知识库以及算法与模型（面向知识获取、融合及应用），在后续的相似场景中都可以复用而不需要从零开始构建；因此，在新的应用场景中进行迁移时能够在此基础上快速地进行知识图谱的构建与应用，降低应用落地的难度与成本。

4.3 业务编排

业务编排是指通过组合基础服务来实现具体业务。实现业务编排的前提是组件微服务化，既包括后台组件的微服务化，也包括将前端组件转化为微服务。前端组件的微服务化需要使用微前端相关的技术实现前端组件的加载、组件注册、页面路由和数据共享。在组件微服务化的基础上，设计与开发适用于知识图谱可视化、推理、问答、统计等应用场景的所见即所得的拖拽式布局编排引擎。

组件微服务化必须建立在数据模型抽象的基础上，这在灵活多变的业务场景中非常难以实现。因

此，业务编排的难点在于业务数据模型的抽象。而知识图谱的可动态定义本体的能力使得数据模型能够动态进行定义与扩展，建立在此基础上的微服务组件极大程度地增强了系统的可编排能力。

4.4 知识图谱中台落地实践

在知识图谱中台上的应用将演变成“大中台+轻前台”的新范式，即重心在于中台的构建，当中台构建成型后，即可快速的实现业务应用场景的构建。同样以金融创投业务应用为例，面向金融领域的知识图谱中台会经过不断的积累得到领域相关的本体、数据和知识、面向金融领域的知识抽取模型等，以及一些经典的企业竞争力分析、企业风险评估算法和模型；在此基础上构建应用时，用户只需要补充特有的内部业务数据（如创业企业的经营数据），这些数据通常是结构化的，通过简单配置即可整合到知识图谱中；接下来可直接利用上述算法（企业竞争力分析算法和风险评估模型等），或是在它们基础上进行微调（如加入特殊数据，改变权重参数等）得到更新的算法和模型，然后利用编排引擎即可实现业务场景的应用。

因此，相较于基于知识图谱平台的应用构建范式，基于中台的应用构建有如下优势：1）在预构建的数据模式、知识库、算法模型等基础上构建，从而避免数据稀疏和冷启动；2）迁移快，能够有效地复用之前积累的能力；3）业务导向，不需要过多地理解构建知识图谱全过程中涉及的复杂技术；4）基于业务编排快速试探应用的构建，缩减开发周期，节约开发成本。

5 知识图谱发展的挑战与机遇

随着行业知识图谱的应用深化，其应用场景呈现出如下特征：数据向多模态化、动态化方向发展，数据类型不断扩展，尤其是深度知识使用需求逐步增加；另一方面，应用所基于的多类型的数据的质量也参差不齐。这使得知识图谱的应用变得越来越复杂，也难有一种方法（包括知识的表示、存储和应用）能够满足所有的应用需求。

5.1 深度知识的表示与获取

在一些专业的领域如智能运维、医疗辅助诊断等，不仅需要概念、实体和关系这些基础的知识作为支撑，对于动态的事件以及深度的业务经验知识和决策过程知识等的需求更加明显（其中的典型代表是密集的业务专家知识）；这对复杂的知识表示与获取提出了更高要求。

业务经验和决策过程等知识是专家经过长期积累形成的，通常隐含在大段的文本中，有些甚至仅存在于专家的脑子里；对于这些知识的获取，知识众包是一种可行的解决方案。同时，图神经网络和知识图谱表示学习的发展也为深度知识的表示与获取提供了解决问题的方法。

5.2 数据稀疏场景下的知识自动获取

深度学习的发展给知识获取带来了机遇，但是它往往需要大规模高质量标注数据，而在企业应用场景中，高质量语料获取通常需要由领域专家手工标注，这使得其构建成本通常非常高。

针对这种数据稀疏场景下的知识获取，弱监督学习、小样本学习等最新的研究成果提供了解决思路。首先，“无监督的预训练语言模型加上特定任务少量语料微调”的文本处理新范式在信息抽取、语义理解等场景得到了广泛的应用；在公开发布的语言模型的基础上，使用少量的行业语料即可完成高可用模型的训练。更进一步，基于知识增强的语言表示模型通过将知识图谱的信息加入到模型的训练中，使模型可以从大规模的文本语料和先验知识丰富的知识图谱中学习到字、词、句和知识表示等内容，从而有助于其解决更加复杂、更加抽象的自然语言处理问题。

