全国智能计算标准化工作组
STANDARDIZATION WORKING GROUP 32 ON INTELLIGENT COMPUTING
OF STANDARDIZATION ADMINISTRATION OF CHINA

Graph+AI：大模型浪潮下的图计算

全国智能计算标准化工作组图计算研究组 编著

2024年11月

版权声明

编制说明

感谢以下专家和学者对本白皮书编制工作的鼎力支持（注：排名不分先后）。

专家顾问

- **林学民** 欧洲科学院院士、IEEE Fellow、上海交通大学讲席教授
- **金耀初** 欧洲科学院院士、IEEE Fellow、西湖大学讲席教授
- **金海** SAC/SWG32 图计算研究组召集人、华中科技大学教授
- **陈文光** SAC/SWG32 图计算研究组联合召集人、蚂蚁集团技术研究院院长/副总裁
- **石川** 北京邮电大学特聘教授、Gamma Lab 实验室创始人
- **陈华钧** 浙江大学教授、OpenKG 牵头发起人
- **邹磊** 北京大学教授、图数据库 gStore 项目负责人
- **张岩峰** 东北大学教授、计算机学院副院长
- **叶小萌** 杭州欧若数网科技有限公司创始人
- **张晨** 浙江创邻科技有限公司创始人
- **梁磊** 蚂蚁集团知识图谱技术总监

编制组组长

- **洪春涛** 蚂蚁集团图计算负责人
- **陈红阳** SAC/SWG32 图计算研究组联合召集人、之江实验室数据枢纽与安全研究中心副主任

编制组成员

蚂蚁科技集团股份有限公司	**之江实验室**
范志东、林恒、桂正科、郭智慧、孙梦姝、陈发强、刘永超、郑达、彭晋、崔安颀、赵培龙、李少衡、吕松霖、何雨潇、历鹏飞、陈梓康	余婷、余磊、杨林瑶、黄丹丹、蒙贵云
北京邮电大学	**浙江大学**
杨成、黄海	张文、张强、王鑫达
西湖大学	**东北大学**
吴泰霖	陈朝亿、付振波、曹春榆、巩树凤
杭州悦数科技有限公司	**浙江创邻科技有限公司**
古思为、鲍翰林、方扬	周研、童冰
北京大学	**北京交通大学**
吴伟	刘钰
复旦大学	**北京海致星图科技有限公司**
郑卫国、张志杰	沈游人、杨帆、王铮
深圳市腾讯计算机系统有限公司	**信雅达科技股份有限公司**
姚亮、何峰、谢思发、程序	林路、嵇津湘、李云波
北京枫清科技有限公司	
吴敏	

推 荐 语

图计算技术与人工智能，特别是大模型的融合，正在为信息处理和知识表示开辟新的前沿。图结构能够有效表达数据的深层关系，图与大模型的结合显著提升了大模型的逻辑推理能力，在解决大模型幻觉等问题上展现出强大潜力。本白皮书梳理了这一领域的最新进展，并对其未来的可能性进行了分析讨论，为读者提供了一个前瞻性的理解视角。

——林学民，欧洲科学院院士、IEEE Fellow、上海交通大学讲席教授

大模型时代，将图计算与 AI 深度融合有着广阔的前景和重要的影响。这一白皮书对图计算与 AI 融合的关键技术、解决方案和应用案例进行了详尽的梳理，尤其在与大模型的融合、可信图计算、科学研究和产业落地等前沿研究和应用方面，做了精彩的阐述。

——金耀初，欧洲科学院院士、IEEE Fellow、西湖大学讲席教授

在信息科技迅猛发展的背景下，图数据和图应用逐步渗透到各行各业，图技术与 AI 的结合正在开创全新的可能性。本白皮书系统回顾了图智能的发展历程，深入讨论了图的核心技术与应用场景，展现了图技术在大模型浪潮中的关键作用。本白皮书旨在帮助读者深入理解图技术的最新进展与未来趋势，期望为读者带来深刻的行业洞察，进一步推动图智能技术的广泛应用与落地。

——金海，SAC/SWG32 图计算研究组召集人、华中科技大学教授

在当前科技飞速发展的时代，图计算与人工智能的结合展现出巨大的潜力与前景。图计算以其天然适应复杂关系网络的优势，为 AI 模型提供了丰富的结构化信息，使得模型不仅能够理解数据的表层特征，更能洞察其内在关联。随着大模型技术的出现，图+AI 的协同效应必将进一步放大，推动智能系统向更高层次发展。

—— 陈文光， SAC/SWG32 图计算研究组联合召集人、蚂蚁集团技术研究院院长/副总裁

大模型浪潮下的 AI 技术快速发展，对图计算也产生了深刻的影响。该白皮书从数据、算法、应用三个层面对 Graph+AI 的结合方式进行详尽的分析，并针对大模型带来的全新学习范式，提出了图计算面临的新问题与挑战。通过总结以往问题的多种解决方案，并在产业落地与科学研究方面提供大量应用案例，该白皮书将为相关研究者如何发展大模型浪潮下的图计算提供有效参考。

——石川，北京邮电大学特聘教授、Gamma Lab 实验室创始人

本白皮书以大模型技术为背景，全面介绍了图技术在数据、模型和应用等方面的发展趋势。内容涵盖图模型的方法论、详细的技术解决方案以及丰富的实际应用案例，为读者提供了全景式的图技术与人工智能融合的深度解析。

——陈华钧，浙江大学计算机科学与技术学院教授、OpenKG 牵头发起人

图计算作为刻画和挖掘万物复杂关联关系的核心技术，已经广泛应用于诸多应用场景。近来大模型的强大的学习和泛化能力为人工智能的发展带来革命性地影响，如何融合图计算和最新的 AI 技术，已经成为业内共识。本白皮书全面、详实地介绍了“Graph+AI”的研究进展和未来展望，值得大家研读与思考。

——邹磊，北京大学王选计算机研究所教授、图数据库 gStore 项目负责人

本白皮书深入探讨了图数据与 AI 结合的关键技术及其在多领域的应用潜力。内容涵盖了图技术在数据挖掘、模型优化和决策增强等方面中的广泛应用场景，以及丰富的案例与详尽的解决方案，为研究者和从业人员提供了系统性指导，揭示了图技术在大模型时代的关键价值。

——张岩峰，东北大学教授、计算机学院副院长

从事图技术领域多年，我们见证了图技术从学术研究到实际应用的飞速发展，本白皮书正是这一领域最新进展的全面展示和深入探讨。本白皮书紧密结合当前 AI 大模型的浪潮，详细阐述了图技术与数据、算力、模型等多个关键技术的结合，无疑是所有对图技术感兴趣的读者的一本宝贵指南。

——叶小萌，杭州欧若数网科技有限公司创始人

人工智能浪潮势不可挡，图技术和 AI 的结合将带来新的机遇。本白皮书详细分享了图模型的建设方案和应用案例，是对 AI 大模型时代图技术发展路径的一次全面综述。期待本白皮书为每一位读者带来具有前瞻性和全局观的产业洞察分析，加速推动图智能的行业应用落地。

——张晨，浙江创邻科技有限公司创始人

序 言

在数字化时代的浪潮中，图计算与人工智能这两项前沿技术在各自的发展与演变中逐渐交织，形成了一幅生动的科技蓝图。

图计算作为处理复杂关系网络的一种高效工具和计算模式，其起源可以追溯到 18 世纪数学家欧拉提出的“七桥问题”。在 20 世纪 60 年代计算机科学发展的早期阶段，图计算就被应用于网络流优化、最短路径寻找等经典问题，为后续的数据挖掘、知识表示等领域提供了基础。随着大数据和互联网的迅猛发展，图计算在社交网络分析、金融风险控制、推荐系统、生物信息学等多个领域展现出了强大的潜力和应用价值。

同时，人工智能的发展也在不断演变。从 70 年代的专家系统、80 年代的机器学习，到近十年来深度学习的崛起，人工智能技术已经渗透至社会生活的方方面面。尤其是在自然语言处理、计算机视觉等领域，深度学习模型所取得的突破性进展，极大加速了人工智能技术的普及与商业化进程。尽管如此，传统的人工智能方法在处理非结构化或高度互联的数据时仍显不足。正是在这种背景下，图计算与人工智能的融合成为了必然趋势。

在图神经网络出现之前，研究者们已经探索了多种将图计算与人工智能相结合的方法，包括图嵌入技术、概率图模型、图核方法等。图神经网络的出现，标志着图计算与人工智能开始深度结合。图神经网络通过在图结构上进行信息传播和聚合，实现了对图数据的高效建模和特征提取。这种结合不仅提升了人工智能模型在处理图数据时的表现，也解锁了图计算技术在智能化应用中的巨大潜能。

近年来，大规模预训练模型的兴起再次引领了人工智能技术的革命。这些模型凭借其卓越的理解和生成能力，展示了向通用人工智能迈进的可能性与“曙光”。同样的，大模型的出现也为图计算与人工智能的结合带来了新的机遇和挑战，比如，大模型的训练通常需要数量庞大且多样化的数据，图计算在捕捉数据深层次关系方面的能力为这一问题提供了潜在解决方案。而如何构建图基础模型以获得类似大语言模型的涌现能力和强泛化能力则是新的挑战。

在大模型的浪潮之下，如何巧妙地整合图计算和人工智能的优势，进一步深化二者的融合，并开拓更广阔的应用前景，已经成为当前学术界和产业界共同关注的焦点。本白皮书旨在全面解析图计算与人工智能（尤其是大模型技术）的交互现状，探讨其背后的原理、面临的问题与挑战、关键技术以及成功实践。希望通过本白皮书的系统梳理和案例阐述，激发更多关于图与人工智能融合创新的思考与探索，为相关领域的研究和应用提供有益的参考和启示，共同迎接一个充满无限可能的图智能未来。

目 录

第 1 章 背景

自 20 世纪中叶人工智能（Artificial Intelligence, AI）概念提出以来，该领域的发展几经跌宕起伏。随着大数据领域的技术持续突破以及硬件算力的不断提升，以神经网络理论为基础的深度学习技术也逐步从“寒冬”走向各行各业。尤其是随着大模型（Large Language Model, LLM）技术的兴起，AI 技术正带着人类社会迈入下一个纪元。

图（Graph）计算领域也拥有着悠久的历史，最早可以追溯到 18 世纪数学家欧拉提出的 “七桥问题”。伴随着大数据时代数据规模的急剧扩张以及数据关联分析复杂度的提升，图计算技术也迎来了飞速发展，并广泛地应用到社交网络、推荐系统、金融风控、生物信息等领域。