5.3 知识的质量与时效性

企业级知识图谱应用通常对知识的质量要求非常高。然而，从不同来源的数据通过不同方法获取的知识，很难保证它们的质量，尤其是那些通过一些机器学习方法从非结构化数据中提取的知识。另一方面，对于知识尤其是高动态知识的时效性保证也面临着巨大的挑战。

知识评估体系相关研究的新进展为知识质量提升提供了评测依据；同时，知识众包形式的知识编辑与校验也是保障知识质量与时效性的有效手段。

5.4 超大规模知识图谱的性能

随着知识图谱在企业中的深度应用，积累的数据日趋庞大，从数据中获取的知识规模从初始的万级别迅速增长到十亿级别，有些大型的企业的数据规模甚至达到了千亿和万亿级别。这种超大规模的数据对知识存储和计算都带来了巨大挑战，传统的图数据库都难以适应这种超大规模的知识。

计算机硬件的快速发展为超大规模知识图谱提供了存储、算力等方面的支撑；同时，大数据时代积累的分布式计算、并行处理等技术，为超大规模知识图谱知识计算提供了丰富的经验。

6 结束语

知识图谱是大数据时代的知识工程集大成者，是符号主义与连接主义相结合的产物，是实现认知智能的基石。近年来，知识图谱技术在互联网及诸如金融、医疗、教育等垂直行业得到广泛应用。该文从工程角度系统地介绍了大规模企业级知识图谱实践技术细节，全面梳理了已有的知识图谱平台，研究了建设知识图谱平台所需的主要过程和关键难点，针对每个环节详细分析了所需的技术和相应的最佳实践。同时，给出了知识图谱平台中台化升级的挑战、相应的技术路线和未来发展方向。随着知识图谱在企业级场景中应用的不断深入，多类型知识的统一表示与自动获取、海量知识的高效推理与计算、知识的质量与时效性等将成为工程与研究中需要解决的问题。

参考文献

[1] NOY N, GAO Y, JAIN A, et al. Industry-scale knowledge graphs: Lessons and challenges[J]. Queue, 2019, 17(2): 48-75.

[2] JI Shaoxiong, PAN Shirui, CAMBRIA E, et al. A Survey on Knowledge Graphs: Representation, Acquisition and Applications [EB/OL]. [2019-03-15] https://arxiv.org/pdf/2002.00388.pdf

[3] WU Tianxing , QI Guilin , LI Cheng , et al. A Survey of Techniques for Constructing Chinese Knowledge Graphs and Their Applications[J]. Sustainability, 2018, 10(9): 3245.

[4] WANG Quan, MAO Zhendong, WANG Bin, et al. Knowledge graph embedding: A survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.

[5] LIN Yankai, HAN Xu, XIE Ruobing, et al. Knowledge representation learning: A quantitative review [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1812.10901.pdf

[6] LIN H, WANG Y, JIA Y, et al. Network Big Data Oriented Knowledge Fusion Methods: A Survey. [J]. 计算机学报, 2017(1):3-29. (in Chinese)
林海伦, 王元卓, 贾岩涛, et al. 面向网络大数据的知识融合方法综述[J]. 计算机学报, 2017(1):3-29.

[7] ZHAO Xiaojuan, JIA Yan, LI Aiping, et al. Multi-source Knowledge Fusion: A Survey[C]//2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC). Hangzhou, China: IEEE Press, 2019: 119-127.

[8] ZOU Lei, ÖZSU M T. Graph-based RDF data management[J]. Data Science and Engineering, 2017, 2(1): 56-70.

[9] WYLOT M, HAUSWIRTH M, CUDRÉ-MAUROUX P, et al. RDF data storage and query processing schemes: A survey[J]. ACM Computing Surveys (CSUR), 2018, 51(4): 1-36.

[10] WANG X, ZOU L, WANG C. Research on Knowledge Graph Data Management: A Survey[J]. Journal of Software, 2019, 30(7) : 2139-2174. (in Chinese)
王鑫，邹磊，王朝坤，等. 知识图谱数据管理研究综述[J]. 软件学报，2019，30(7) : 2139-2174

[11] GUAN Saiping, JIN Xiaolong, JIA Yantao, et al. Knowledge Reasoning Over Knowledge Graph: A Survey[J]. Journal of Software, 2018, 29(10) :2966-2994. (in Chinese)
官赛萍，靳小龙，贾岩涛，等. 面向知识图谱的知识推理研究进展[J]. 软件学报, 2018, 29(10) : 2966-2994.