图数据模型在描述复杂数据关联关系以及计算可解释性上有着天然优势，将图计算技术与 AI 技术相结合，并从中发掘出新的技术方向和应用场景，是非常有价值的研究课题。

数据层面，传统的机器学习方法对欧几里得数据有着较好的处理，但在非欧几里得数据上性能不佳，在模态与模型的适配上存在问题。因而我们需要针对性的设计合理的数据形式及处理模型。基于图论的图计算建模方法处理非欧几何数据是合理且自然的，其以节点表示实体，将实体与其特征一一对应，以边表示关系，将实体间的关系显式表示出来。知识图谱（Knowledge Graph）则进一步在图数据上层构建了语义网络，将复杂关系建模为有标签的有向图，以表示事物之间的复杂关系。

算法层面，随着深度神经网络的迅猛发展，以图神经网络（Graph Neural Network, GNN）、图表示学习为代表的方法为机器学习领域带来了新的进展。众多学者尝试将深度神经网络进行合理的改造以适应图的特殊结构，借助其强大的模型性能挖掘更深层次的信息，减少参数量并提高泛化能力。受到大语言模型的启发，图基础模型通过预训练和适应性方法提升模型在各种任务中的表达能力和泛化能力。通过在广泛的图数据上进行预训练，图基础模型能够适应多种下游图任务并具备两种核心能力：涌现和同质泛化。涌现能力意味着当模型参数足够多时，会出现新的功能。同质泛化能力表明模型具有通用性，能够适应多种图任务和不同领域的应用。与语言基础模型相比，图基础模型在数据和任务上存在显著差异。图数据的通用性和多样性使得开发一个“通用图模型”具有挑战。

应用层面，以 LLM 为核心，结合图计算的技术方案和应用场景也在如火如荼的发展，包括但不限于知识图谱、自然语言转图查询（Text2GQL）、图系统优化、图检索增强生成（GraphRAG），以及结合图技术的智能体（Agent）系统等。

- 知识图谱的概念最早源自语义网的研究，目的是让计算机理解互联网中信息的语义，经过多年的发展，知识图谱已经广泛应用于医疗、金融、电商等领域。在实际应用中，知识图谱常用于存储领域知识，包括领域应用中的重要概念以及概念之间的上下位关系。

构建好的领域知识图谱可以服务于各种任务，帮助算法更好地挖掘数据中的隐形关系，实现更智能的推理和决策。

- Text2GQL 是一种将自然语言查询转换为图查询语言（GQL）的技术，旨在帮助开发者和非技术用户更便捷地从图数据库中获取所需数据。通过理解用户的自然语言输入，Text2GQL 能够自动生成相应的 GQL 查询语句，可以简化数据检索的过程，提高效率和准确性。
- 图系统优化是构建工业级的图计算系统过程中需要持续解决的问题，结合 LLM 的优势，可以实现更高效的数据处理和分析、更深入的语义理解、更高效的信息检索和个性化交互等，为各种应用场景提供更有价值的洞察和决策支持。
- GraphRAG 在 RAG 的基础上进行了改进，引入了图结构来构建知识库，并利用图中节点和边的关系来改进信息检索和生成，从而能够捕捉和处理复杂的关系和事务关联，提供更准确、更全面的问答结果。
- KAG 充分融合知识图谱的符号决策和 RAG 的向量检索的优势，通过知识对齐进一步克服 GraphRAG 信息抽取引入的噪声问题，参考 DIKW 知识分层架构构建了知识与 Chunk 互索引结构，在推理问答阶段使用符号逻辑引导的推理和检索有效平衡了复杂决策和信息检索。
- Agent 将 LLM 与现实世界打通，让 LLM 具备类人的自主工作能力，通过图计算技术可以进一步改进智能体的记忆、思考、规划以及行动能力，同时利用多智能体技术，可以进一步改进图应用场景的解决方案生成，为图计算业务带来更多的价值和可能。

总的来看，图计算技术与 AI 技术的结合是一个相互增强的过程。图计算的关联分析性能优势和计算可解释性可以促进 AI 领域的数据质量提升、训练推理加速，以及降低模型幻觉。AI 技术，尤其是大模型技术，可以辅助图计算系统持续的性能改进，降低图计算产品的使用门槛。

第 2 章 问题与挑战

AI 技术使得我们能够更好地处理复杂的图数据，推动了社交网络分析、推荐系统和生物信息学等领域的发展。尽管图计算技术和 AI 技术结合已经取得了显著的进展，但依然面临着诸多的挑战。随着大规模技术的崛起，图技术与大模型的结合有望成为解决这些挑战的重要途径。大模型为图数据的处理和分析提供了新的方法和视角，推动了知识图谱、图神经网络等领域的创新，但同时也带来了新的问题和挑战。

图数据

图数据的收集、存储和使用面临显著挑战。首先，图数据在收集过程中容易受到噪音的影响，这些噪音会沿着边传播，导致更大的危害。动态图和异质图增加了时间维度和节点、边的种类，使得存储和计算要求更高。图数据不仅需要存储节点的特征和标签，还需要存储边及其标签，这使得图的存储更占空间。此外，图数据的标注成本高，标注数据相对较少，进一步增加了处理难度。单一节点特征的信息密度高，处理难度较大，而多模态数据的统一处理也面临巨大挑战。图数据的复杂网络结构和多样性导致任务需求不同，模型需要关注的信息粒度也不同。传统的数据增强方法不适用于图数据，需要针对图数据的特征、结构、标签进行分别增强。图数据的长尾效应导致度数较高的枢纽节点容易被蓄意破坏，造成较大危害。全图的存储和计算不可行，需要平衡采样大小与计算成本，针对不同特性及任务需求采取不同的采样方法才能高效计算。针对这些问题，研究者们提出了多种解决方案。例如，针对动态图的构建需要有效捕捉节点和边的时间变化的问题，研究者们提出了包括基于增量式构建的图流算法的多种动态图数据集构建方法。图数据增强面临的挑战促使研究人员开发专门针对图数据的增强技术，图结构学习就是其中一种重要的方法。为了避免在全图上进行计算，研究者们发明了多种采样技术，包括随机游走采样、邻域采样、聚合采样等等。

图神经网络

图神经网络为图分析提供了一个有效的解决方案，然而，它们在实际应用中仍面临一些关键的挑战。例如，大规模图计算在性能方面存在显著不足，采样方法尚未统一，处理大规模图数据需要更高效的算法。图数据种类繁多，包括同质图、异质图模型存在本质区别，动态图和文本图带来了额外的信息处理需求，这使得模型之间的迁移和泛化能力面临严峻挑战。节点分类的不平衡问题难以解决，采用欠采样和过采样的方法获得的样本在连边上不够真实，严重影响了模型的性能。图神经网络的可解释性较差，继承了神经网络的非线性特性，加之其自身复杂的结构信息，使得预测结果更加难以解释。此外，图神经网络中的节点信息会进行传递和迭代，导致梯度比其他神经网络更容易爆炸。这些挑战表明，图神经网络在处理复杂结构化数据时仍需进一步优化和提升。为了提高大规模图数据的训练和推理，通常需要分布式或/和 CPU-GPU 异构架构下的训练推

理方法。为了提升图神经网络的可信性，研究者们对图神经网络的鲁棒性、公平性以及分布外泛性等多个方面进行了深入的研究。

图基础模型

图基础模型的发展也面临诸多问题和挑战。首先，大规模图数据不易获取，图数据集的规模和多样性不足以支持大图模型的训练。其次，图任务类型多样化，节点级、边级和图级任务的差异性增加了模型设计的复杂性。安全与隐私问题也是一个重要挑战，图基础模型可能面临与大语言模型类似的安全问题，如生成幻觉和隐私风险。提高模型的可信度和透明度，保护隐私是亟待解决的问题。目前，图基础模型缺乏统一的范式，尚未展现出涌现能力和强泛化能力。鉴于大语言模型在自然语言处理中的成功应用，探讨图基础模型如何获取大语言模型的涌现和强泛化能力成为一个重要的研究方向。

知识图谱

以大语言模型为代表的大模型展现了很好的自然语言理解泛化能力，并且被公认掌握了一定的世界知识，这些知识以参数化的形式存储于模型的参数中，并在推理过程中得到应用。大模型给知识图谱技术的发展带来了机遇，也带来了新的问题和挑战。首先，从知识图谱构建的角度来看，大语言模型的语言理解能力是否能够降低知识图谱的构建成本，并提升其规模和质量，使得知识图谱的发展进入一个新的阶段，这是一个值得深入研究的问题。其次，大模型是一种参数化的知识表示和推理技术方案，而知识图谱是一种符号化的知识表示和推理技术方案。在大模型出现之后，如何从知识表示和推理的角度进行协作？哪些知识应该存储于大模型中，哪些知识应该存储于知识图谱中，这些都是需要解决的重要问题。大模型具有很强的任务泛化能力，可以完成许多任务，在大模型时代背景下，如何提升知识图谱技术的泛化性，以便更好地与大模型配合并保留其强大的任务泛化能力，也是一个关键挑战。总的来说，大模型的出现为知识图谱的构建、推理和服务带来了新的视角，有望促使知识图谱技术在未来实现重大突破，与大模型结合，完成大模型时代之前难以实现的任务和目标。

图应用

首先，自然语言转图查询（Text2GQL）面临着诸多挑战。相比于相对成熟的 SQL 语法标准，图查询语言标准（ISO/GQL）尚未全面普及，目前存在多种查询语法并存的状态（如 GQL、PGQ、Cypher、Gremlin、GSQL 等），导致图数据库的使用门槛较高。Text2GQL 研究方向发展较晚，面临几个主要困难：缺乏海量数据集，鲜有公开的 Text2GQL 数据集；缺乏如 Spider 数据集那样的评测标准和对应的评测数据；由于数据集和评测标准的欠缺，各种大模型微调方法的效果难以在 Text2GQL 领域得到验证。可喜的是，在科研工作者不断的探索之下，Text2GQL 已取得到了不错的进展，在数据集方面提出了通过语法制导的生成语料方法，并构建了对应的评测数据，在大模型微调方面，也发展出了多种技术。

图系统优化方面，尽管图计算系统在关联性数据分析性能上有天然优势，但在系统的成熟度、计算存储性能、运维自动化、产品安全性和使用门槛上，仍有巨大改进空间。已有大量的研究将图系统与 AI、LLM 相结合，这样可以充分发挥三者的优势，实现更高效的数据处理和分析，为各种应用场景提供更有价值的洞察和决策支持。

近年来，大规模语言模型在自然语言处理领域取得了显著进展，提升了许多应用场景的智能水平。然而，它们在处理涉及专业领域时仍面临巨大挑战，如生成幻觉、缺乏专业领域知识、信息时效性不足、计算成本高、缺乏可解释性等问题。业界通过检索增强生成（RAG）技术对此做了一定优化，但是通用的 RAG 方法在处理文本分割与索引时无法满足商业场景下的复杂需求，如数据分块（Chunking）的粗粒度方式天然会导致分散的知识丢失，信息间跨相邻分布的关系上下文因为分割而消失，基于字面语义、通识的嵌入（Embedding）易造成误解等，因此，需要一种更精炼、准确、高效、灵活的知识获取方式，如 GraphRAG。

在智能体方面，尽管大语言模型已经具备了一定的思考与决策能力，但要实现与现实世界的交互，具备类人的自主工作能力，还需要大量工作，包括角色设定、记忆、思考规划以及行动等。通过工作流编排单智能体的行为是当前主流的实践手段，但依赖于人工进行的工作流编排，对用户的专家经验有较高的要求。另外，单智能体在处理复杂任务时效果不尽如人意，而采用多个智能体协同工作的策略虽能提高效率，却也带来了系统复杂性和控制难度的增加。目前，设计高效的多智能体系统尚缺乏坚实的理论基础和成熟的应用实例，不过图计算技术可能为这一挑战提供解决方案。

图技术与 AI 技术，尤其是大模型的结合，为信息处理和知识管理开辟了新的路径。尽管面临多重挑战，图技术在大模型时代背景下有望实现重大突破。通过不断优化和创新，图技术和 AI 技术的协同发展将推动更多复杂任务的实现，为各领域带来深远影响。在未来，图技术与 AI 技术的深度融合将进一步提升图数据处理的效率和效果，推动各行业的智能化和数据驱动发展。

第 3 章 关键技术

3.1 图数据处理

3.1.1 图数据构建

在现实世界中，图数据可以用来描述不同领域的关系结构，包括社会科学、化学、生物学等。图数据构建是图计算的关键步骤，其任务是将复杂的现实世界关系建模为计算机可处理的数据结构，这一过程涉及对节点、边以及其属性的合理抽象和表示[15]。节点通常表示图中所描述的对象或实体，边则表示这些对象之间的关系或交互，以社交图为例，节点表示人，边表示社交关系。节点和边通常附带有特定的属性信息，例如在社交图中，人作为节点，其属性可能包括年龄、职业等；而在分子图中，边可能表示化学键，并包含单键、双键等属性信息。节点和边的属性为图模型提供了上下文信息，使算法在计算节点或边的表示时能结合更多维度的数据，通过对这些属性信息的充分利用，图计算可以更好地刻画出节点及其关系的本质特征，从而提升模型在节点分类、链路预测等任务中的性能[16]。

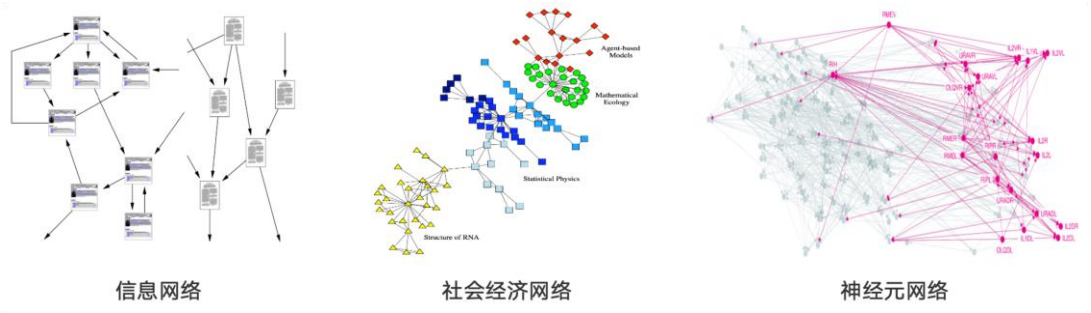

图 3.1 不同领域图数据的构建[53, 54, 55]

图数据构建主要包含数据采集与预处理、节点与边的信息抽取、数据存储与转换这几个关键步骤。

1）首先，需要从现实世界中收集合适的数据，这些数据可以来源于数据库、传感器等多种途径。由于收集到的数据包含噪声、不完整或冗余的信息，因此通常需要清洗和过滤，从而保证数据的质量。

2）在节点与边的信息抽取阶段，需要将数据中的实体和关系映射为节点和边，并提取保存节点和边的属性信息，必要时还需要对边进行加权处理，以反映关系的重要性或强度。

3）构建好图数据后，通常会将其存储为标准的图数据格式，例如邻接矩阵或边列表。

在实际应用中的交互系统非常复杂，因此图数据的构建面临着多重挑战。例如，即使经过预处理，图数据中仍可能包含难以识别和处理的噪声。为了解决这一问题，研究者提出了基于统计特性的噪声检测和清洗方法，如图数据去噪中的图平滑技术和鲁棒图建模算法[47]，这些方法可以

通过检测异常的节点和边来移除噪声，从而提高图数据的质量。此外，许多应用中的数据往往是动态变化的，因此动态图的构建需要有效捕捉节点和边的时间变化。针对这一问题，研究者们提出了多种动态图数据集构建方法，如基于增量式构建的图流算法（Graph Stream Algorithms），它能够在数据更新时动态地调整图结构[48]。同时，时间维度的建模方法（如基于时序的图数据生成技术）[49]，能够将时间信息整合到图数据集中，以捕捉节点和边随时间变化的特性。在复杂系统中，图数据往往由多种类型的节点和边构成，形成异质图。异质图的构建与存储挑战更大，因为需要合理地抽象和建模不同类型的关系。为应对这一挑战，当前提出了异质图数据集构建框架，如 HIN-Mine[50]，它通过对不同类型节点和边的特征提取和语义关系建模，有效构建和存储异质图数据集。通过这些解决方案，图数据集构建能够更加准确地处理现实世界的复杂数据，确保生成的图数据能够有效支持后续的图计算任务。

此外，现实世界的数据常具有多模态特性，例如网页数据可能同时包含文本、图像、视频和音频等。为了综合考虑不同模态中的丰富信息，可以通过图建模的方式进行有效地整合，从而帮助挖掘多模态数据间的复杂关联信息。在多模态图学习中[60]，首先需要对多模态数据进行异质图建模，将各模态的数据视为不同类型的节点，并根据数据点之间的关联关系构建节点间的边。例如，对于一个包含文本、图像、视频和音频的网页数据，可以为每种模态构建对应类型的节点：文本节点、图像节点、视频节点和音频节点；不同模态数据间的关联通过边来表示，例如一段文本对应一张图片时，在文本节点和图像节点之间添加一条边；如果两段文本存在关联，也可在相应的文本节点之间添加边。与传统多模态学习方法多聚焦于两种模态的关系不同，图建模能够灵活地处理多种模态数据，能有效避免训练中对某一模态的过度关注或忽视。考虑到多模态数据的复杂特性，如时序动态性等，如何对这些特性进行有效建模，在确保模态信息完整和关联关系精准捕捉的前提下，使其在各种变化中具有更好的鲁棒性和持续学习能力，也是未来多模态图数据构建的重要研究课题。

3.1.2 图数据增强

图数据增强是一种通过生成、修改数据来扩展现有训练数据的技术，旨在提升图计算的性能[17,61]。与图像或文本数据的增强不同，图数据的结构是非欧几里得的，因此传统的数据增强操作，如裁剪或翻转，不能直接应用于图数据。这使得图数据增强面临更多挑战，并促使研究人员开发专门针对图数据的增强技术。

根据增强对象的不同，图数据增强可分为结构增强、特征增强和标签增强。结构增强通过添加、删除图中的节点或边来改变图结构，从而生成新的图数据。例如，DropEdge 通过随机移除部分边来增强现有数据集，从而缓解图神经网络的过平滑问题[19]。特征增强则通过随机掩码或添加扰动等方式修改节点特征。标签增强在训练数据的基础上生成新标签，如混合不同类的图数据并为新生成的数据分配新的标签。根据增强方法是否需要学习，又可分为基于规则的增强方法和可

学习的增强方法。基于规则的数据增强通过预定义规则来修改图数据，无需学习任何参数，其优点在于实现简单且效率高；可学习的数据增强通过学习优化图结构或特征来生成增强数据，其通常通过模型训练迭代优化图数据结构，并在增强过程中不断改进[20]。图数据增强技术可以应用于有监督学习和自监督学习场景。在监督学习中，数据增强主要用于缓解模型的过拟合现象，提升模型的泛化能力；在自监督学习的对比学习等框架中，图数据增强可以用于生成正负样本，通过拉近与正样本的距离、最大化与负样本的差距来训练模型。

总的来说，图数据增强技术在不增加额外标注成本的前提下，生成更多训练数据或提升图数据质量，从而有效提高了图计算的性能和鲁棒性。

3.1.3 图采样

由于图通常包含大量节点和边，直接在全图上进行计算可能会带来巨大的时间和空间开销，因此采样技术成为了图计算中的关键技术。图采样通过选择部分节点或子图，构造能够代表原始图全局或局部特征的子集，确保在减少计算成本的同时，依然能够得到有效的学习效果。

常见的图采样方法有随机游走采样、邻域采样和聚合采样等。随机游走从一个节点出发，随机选择相邻节点进行访问，从而有效保留图的局部结构信息，能够灵活捕捉图的不同模式[21]；邻域采样从节点邻居中随机采样一部分节点，然后对这些节点的特征进行聚合计算[18]，能够减轻全图计算的压力，并且通过采样保持了图的局部结构信息；层次聚合采样是对图的多个层次结构进行抽象和采样，使得每个层次都保留原图的关键信息，能够在保持全局图结构的同时有效减少冗余计算，从而在大规模图上表现出良好的性能。这些方法通过不同的策略提取图的局部结构信息，从而有效捕捉图的全局特征。

图采样需要考虑如何平衡样本大小和计算开销之间的关系。此外，在实际应用中，图的不同特性和任务需求可能需要不同的采样策略，选择适当的采样策略，才能在实现高效的同时保证图计算的性能。

3.2 图神经网络

3.2.1 图神经网络

图作为一种非欧几里得数据结构，具有强大的表达能力。随着图在各个领域的应用越来越广泛，对利用机器学习分析图的需求也日益增长。传统的机器学习方法在处理图数据时往往依赖于手工设计的特征，这不仅增加了数据处理成本，也限制了模型的灵活性。GNN 的出现为图分析提供了一个有效的解决方案，通过深度学习的方法自动学习图的结构特征，从而提高了模型的性能和泛化能力。