[12] LI Wenzhuo, QI Guilin, JI Qiu. Hybrid reasoning in knowledge graphs: Combing symbolic reasoning and statistical reasoning[J]. Semantic Web, 2020, 11:53-62.

[13] HITZLER P, BIANCHI F, EBRAHIMI M, et al. Neural-symbolic integration and the Semantic Web[J]. Semantic Web, 2020, 11(1): 3-11.

[14] PAULHEIM H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic web, 2017, 8(3): 489-508.

[15] KEJRIWAL M, KNOBLOCK C, SZEKELY P. Constructing domain-specific knowledge graphs[C]// Proceedings of the 16th International Semantic Web Conference (ISWC2017). Vienna, Austria: Tutorial, 2017.

[16] GAO Y, LIANG J, HAN B, et al. Building a large-scale, accurate and fresh knowledge graph[J]. KDD-2018, Tutorial, 2018, 39.

[17] SINGHAL A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012, 16.

[18] HE Shizhu, Liu Kang, JI Guoliang, et al. Learning to represent knowledge graphs with gaussian embedding[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia: ACM Press. 2015: 623-632.

[19] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems (NIPS 2013). Stateline, USA. 2013: 2787-2795.

[20] WANG Zhen, ZHANG Jianwen, FENG Jianlin, et al. Knowledge graph embedding by translating on hyperplanes[C]//Twenty-Eighth AAAI conference on

artificial intelligence. Québec, Canada : AAAI, 2014: 1112-1119.

[21] JI Guoliang, HE Shizhu, XU Liheng, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015: 687-696.

[22] NICKEL M, TRESP V, KRIEGEL H P. A three-way model for collective learning on multi-relational data[C]// The 28th International Conference on Machine Learning. Bellevue, USA, 2011, 11: 809-816.

[23] YANG Bishan, YIH Wen-tau, HE Xiaodong, et al. Embedding entities and relations for learning and inference in knowledge bases [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1412.6575.pdf

[24] TROUILLON T, WELBL J, RIEDEL S, et al. Complex embeddings for simple link prediction[C]// Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, USA, 2016.

[25] LIU Quan, JIANG Hui, EVDOKIMOV A, et al. Probabilistic reasoning via deep learning: Neural association models [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1603.07704.pdf

[26] CUI Peng, WANG Xiao, PEI Jian, et al. A survey on network embedding[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(5): 833-852.

[27] ZHOU J, CUI G, ZHANG Z, et al. Graph neural networks: A review of methods and applications[EB/OL]. [2019-03-15] https://arxiv.org/pdf/1812.08434.pdf

[28] WU Zonghan, PAN Shirui, CHEN Fengwen, et al. A comprehensive survey on graph neural networks [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1901.00596.pdf

[29] WANG Daixin, CUI Peng, ZHU Wenwu. Structural deep network embedding[C]//Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. San Francisco, USA. 2016: 1225-1234

[30] HUANG Xiao, LI Jundong, HU Xia. Label Informed Attributed Network Embedding[C]// Proceedings of the Tenth ACM International Conference on Web Search and Data Mining.Cambridge, UK. 2017: 731-739.

[31] LIU Zhengming, MA Hong, LIU S, et al. Network representation learning algorithm incorporated with node profile attribute information[J]. Compute Engineering, 2018, 44(11): 165 - 171. (in Chinese)
刘正铭，马宏，刘树新，等. 一种融合节点文本属性信息的网络表示学习算法[J]. 计算机工程，2018，44(11): 165 - 171.

[32] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//European Semantic Web Conference. Heraklion, Greek: Springer, 2018: 593-607.

[33] CHIU J P C, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[34] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1603.01360.pdf

[35] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1810.04805.pdf

[36] CRAVEN M, KUMLIEN J. Constructing biological knowledge bases by extracting information from text sources[C]// proceedings for the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB-99). Heidelberg, Germany. 1999: 77-86.

[37] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore. 2009: 1003-1011.

[38] ZENG Daojian, LIU Kang, CHEN Yubo, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon. Portugal: 2015. 1753-1762.