图表示学习方法的兴起，特别是 DeepWalk、Node2Vec 和 LINE 等，为 GNN 的发展提供了基础[62] [63] [64]。这些方法通过学习低维向量表示，捕捉了图中的结构信息。但是其更多地依赖于随机游走或预定义的采样策略，这可能无法充分利用图的局部和全局结构信息。CNN 在图像领域内取得了不错的成就，但它们的通用性受到限制。图像数据等欧式数据可以认为是图数据的一个特例，如何将图像领域的成果迁移到更复杂的图网络也越来越受到了人们的关注，但是将深度神经模型扩展到非欧数据上很难定义局部卷积过滤器和池化算子，这阻碍了从欧几里得域到非欧几里得域的 CNNs 的转化[65]。本节依次介绍图神经网络类几种经典的卷积、池化算子并简要概述图神经网络的前沿相关的开放问题，展望未来图神经网络的发展。

3.2.1.1 卷积算子

图神经网络的卷积算子根据操作域和图结构类型可分为频域（或称为谱域）和空间域卷积，以及同质和异质卷积。频域卷积利用图拉普拉斯矩阵的特征分解在谱域上定义滤波器，以捕捉图的全局结构信息，而空间域卷积直接在图的结构空间中进行局部邻域的信息聚合[65]。在同质图中，所有节点类型相同，卷积算子简单一致，而在异质图中，节点类型多样，卷积算子需要处理不同类型节点间的复杂关系。如下依次介绍几个比较经典的卷积算子。

1、GCNConv

GCN（Graph Convolutional Network）是一种经典的谱域的图卷积算子[66]，其是基于图信号处理理论的一种方法。谱域卷积网络是通过在图的谱域上进行操作来实现卷积的，类似于传统卷积网络中的频率卷积。在图上，节点和边的关系可以用“频率”来描述，类似于我们用频率分析声音或图片。我们通过图的拉普拉斯矩阵来计算这些频率。在频率空间上卷积，可以理解为用某种“滤镜”处理图上的数据，提取出有用的信息。图的卷积操作就是将图的信号（节点特征）在频率空间上进行滤波。但是直接做频率计算很慢，因此 GCN 使用近似方法来加速。这个近似通过数学方法把复杂的操作简化为图上节点和邻居之间的“信息传递”。GCN 的操作可以看作是每一层，节点和它的邻居交换信息，通过权重矩阵和非线性激活函数来更新节点的特征。

2、SAGEConv

GraphSAGE（Graph Sample And Aggregation）是一种基于 MPNN（Message Passing Neural Networks）架构改进的图卷积方法，特别适合处理大规模图[67]。它的关键特点是通过采样和聚合节点的邻居来进行特征更新，在大图中，每个节点可能有成百上千的邻居，直接使用所有邻居更新特征代价太大。GraphSAGE 通过随机采样每个节点的一部分邻居，减少计算负担。每个节点通过它采样到的邻居节点进行特征聚合。聚合方式可以有多种，比如求平均（mean）、求和（sum）、最大值（max）等。聚合邻居特征后，节点会结合自己的特征来更新，类似于将“邻居的影响”和“自身的信息”一起考虑。GraphSAGE 的设计让它非常适合在超大图上使用，因为它只采样部分邻居，所以计算量不会随着图的大小成比例增加。

3、GATConv

GAT（Graph Attention），图注意力网络是通过注意力机制在图结构数据中进行节点特征更新的[68]。与其他图卷积网络不同，GATConv 通过自适应地为每个邻居分配权重，重点关注对节点最重要的邻居，GATConv 引入了注意力机制，允许每个节点赋予不同邻居不同的重要性。在传统的 GCN 和 GraphSAGE 中，节点与所有邻居的影响通常是均等或固定的（例如通过平均聚合），但在 GAT 中，每个邻居会被分配一个自适应的权重，反映它们对当前节点的重要程度。节点的特征聚合不再是简单的平均或求和，而是通过加权求和。每个邻居的特征都会乘以一个注意力权重，这个权重是通过节点之间的特征相似性计算得到的，每对节点的注意力分数是通过一个可学习的注意力函数计算的，计算出它们的相似度，并用这个相似度作为注意力权重。

4、RGCNConv

RGCN（Relational Graph Convolution）是图卷积网络的一个扩展，专门用来处理异构图，即图中的边有不同的类型或关系[69]。在 RGCN 中，节点之间的连接不仅仅表示简单的邻居关系，还表示不同类型的关系。RGCNConv 通过引入关系类型的概念，帮助网络处理更加复杂的图结构数据，特别适合像知识图谱这样的场景。在普通的图卷积网络中，所有节点的连接边都是相同的，没有区分不同的关系类型。而在 RGCN 中，每条边都表示一种特定的关系类型，例如在知识图谱中，“人”可以通过“朋友关系”连接到其他“人”，也可以通过“工作关系”连接到一个“公司”。RGCN 通过对不同关系类型分别处理，使得模型能在复杂的异构图中工作。对于每一种关系类型，RGCN 会为其单独计算一个卷积操作。这意味着在 RGCN 中，每个节点的特征更新要考虑到所有不同关系类型的影响。与标准的 GCN 类似，RGCN 也是通过邻居节点的信息来更新每个节点的特征。不同之处在于每个邻居节点的特征聚合过程要根据关系类型来区分。

3.2.1.2 池化算子

在计算机视觉领域，卷积层通常跟随一个池化层以获得更通用的特征。复杂和大规模的图通常具有重要的分层结构，对于节点级和图级分类任务非常重要。池化算子主要用于对图进行下采样和特征聚合，帮助模型从复杂的图结构中提取更具全局性的表示。在图神经网络中，池化层通过减少节点或边的数量，对图进行下采样。这种降维操作帮助简化图结构，降低图的复杂度，保留重要的子结构，从而使模型能够在更低维的空间中进行学习，池化层可以通过在图的不同区域进行聚合，帮助模型从局部信息转向全局信息。对于大型图，逐层池化可以使模型获得更加抽象和全局的图表示，进而捕捉图的宏观结构，提升模型在图分类等任务中的性能。

1、SimplePool

SimplePool 通过不同的节点选择策略直接学习图级别的表示。在一些变体中，这些模块也被称为读出函数。一些模型使用简单节点池化方法。在这些模型中，对节点特征进行节点最大值/平均值/求和/注意力等操作，以获得全局图表示。

2、DiffPool

DiffPool（Differentiable Pool）是图神经网络中的一种经典的分层池化方法，它通过可微分的方式学习图的层次化结构，从而实现图的多层级抽象和下采样[70]。相比于简单的池化方法（如最大池化或平均池化），DiffPool 不仅仅是简单地聚合邻居节点的特征，而是动态地学习如何将节点聚类到某些超节点上，形成图的更紧凑表示。DiffPool 的关键在于通过神经网络直接学习图的层次结构，并且这个结构可以在模型的训练过程中动态调整。它通过学习一个软分配矩阵，将图中的节点映射到若干聚类，然后在每个聚类中进行特征的聚合。DiffPool 能够自动学习图中节点之间的聚合关系，从而动态生成更小的图。它能够处理具有复杂拓扑结构的图，而无需事先指定图的层次信息。通过逐层池化和聚合，DiffPool 可以捕捉到图的全局结构。每一层都对图进行下采样，使得最终的输出是图的紧凑、高层次表示，有助于提升图分类、聚类等任务的性能。对于节点数不固定或结构多样的图，DiffPool 提供了灵活的处理方式，通过学习层次结构来适应不同的图结构，尤其适用于图分类任务。

3、gPool

gPool（Graph Pool）是图神经网络中的一种经典的分层池化方法，它通过学习节点的重要性得分来选择节点，并动态地对图进行下采样[71]。gPool 的核心是通过一个可训练的得分函数来计算每个节点的重要性分数。这个分数用于选择节点，从而将图的结构和节点特征压缩为更简洁的形式。它不仅可以减小图的规模，还能保留重要的结构信息，增强图神经网络的全局表示能力。gPool 使用一个可训练的投影向量来计算每个节点的得分。得分通过节点特征与投影向量的内积来计算，用来衡量它的重要性。gPool 按得分从高到低排序，并选择得分最高的前 k 个节点。这个 k 通常为总节点数的一个固定比例。通过这种方式，gPool 保留得分最高的节点，并丢弃得分较低的节点。

4、SAGPool

SAGPool（Self-Attention Graph Pool）是一种基于自注意力机制的图神经网络池化方法。它通过学习节点的重要性得分来对图进行下采样，并保留图中的关键结构。SAGPool的主要贡献在于，它将图卷积与自注意力机制相结合，动态选择图中的重要节点，从而在降低图的复杂度的同时，保留图的全局和局部信息[72]。SAGPool 利用图卷积层（GCN）来计算每个节点的重要性得分。通过图卷积操作，每个节点不仅考虑了自身特征，还聚合了其邻居节点的信息，从而形成一个全局性的节点表示。SAGPool使用自注意力机制为每个节点分配得分，根据得分的大小对节点进行排序，并选择得分最高的前 k 节点。k 通常是节点总数的一个比例。这一过程可以通过阈值控制或动态比例来实现，被选择的节点会形成一个新的子图，保留的节点的特征和结构将继续用于后续的网络层处理。特征矩阵和邻接矩阵会根据选中的节点进行更新，以仅包含这些关键节点及其对应的边。SAGPool的自注意力机制允许每个节点通过邻居的特征计算其重要性，这种机制能在池化过程中保

留全局上下文信息。因此，SAGPool 不仅通过图卷积捕获局部信息，还通过自注意力机制为每个节点分配权重，增强池化操作的表达能力。

5、EdgePool

EdgePool 是一种图神经网络中基于边坍缩的经典分层池化方法，它主要通过对边进行池化来减少图的复杂性。这种方法在图的降维过程中不同于传统的节点池化方法，而是通过学习重要的边来优化图结构，从而得到一个精简但具有重要结构信息的子图[73]。EdgePool 的核心在于通过对图的边进行池化来实现图的下采样。它通过学习每条边的重要性来选择保留的边，从而得到一个更加紧凑的图表示。相较于节点池化方法，EdgePool 专注于保留图的关键边，保持图的结构完整性。EdgePool 使用一个学习到的边权重来评估每条边的重要性。这些边权重可以通过神经网络计算得到。根据边的重要性得分，EdgePool 按得分从高到低排序，并选择得分最高的前 k 个边。边的选择过程可以通过设置阈值或按比例选择来实现，在选择了重要的边之后，EdgePool 会更新图的邻接矩阵，保留这些关键边。更新后的邻接矩阵仅包含保留的边的信息。节点的特征矩阵 X 也会保持不变，但图的结构被简化为仅包含重要的边。在更新图的结构后，EdgePool 会根据保留的边来重新聚合节点特征。节点的特征通过邻接矩阵中的边信息来重新计算。

3.2.1.3 展望

尽管图神经网络（GNN）在各个领域取得了显著成功，但它们在实际应用中仍面临一些关键挑战和开放问题。

鲁棒性：GNN 易受到对抗攻击，这些攻击不仅针对节点特征，还涉及图结构信息。尽管已有一些防御方法被提出，但仍需进一步增强模型的鲁棒性，以应对复杂的对抗攻击。

可解释性：GNN 通常被视为“黑匣子”，缺乏明确的解释能力。虽然已有少数方法尝试为 GNN 模型提供示例级别的解释，但在现实应用中，提升 GNN 的可解释性仍然至关重要。

图预训练：类似于计算机视觉和自然语言处理中的预训练方法，图数据的自监督学习和预训练也显示出潜力。然而，目前在图预训练领域仍面临许多挑战，如设计有效的预训练任务和评估现有模型的学习能力。

3.2.2 训练和推理

3.2.2.1 图神经网络执行模式

图神经网络是一种用于处理图结构数据的深度学习模型，旨在同时捕捉拓扑信息和特征信息。图神经网络通过堆叠多个图广播层为图中的每个节点生成一个包含聚合邻居信息和特征信息的嵌入表示。具体来说，每一层的计算模式可以被抽象成四个计算步骤[81] [82] [83]：ScatterToEdge,

EdgeForward, Gather&Aggregate, Vertex Forward。下图是一个单层计算模式的示例（以节点 2 为例）。

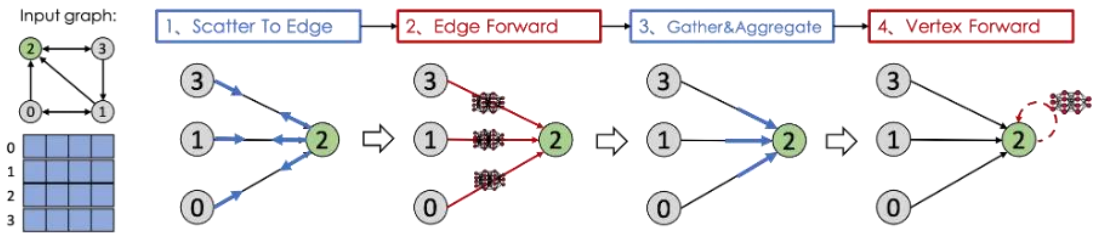

图 3.2 图神经网络计算过程

ScatterToEdge 是一个边消息生成操作，用于将源节点和目的节点的表示向量（属性向量）发送到边上用于参数化的神经网络计算；EdgeForward 是一个定义在边上的神经网络计算，通过合并源节点和目的节点的表示来计算边上输出消息；Gather&Aggregate 是一个聚合计算，用以接收边上的信息并进行聚合（如求和、平均等）以更新自身表示 VertexForward 是一个定义在节点上的神经网络计算，通过神经网络来更新聚合的信息节点表示。

新的节点表示再进入下一层执行计算。通过堆叠多个图广播层可以构建一个深层的图神经网络任务以扩大节点聚合信息的范围。最终，经过多层计算得到的节点嵌入可以用于下游计算任务，例如节点分类，图分类等。

根据运行环境和应用场景，GNN 的训练和推理可以分为单机 GNN 和分布式 GNN。单机 GNN 的训练和推理是指在单个计算节点上完成 GNN 模型的训练和推理，适用于数据规模相对较小或计算资源较为有限的场景。它具有实现简单、易于调试的优点，但在处理大规模图数据时会面临计算瓶颈和内存限制的问题。分布式 GNN 的训练和推理则通过将计算任务划分到多个计算节点上进行并行处理，以应对大规模图数据和复杂的模型训练。分布式 GNN 能够显著提升计算效率和模型性能，但其实现相对复杂，需考虑计算节点之间的数据传输和同步问题。总体而言，选择使用单机或分布式 GNN 取决于具体的应用需求和计算资源的可用性。

3.2.2.2 CPU-GPU 异构架构下的训练推理

在 CPU-GPU 异构框架下的训练和推理中，全图训练与微批量训练有着不同的挑战。全图训练指的是使用数据集中全部节点的全部邻居执行 GNN 训练；而微批量训练只针对部分有训练标签的节点，对它们执行采样算法，即只使用部分邻居进行 GNN 训练。

全图训练中，内存资源是首要限制。有限的 GPU 内存可能无法一次性加载整个大图。NeutronStar[84]使用分块的技术，将大图分为多个子图，在训练过程中逐块加载到 GPU 进行训练。然而，全图训练使用全邻居聚合范式以及全局梯度下降算法，子图之间的共同邻居需要被频繁传输，并且子图训练产生的中间结果需要持续累积，直到反向传播阶段才能释放。因此，ROC[85]提

出将中间数据从 GPU 传回 CPU，但这严重增加了传输量。HongTu[86]使用了重计算技术通过重新计算代替存储（传输），并且通过缓存邻居减少了频繁的信息传输。

微批量训练中，采样-聚合-训练的范式已经成为广泛应用的策略。该范式将训练过程分解为三个独立的步骤：图采样、特征提取和训练，并将这些步骤部署在不同的计算设备上，以实现高效的计算性能。异构环境下的 GNN 训练包括以下方法：第一，CPU 采样、特征提取，GPU 训练。这种方法引入了显著的内存访问开销。第二，将采样、提取放置在 GPU 执行，CPU 仅负责存储全图的特征。这种方法仍然存在频繁的数据传输。

此外，为了提高大规模图数据的训练效率，CPU-GPU 之间的数据传输，缓存以及流水线并行等策略被频繁应用。

数据传输是指数据在 CPU 和 GPU 之间的交换，主要通过 PCIe 或 NVLink 等高速总线完成。通常，图数据和节点特征存储在 CPU 的内存中，训练时，CPU 将需要的数据传输到 GPU。这包括采样后的子图结构和相应的节点特征。在这个过程中，通信的效率直接影响系统的整体性能。因此，一些优化传输的策略如下：第一，将频繁使用的节点特征缓存到 GPU 内存中，避免重复传输数据。第二，将较简单的任务（如采样）放在 CPU 执行，而将计算量大的任务放在 GPU 执行，以此均衡两者的负载，减少资源争用问题。

缓存技术指的是将频繁访问的节点特征、邻接关系或嵌入预先存储在 GPU 内存中，以减少频繁的 CPU-GPU 数据传输，有效减少了 CPU-GPU 之间的通信负担，从而提升性能。例如，NeutronOrch[87]通过热度感知的嵌入重用技术可以识别训练中频繁访问的“热节点”，并将这些顶点的嵌入数据预先存储在 GPU 中，从而提高训练效率。DUCATI[29]不仅缓存节点特征，还缓存一部分常用的图拓扑进一步提高采样、训练的效率。

流水线技术是指异构设备并行处理不同的任务。在分批次训练时，数据通常是逐批加载的，缓存部分子图或节点特征到 GPU 有助于加快每批次的处理速度。为了进一步优化，NeutronOrch[1]通过超批次流水线技术将多个批次组合在一起，让 GPU 和 CPU 任务并行执行，以减少空闲等待时间。

3.2.2.3 分布式训练推理

在分布式图神经网络系统训练和推理中，为了提高大规模图数据的训练效率，通常需要结合多种策略来优化计算和通信性能。这些策略主要包括并行加速、图划分、通信优化和迭代加速等方法，它们从不同的角度解决分布式环境下的计算负载、通信开销和模型收敛性问题。

并行加速策略主要包括流水线并行[88]、数据并行[89]和张量并行[90]。流水线并行将模型按层划分，不同设备同时处理不同批次不同层的数据并更新各自的参数；数据并行则将数据划分给多个设备，每个设备拥有完整的模型副本，独立进行前向和反向传播后汇总梯度更新模型；张量

并行通过将节点特征或嵌入按维度切分到多个设备，每个设备处理一部分张量并同步必要信息，最终汇总梯度完成模型更新。

图划分策略则包括哈希、Metis[91]、Metis-extend 和流式划分四种方法。哈希划分通过随机映射顶点以平衡负载，但未考虑图神经网络的 L 跳邻居关系，通信负载较重；Metis 通过最小化割边将图划分为大小相等的子图，并尽可能减少子图之间连边，从而减少通信；Metis-extend 进一步优化了 Metis 算法，使用聚类算法和额外约束，确保子图中的邻居集中同时也平衡不同子图的节点和边的数量；流式划分则采用动态策略，虽然其优先考虑减少子图间连边从而减少通信开销，但未能充分考虑图的密度和 L 跳邻居的分布，可能导致计算和通信负载不平衡。

通信优化算法旨在通过提升通信效率来改善训练性能，分为无损和有损两类。无损通信优化通过优先级缓存[92]和部分缓存[93]等技术对节点特征数据进行缓存，显著提高了数据缓存利用率和传输效率，并确保模型的准确性不受影响。有损通信优化则通过边界节点的随机采样、选择性丢弃部分节点数据[94]以及对通信数据进行量化[95]，减少了传输数据量，不过也降低了数据的精度。尽管有损策略引入了精度损失，但适度的削减在保证模型性能的同时，有效缩短了训练时间，显著加速了整个训练过程。

迭代加速策略通过同步异步混合模式优化训练效率与模型准确性之间的平衡。该模式结合了同步和异步机制，适应不同的网络和计算需求，提升训练性能。陈旧的同步并行（SSP，Stale Synchronous Parallel ）中的有界陈旧性允许异步训练[96]，并在固定迭代次数后进行同步更新。这样既能利用异步训练的高效率，又能通过定期同步保证模型的收敛性和稳定性。SSP 的这种灵活性使得它能够更好地适应不同的硬件和网络环境，在多节点分布式训练中有效平衡性能与准确性。