[39] BAO Kaifang, Gu Junzhong, YANG Jing. Knowledge Graph Completion Method Based on Jointly Representation of Structure and Text[J]. Computer Engineering, 2018, 44(7): 205 - 211. (in Chinese)
鲍开放，顾君忠，杨静. 基于结构与文本联合表示的知识图谱补全方法[J]. 计算机工程，2018，44(7): 205 - 211.

[40] TAI Chih-Hua, CHANG Ching-Tang, CHANG Yue-Shan. Hybrid knowledge fusion and inference on cloud environment[J]. Future Generation Computer Systems, 2018, 87: 568-579.

[41] SMIRNOV A, LEVASHOVA T, SHILOV N. Patterns for

context-based knowledge fusion in decision support systems[J]. Information Fusion, 2015, 21: 114-129.

[42] HJØRLAND B. Facet analysis: The logical approach to knowledge organization[J]. Information processing & management, 2013, 49(2): 545-557.

[43] DONG Xin Luna, GABRILOVICH E, HEITZ G, et al. From data fusion to knowledge fusion [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1503.00302.pdf

[44] ANGLES R, ARENAS M, BARCELÓ P, et al. Foundations of modern query languages for graph databases[J]. ACM Computing Surveys (CSUR), 2017, 50(5): 1-40.

[45] YAN Da, BU Yingyi, TAN Yuanyuan, et al. Big graph analytics platforms[J]. Foundations and Trends® in Databases, 2017, 7(1-2): 1-195.

[46] MCCUNE R R, WENINGER T, MADEY G. Thinking like a vertex: a survey of vertex-centric frameworks for large-scale distributed graph processing[J]. ACM Computing Surveys (CSUR), 2015, 48(2): 1-39.

[47] CHEN Yang, GOLDBERG S, WANG D Z, et al. Ontological pathfinding[C]//Proceedings of the 2016 International Conference on Management of Data. San Francisco, USA. 2016: 835-846.

[48] COHEN W W, YANG Fan, MAZAITIS K R. Tensorlog: Deep learning meets probabilistic dbs [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1707.05390.pdf

[49] SHI Baoxu, WENINGER T. ProjE: Embedding projection for knowledge graph completion[C]//Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, USA. 2017:1236-1242.

[50] HOHENECKER P, LUKASIEWICZ T. Deep learning for ontology reasoning [EB/OL]. [2019-03-15] https://arxiv.org/pdf/1705.10342.pdf

[51] GUO Shu, DING Boyang, WANG Quan, et al. Knowledge base completion via rule-enhanced relational learning[C]//China Conference on Knowledge Graph and Semantic Computing. Beijing, China: Springer, 2016: 219-227.

[52] BERANT J, CHOU A, FROSTIG R, et al. Semantic parsing on freebase from question-answer pairs[C]// Proceedings of the 2013 conference on empirical methods in natural language processing. Seattle, USA. 2013: 1533-1544.

[53] YAO Xuchen, VAN Durme B. Information extraction over structured data: Question answering with freebase[C]// Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA. 2014: 956-966.

[54] DIEFENBACH D, LOPEZ V, SINGH K, et al. Core techniques of question answering systems over knowledge bases: a survey[J]. Knowledge and Information systems, 2018, 55(3): 529-569.

[55] CHEN Yubo, XU Liheng, LIU Kang, et al. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015: 167-176.

[56] BIZER C, SEABORNE A. D2RQ-treating non-RDF databases as virtual RDF graphs[C]//Proceedings of the 3rd international semantic web conference (ISWC2004). Hiroshima,Japan. 2004.

[57] BIZER C, CYGANIAK R. D2r server-publishing relational databases on the semantic web[C]//Poster at the 5th international semantic web conference. Athens, USA: Springer, 2006: 175.

[58] BAUMGARTNER R, FLESCA S, GOTTLOB G. Visual web information extraction with lixto[C]// Proceedings of the 27th VLDB Conference. Roma, Italy: 2001.

[59] DONG Xin, GABRILOVICH E, HEITZ G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion[C]// Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA. 2014: 601-610.

[60] NENOV Y, PIRO R, MOTIK B, et al. RDFox: A highly-scalable RDF store[C]//International Semantic Web Conference. Bethlehem, USA: Springer, 2015: 3-20.