3.2.3 可信图机器学习

随着图神经网络的迅速发展，它们在处理图结构数据方面显示出了卓越的能力，被广泛应用于金融分析、交通预测、药物发现等高风险场景。然而，尽管图神经网络在真实世界中具有巨大的潜力，最近的研究显示它们可能泄露私人信息、易受对抗性攻击、可能从训练数据中继承并放大社会偏见，并且难以泛化到分布外数据，这些风险可能无意中对用户和社会造成伤害。例如，已有研究表明，攻击者可以通过在训练图上进行微小的扰动来欺骗图神经网络，使其产生他们期望的结果；在社交网络上训练的图神经网络可能将歧视嵌入其决策过程中，加强了不希望看到的社会偏见。因此，从多个方面提升图神经网络的可信性，如图神经网络在鲁棒性、公平性、以及分布外泛化等方面，以防止这些潜在的伤害，并增加用户对图神经网络的信任变得尤为重要。

3.2.3.1 图神经网络的鲁棒性

深度学习模型通常缺乏对抗鲁棒性，即模型很容易误分类对抗样本。对抗样本是经过精心设计或修改的输入样本，目标是误导模型产生错误的预测结果或降低模型的性能。只有模型对对抗

攻击能够保持稳定的性能，模型才是对抗鲁棒的。对于图像分类任务，攻击者可利用梯度信息构造微小扰动，添加到原始图片以生成对抗样本，使人眼难以发现对抗样本与原始样本的区别，但深度学习模型会以很高的概率将对抗样本错分为其他类别。这表示深度学习模型的假设或设计存在漏洞，依赖于一些非本质的特征，例如模型通过复杂深度模型建模的数据间的统计特征。这将阻碍深度学习模型在法律、金融、医药、军事、人脸识别、自动驾驶等安全敏感领域的应用。为此，对抗攻击作为一个强大的安全分析工具，常被用于探测深度学习模型的漏洞、发现安全隐患，构建可信的人工智能系统。随着对抗攻击的发展，揭示出了模型的脆弱性，而相应的多种防御技术也相继被提出。这个领域在攻防竞赛过程中进一步深入探索了深度学习鲁棒性[3]。

作为深度学习在图上的扩展，图神经网络也可能存在着对抗风险，考虑到图神经网络已在各个领域被广泛应用，研究其对抗鲁棒性具有重大实际意义。然而，图神经网络有着不同于深度学习的对抗鲁棒性：一方面，不同于图像具有连续的像素特征空间，图神经网络应用的图数据包含着特征、拓扑以及标签等多类型数据，且拓扑结构信息是离散的，这给扰动的生成以及不可见扰动的定义带来巨大挑战；另一方面，图数据中不同实例（节点）之间并非完全独立，实例之间存在着关联关系（边），即操纵一个实例可能通过消息传递影响到其他实例。因此一些研究者开始深入探索图神经网络的鲁棒性，如图所示，在原始图上生成微量的拓扑扰动和特征扰动，使得图神经网络错误预测目标节点的标签。具体而言，研究者尝试向拓扑攻击模型中引入更精确的梯度近似方式以生成高效离散拓扑扰动，并重新定义了拓扑结构下的隐蔽性，例如通过限制扰动边总个数来达到隐蔽扰动。

随着人们对于图神经网络安全性的关注，图对抗攻防研究不断取得新的进展。主要研究方法有：

对抗训练：对抗训练是一种流行且有效的方法，广泛应用于计算机视觉中防御逃避攻击。这种方法同时生成可以欺骗分类器的对抗样本，并让分类器对原始样本及其扰动版本给出相似的预测，从而提高分类器的鲁棒性。同时，对抗训练这一方法也被用于防御图对抗攻击。

认证鲁棒：虽然多种方法如图对抗训练可以提高对对抗样本的鲁棒性，但总有可能会开发出新的攻击方法使得防御措施失效，导致一场无休止的攻防赛。为了解决这个问题，最近的工作开始分析图神经网络的认证鲁棒性，以了解最坏情况下的攻击将如何影响模型。认证鲁棒性旨在为潜在扰动下仍然鲁棒的节点提供证书。这些证书通过解优化问题获得。此外，还可以通过随机平滑技术注入噪声到测试样本中以减轻对抗性扰动的负面效应，并提供认证保证。这种方法证明了在特定条件下图神经网络的预测是稳定的。

3.2.3.2 图神经网络的公平性

公平性是可信图神经网络中最重要的方面之一。随着图神经网络的迅速发展，图神经网络已被应用于多种场景。然而，近期的研究表明，类似于传统机器学习模型处理独立同分布数据时所

表现出的问题，图神经网络也可能因数据中存在的社会偏见而给出不公平的预测结果。例如，在图神经网络的书籍推荐系统中，因为男性作者较多，图神经网络可能偏向于推荐男性作者的书籍，表明图神经网络可能对少数群体存在歧视，从而导致社会问题。此外，这种歧视可能严重限制图神经网络在其他领域的广泛应用，如职位申请者排名和贷款欺诈检测，并可能引起法律问题[132]。

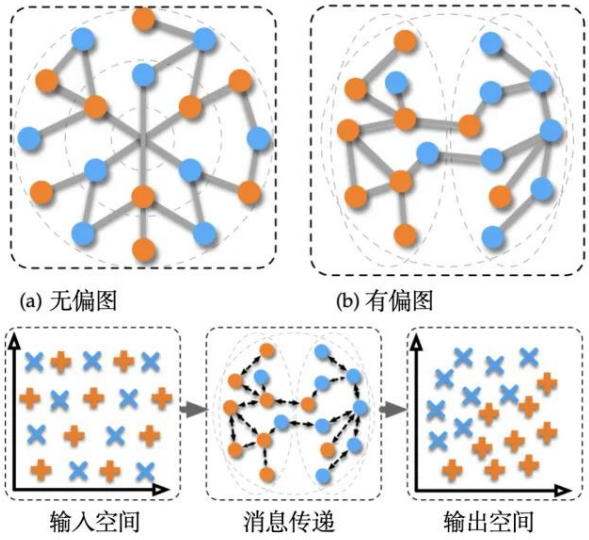

图 3.3 图神经网络增大偏见[132]

训练数据中的偏见甚至可能通过图神经网络的图拓扑结构和消息传递机制被放大，如图所示，不同颜色的节点代表具有不同敏感属性的节点，（a）中不同敏感属性的节点在图上均匀分布，是无偏图，（b）中属于同一敏感属性的节点更容易聚集在一起，是有偏图。在有偏图上经过消息传递后，特征输出空间中属于同一敏感群体的节点的特征聚在一起，不同敏感群体的节点的特征更加区分开，所以模型可以根据某个节点的特征判断该节点属于哪一个敏感群体，从而利用这一信息做出歧视性的预测。因此，确保图神经网络不对用户表现出歧视至关重要。为此，最近涌现了许多研究，旨在开发公平的图神经网络，以实现不同任务上的各种公平性。主要研究方法如下：

对抗去偏：对抗学习最初被用于处理独立同分布数据的公平机器学习模型中，目的是消除偏见。这种方法已被扩展到图结构数据。在对抗性去偏的过程中，使用一个对抗模型来预测编码器生成的表示中的敏感属性。编码器的目标是生成能够欺骗对抗模型并且能够准确预测任务结果的表示。通过这种方式，最终的表征将不包含任何敏感信息，从而确保预测结果与敏感属性无关。

添加公平性约束：除了对抗性去偏之外，直接在机器学习模型的目标函数中添加公平性约束也是一个常用的方法。这些约束通常基于公平性定义。这些公平性约束作为正则化项存在，目的是在保持预测性能的同时，平衡模型的公平性。这样的目标函数结构有助于在不牺牲模型实用性的前提下，实现预测的公平性标准。

3.2.3.3 图神经网络的分布外泛化

尽管图神经网络取得了显著的成功，现有文献普遍假设测试和训练图数据来自相同分布，即分布内假设。然而，在现实世界中，这种假设很难得到满足，测试与训练图之间的分布偏移不可避免，这些经典的图神经网络缺乏分布外泛化能力，在分布偏移下性能显著下降。因此，开发能够在图上进行分布外泛化的方法显得尤为重要，特别是对于高风险的图应用，例如分子预测、金融分析、刑事司法、自动驾驶、粒子物理学、疫情的流行预测、医疗检测，以及药物重定位等。

分布外泛化算法旨在未知分布偏移下实现令人满意的泛化性能。由于越来越多的处理实际场景中未见过的数据的需求，图上的分布外泛化自然成为一个有前景的研究方向，以促进图机器学习模型在现实世界场景中的部署。图分布外泛化的主要研究方法如下：

图数据增强技术：图数据增强技术依赖于训练数据的多样性和质量，以提高图模型的泛化性能。通过适当的图增强技术，可以简单地获得更多的图实例进行训练。图数据增强的方法通常归纳为三种策略：结构增强、特征增强以及混合类型增强。结构增强涉及修改图的拓扑结构，例如添加或删除节点和边；特征增强则是修改节点或边的特征；混合类型增强同时结合结构和特征的修改。这些增强方法旨在通过增加训练数据的代表性和丰富性，提高模型在未见过的数据分布上的表现。

特定图模型设计：除了通过增强输入图数据以实现良好的分布外泛化外，还有一些研究专门设计新的图模型，引入一些先验知识到模型设计中，使得图模型具有改善分布外泛化的图表征的能力。在这一类方法中，两种流行的技术是基于解耦的图模型和基于因果关系的图模型。基于解耦的图模型通过分离表征中的相关因素来提高泛化能力；而基于因果关系的图模型则利用因果推断原理来设计图结构，从而使模型能够更好地理解和适应数据分布的变化。这些技术通过在模型设计阶段引入结构化的知识，助力模型在面对实际应用中数据分布变化时，依然能保持较好的预测性能。

3.3 图基础模型

近年来，图神经网络和大型语言模型的融合引起了广泛的关注。图大模型旨在处理大规模的图数据，为复杂的图推理任务提供强大的工具。然而，由于图数据的复杂性和非结构化特点，构建高效、可扩展的图大模型面临诸多挑战。首先，大规模图的存储和计算需求巨大。在模型训练和推理过程中，计算复杂度高，容易导致内存和时间成本过高。这对硬件资源和算法效率提出了更高的要求。其次，在处理不同类型的图数据时，模型需要具备良好的泛化能力，能够适应不同规模和结构的图，同时保持高效的性能。这对于模型的架构设计和训练方法都是一大挑战。此外，相较于自然语言处理领域，图数据集的规模和多样性较为有限，缺乏统一的评估基准。这使得模型性能的客观评估和比较变得困难，阻碍了领域的进一步发展。

3.3.1 图基础模型概念

图基础模型的具体定义是指在广泛的图数据上进行预训练并能够适应多种下游图任务的模型[97]。图基础模型应具有以下四方面的核心能力[57]：

1、缩放法则：模型性能随着参数规模、数据集规模和训练计算量的增长而持续改进，预期大图模型也应展现出小规模或中等规模图学习模型所不具备的新能力。

2、同质泛化能力：具备同质泛化能力的预训练的大型图模型，能统一处理不同领域的图数据和任务。模型需理解图的内在结构，拥有图的常识知识。图基础模型应理解图上下文（节点、边、子图和全图），无需过多修改。此能力与少样本/零样本学习、多任务学习和分布外泛化相关，使模型利用预训练知识快速适应新数据。

3、多任务适应性：图数据中的任务类型多样化，主要可以分为三大类：节点级任务、边级任务和图级任务，每一类任务都涉及广泛的应用领域。每类任务在数据结构、目标函数以及优化方式上都有显著差异。能够有效处理并统一不同任务是图基础模型真正同质泛化和普适化能力的关键。

4、图推理能力：图基础模型需理解图拓扑结构，如大小、度数、节点连通性，并进行多跳推理以利用高阶信息。这能力增强决策可解释性，类似思维链，还需处理全局结构与复杂模式，如中心度和动态图演变。

3.3.2 图基础模型研究路径

虽然图基础模型有许多值得期待的能力，但目前尚未出现如 ChatGPT 一样成功的图基础模型。现有工作主要从以下几个方面来推进图基础模型的发展。

1、图数据资源：构建大规模、多样化的图数据集对于训练稳健模型至关重要。图基础模型的构建必须考虑图数据的独特特性。首先，根据不同的数学建模方法，图数据可以分为同质图和异质图。对于图基础模型来说，处理异质图的难度更大，这需要对主干网络进行特定的设计和优化。其次，现实世界中的图数据集规模可能非常庞大，处理如此大规模的图数据一直是图学习领域的挑战。对于图基础模型来说，海量且高度互联的图数据对模型的能力提出了更高要求。此外，图数据所涵盖的领域多样性也是一个显著特征。图基础模型需要能够处理跨领域的数据，并理解不同领域中图的底层语义信息。

2、图表示基础：研究如何有效地表示图结构，平衡表达能力和计算效率是图基础模型深入理解图结构本质及规律的前置基础。图嵌入、图卷积网络、图注意网络、图同构网络等技术能实现图结构的基础表示能力。社区检测、子图匹配等分层和局部表示技术能通过识别和利用图中的重复模式和结构，能够在保留关键特征的同时降低计算复杂度。稀疏化、节点抽样和图近似等图降维与压缩，确保在减少数据规模的同时尽可能保留重要信息。

3、图基础模型的开发：探索大规模图数据的架构、预训练和后处理技术，增强 LLM 的图理解和推理能力。指令微调和提示策略有望弥合文本模型与图推理任务间的差距，通过指令微调将图领域知识融入 LLM，提升图任务表现，为结合 LLM 和图推理提供新途径[58]。代表性的图基础模型开发技术包括提示学习（prompting）、高效参数微调（parameter-efficient fine-tuning）、模型对齐（alignment）和模型压缩（model compression）等。下面简要总结用于图模型的适配技术[57]。

4、基准和标准：NLGraph 是一个用于评估语言模型在纯自然语言描述下解决基于图的问题的基准。该基准包含 29,370 个问题，涵盖了八个不同复杂度的图推理任务，例如最短路径寻找、连通性检查和图同构[58]。像 NLGraph 这样的基准的引入对于评估进展和确定改进领域至关重要。标准化的数据集和评估指标使社区能够在不同模型和方法之间进行有意义的比较。

3.3.3 图基础模型发展方向

3.5.3.1 技术发展方向

未来图大模型的研究可在以下几个方面展开：

1、跨学科融合：结合自然语言处理、图论和机器学习等领域的优势，开发更全面的模型，促进知识的交叉融合，构建具有强大图推理能力的模型，创建能够理解复杂图结构和语言指令的模型。

2、丰富图数据集：构建大规模、多样化的图数据集，涵盖不同领域和应用场景，为模型训练提供坚实的数据基础。同时，建立标准化的评估基准，促进模型性能的客观比较。

3、模型架构创新：设计适合处理非欧几里得结构的高效神经网络架构，使模型适用于不同类型和结构的图数据，同时充分利用 LLM 的上下文理解能力。创新的模型架构将提高模型的性能和可扩展性，开发能够处理大型图的高效算法和架构。

4、优化计算效率：开发新的算法和技术，降低大规模图模型的计算和存储成本，提高模型的实际应用价值。这包括分布式计算和高效的数据处理方法。

5、应用拓展：将图大模型应用于社交网络分析、生物信息学、知识图谱等复杂领域，验证模型的实用性和有效性。真实世界的应用将推动模型的进一步改进。

6、模型可解释性与安全性：加强对图大模型的可解释性研究，确保模型决策的透明度。同时，关注数据隐私和模型安全问题，确保模型的可靠性和可信度。

3.5.3.2 未来应用方向

与语言基础模型在文本翻译、生成等任务中取得的显著成就相比，图基础模型在图任务中的影响尚不确定。然而，在图神经网络已经展现出有效性的领域，如电子商务和金融，将图基础模

型与大语言模型相结合，可能在开放性任务中进一步提升性能。特别是在新兴领域，如药物研发方面，图基础模型展现出了巨大的潜力。

在药物开发这一复杂且昂贵的过程中，语言模型已经在诸如靶点识别、副作用预测等任务中提供了重要的帮助。然而，由于蛋白质等生物分子具有复杂的三维结构，基于文本的数据并不足以充分表达其特性。图基础模型通过对图结构信息进行建模，可以更好地捕捉蛋白质分子的结构和相互作用，有望对药物发现过程带来革命性变化，极大加速新药研发进程。

此外，在城市计算领域，传统的交通预测往往关注孤立的任务，而忽略了整个交通系统的综合性。通过将交通系统视为时空图，图基础模型能够为交通系统中各参与者的行为提供更全面的理解。借助图基础模型，研究者能够在分析不同交通节点、路线、参与者行为的基础上，提出统一的解决方案，以应对各种城市计算中的挑战。例如，在复杂的交通网络中，不同的路段、信号灯、交通工具等都可以被视为节点和边，通过图基础模型的分析，可以优化整个系统的运作，从而提升交通管理的效率和预测准确性[44]。

总的来说，虽然图基础模型在许多任务上的潜力尚需进一步验证，但在一些特定领域，尤其是结合语言模型时，图基础模型有望带来显著的性能提升，特别是在那些需要对结构化信息进行深入理解的任务中，例如药物开发和城市计算。

3.4 知识图谱工程

知识图谱利用三元组描述事物之间的复杂关系。从图的技术角度来看，大量三元组构成的知识图谱可以看作是一个有标签的有向图，图技术如图神经网络、图表示学习等在知识图谱中有大量的应用。从人工智能的角度来看，知识图谱中包含图结构数据、文本数据、逻辑规则等，涉及多样的人工智能技术应用，是典型的图与人工智能融合的研究领域。本小节将从知识表示、知识抽取、知识补全、和知识服务四个方面对知识图谱工程展开介绍。

3.4.1 知识表示

知识图谱作为符号化的知识表示体系，具备高阶语义、结构严谨、复杂推理等能力。在大语言模型（LLM）飞速发展的时代，知识图谱与 LLM 之间有丰富的互动关系，一方面 LLM 为低成本构建大规模知识图谱提供了有力工具；另一方面知识图谱的高质量、可解释的知识表示和推理能力，也为解决 LLM 的幻觉问题提供了新的方向。

传统知识语义框架，如 RDF、OWL 及 LPG 等在知识管理方面有显著不足，很难支撑 LLM 时代的知识图谱构建与应用。大模型时代的知识图谱，可以从 DIKW 层次范式出发，提供从数据（Data）、信息（Information）、知识（Knowledge）的完整表示能力，以实现信息完备性、知识精准性、逻辑严谨性的有机统一。

3.4.1.1 知识分层

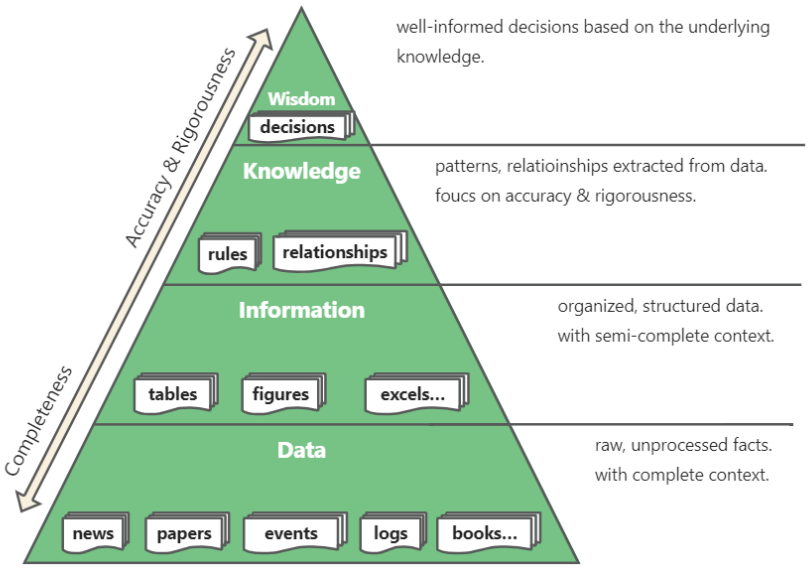

图 3.4 DIKW 知识分层

在 DIKW 知识分层中，从低到高依次是数据（Data）、信息（Information）、知识（Knowledge）、智慧（Wisdom）。

数据（Data）表示原始未处理过的数据，比如新闻、文章、事件、日志、数据等。信息（Informatioin）表示良好组织的结构化数据，比如关系数据库、表格、图表等。知识（Knowledge）是指从信息中总结出的规律、模式、关系，聚焦在知识在精准性与逻辑严密性。智慧（Wisdom）源自基于知识的推理决策，以及由此触发的行动规划。

在 DIKW 金字塔结构中，越往下，上下文信息越完整，但是知识的精准性与逻辑性越差；对应的，越往上，知识的精准性与逻辑性越强，但上下文信息缺失越严重。

3.4.1.2 知识分类

按主体类别粒度，知识可以划分为概念类型、实体类型、事件类型、标准类型、关系类型等。

主体分类模型的简要解释如下：

- **实体**：业务相关性比较强的客观对象，多属性、多关系刻画的多元复合结构类型，如用户、企业、商户等。考虑到对于 DIKW Data 层原始数据存储的诉求，Data 中的文件、文件中的段落应划分到实体类型的范畴。
- **概念**：实体从具体到一般的抽象，表述的是一组实体实例或事件实例的集合，是一种分类体系。相对静态，也是常识知识，具有较强复用性，如人群标签、事件分类、行政区划分类等。为简化企业应用，标准类型可划分到常识概念中。

- **事件**：加入时间、空间等约束的时空多元类型，如通过 NLP、CV 等抽取出来的行业事件、企业事件、诊疗事件或因购买、核销、注册等行为产生的用户行为事件。
- **属性**：属性是实体、事件、概念等的组成要素，用以表述一个复杂结构的各个独立要素，每个属性要素又会关联为一个具体的简单或复杂结构，如基础类型、标准类型、概念类型等。
- **关系**：关系的定义和属性基本一致，表达同一个复杂对象与其他对象之间的关联，关系和属性的区别是，若关联对象为实体类型则为关系。

3.4.1.3 逻辑规则

- 除实体、概念、事件、属性、关系外，业务专家基于特定业务场景总结的各种规则、模式、触发条件（如保险理赔规则、疾病诊断规则等），也属于知识的一种，逻辑规则可以采用三段式语法表示，例如其语法结构可以定义为：

```
#Structure：定义匹配的子图结构。
Structure {
    // path desciption
}
#Constraint：定义上述Struct中，对实体和关系的约束条件、以及规则计算的表达式。
Constraint {
    // rule express
}
#Action：指定了对符合Structure和Constraint的结果进行的后置处理。
Action {
    // action desciption
}
```

- 定义新的逻辑规则的语法结构，如下：

```
#Define用于定义新的逻辑谓词。它允许您创建符合特殊Structure和Constraint限制的自定义谓词。
Define (s:sType)-[p:pType]->(o:oType) {
    Structure {
        // path desciption
    }
        Constraint {
        // rule express
    }
}
```

逻辑规则语法结构中，包含 Structure、Constraint、Action、Define 等模块。

- **Structure**

路径的基本单元是边，多种边组合起来的连通图成为路径，Structure 中可以描述多个路径，方便在不同场景下使用。路径描述按照 ISO GQL 方式进行描述：

```
Structure {
    (s:User)-[p:own]->(o:Shop)
}
Structure {
    (s:User)-[p:own]->(o:Shop), (s)-[c:consume]->(o)
}
```

- **Constraint**

Constraint 中支持单规则语法、规则组语法、聚合语法。单规则语法中，Constraint 中每一行作为一个规则，包括逻辑规则、计算规则、赋值规则等。

```
1、逻辑规则
以 规则英文名("规则说明"): 表达式 这种格式进行表达，输出为布尔值。常用运算符有>、<、==、>=、<=、!=、+、-、*、/、%等，运算符可以进行扩展。
2、计算规则
以 规则英文名("规则说明")= 表达式 进行表达，输出结果为数字或者文本，取决于表达式内容。
3、赋值规则
以 别名.属性名= 表达式 没有规则名，仅允许Define中定义的别名进行属性赋值表达。此类规则仅在特定谓词的规则定义中有效。
```

规则组可以将逻辑规则进行组合，主要目的是将逻辑计算层次化，例如：

```
Structure {
     （s:User）
}
Constraint {
    R1("成年人"): s.age > 18

    R2("男性"): s.gender == "男"

    // 下面这句是正确的，R3由R1和R2组合而成，就被视为是一条规则组
    R3("成年男性"): R1 and R2

    // 下面这句是错误的，规则组里不允许有非规则的变量
    R3("成年男性"): R1 and s.gender == "男"
}
```

聚合语法指的是对 groupby、sum、avg 等聚合算子的支持。

- **Action**

通常 Action 中支持多种操作：

- createNodeInstance/createEdgeInstance：用于因果的逻辑结果的语义表达
- get ：输出匹配的结果，包括实体、关系以及属性等内容。

3.4.1.4 互索引结构

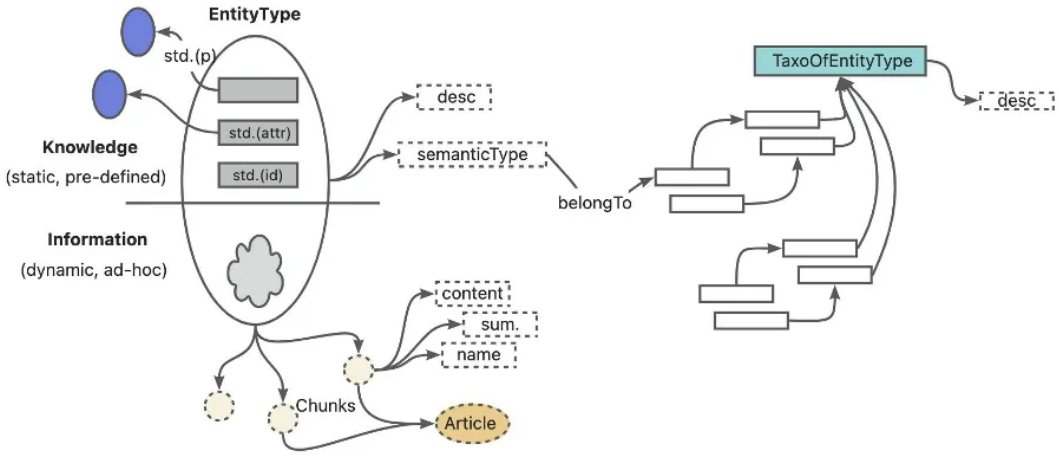

图 3.5 文本和图结构互索引的知识表示[150]

图谱索引是一种基于图谱的文档脉络索引，包含 Chunk 段落、具体业务实体、通用概念知识三部分内容。这样一种图和文本混合的互索引结构，使得既可以在图上进行遍历，也可以检索文本块，并进行有效的分析。如上图左侧所示，通过图结构可以更有效地组织文档间的关联。

根据文档的篇章结构，段落间内在的逻辑关联，实现 Chunk 段落的语义分块。语义分块的结果兼具长度限制和语义约束，以适配大模型对窗口长度的限制，并实现同一分块内的内容高度内聚的要求。文档语义切分产出的分块，包含 id、摘要、正文等信息；id 由文档 id、篇章结构和顺序编码组成，文档中相邻的内容其 id 也是连续的。同时，文档与切分出的语义分块之间，也是互相关联的。

业务实体、实体间关系抽取自 Chunk 段落，通过实体消歧、实体归一、实体融合、概念图挂载、语义构图等图谱技术栈，实现从歧义化、模糊化、碎片化的信息到明确化、标准化、网络化的知识的转变。实体中包含 knowledge、information 两部分信息。knowledge 指由业务专家预定义的，高频、常见的知识，对应的特征为强 schema 约束、结构化数据、属性和关系标准化；information 指由大模型开放抽取得到的动态知识，特征为弱 schema 约束、非结构化数据（文本和向量等）。

实体类型包括预定义类型 EntityType 和 semanticType；EntityType 属于高层级的分类，包括 Person、Organization、GeoLocaltion、Date、Creature、Works、Keyword 等；semanticType 属于低层级的分类，比如 Person 类别下的 biochemist, musician 等。高层级的分类，目的是知识存储的便利化；低层级的分类，目的是业务应用的精准性。

概念图作为领域专家知识内嵌到图谱系统中，基于实体的 semanticType、desc，通过概念挂载实现实例图与概念图的融合。

3.4.2 知识抽取

知识抽取是从非结构化或半结构化数据中识别、提取和组织有价值的信息和知识的过程。其目标是将各种形式的原始数据（如文本、图像、音频、视频）转化为结构化的数据，以便于计算机系统理解、分析和利用。知识抽取的方法经历了多个发展阶段：从早期依赖规则和模板的方法，到后来基于统计机器学习的技术，再到深度学习方法的应用，最终发展到如今使用预训练模型的方法。

3.4.2.1 知识抽取任务

知识图谱的构建和维护涉及多个知识抽取任务，其中实体抽取、关系抽取和事件抽取是最核心和直接相关的任务。

3.4.2.1.1 实体抽取

实体抽取，也称为命名实体识别，用于识别数据源中的命名实体（包括人名、地名、组织名等），这些实体通常作为知识图谱中的节点，是知识图谱中最基本的元素。例如，“2010 年 9 月 24 日，马青骅代表北京现代车队参加中国房车锦标赛，获得鄂尔多斯站冠军”中的信息可以通过其包含的时间实体“2010 年 9 月 24 日”，人员实体“马青骅”，参赛队伍实体“北京现代车队”，赛事类型实体“中国房车锦标赛”，地点实体“鄂尔多斯”和荣誉类型实体“冠军”来直接表达。知识图谱的质量与实体抽取的完整性、准确率和召回率息息相关。早期的实体抽取方法包括依赖规则和模板的方法以及利用统计机器学习的方法。基于规则和模板的方法依赖于预定义的规则和模板，当所选用的规则能够很好地反映文本信息时，通常效果不错。例如，定义规则人名是两个连续的首字母大写的单词，然后将符合规则的文本字符抽取为实体；基于统计机器学习的方法的核心想法是从标注好的数据来中学习和推断规律，以进行实体抽取。近年来，随着深度学习方法在自然语言处理、计算机视觉等领域取得显著的突破，深度学习方法成为了实体抽取的主流方法。用于实体抽取任务的深度学习模型涵盖了多种架构，包括卷积神经网络（Convolutional Neural Network, CNN）、循环神经网络（Recurrent Neural Network, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM）、基于 Transformer 的预训练模型和图神经网络（Graph Neural Network, GNN）。 CNN 通过一系列卷积和池化操作，能够有效地提取文本中的局部特征，随后通过全连接层进行实体识别和分类；RNN 逐个处理文本中的每个词，利用其循环结构保留并处理词与词之间的时间依赖信息，从而实现命名实体识别；GRU 利用门控机制调节信息流动，能够捕获文本中长距离依赖关系，逐词处理文本以实现命名实体识别；基于 Transformer 的方法采用多头自注意力机制，可以并行处理序列中的所有词，并直接在编码器中获取上下文信息；GNN 将文本转化为图，通过迭代更新节点向量来聚合上下文信息[2]。

3.4.2.1.2 关系抽取

通过实体抽取获取的实体之间是离散且无关联的。关系抽取用于识别实体之间的关系并建立起实体之间的语义链接。这些关系通常作为知识图谱中的边。例如，在句子“ChatGPT 是由 OpenAI 开发的一种大语言模型”中，关系抽取任务会识别出（OpenAI，开发，ChatGPT），（ChatGPT，是，大语言模型）这样形式的三元组关系，从而构建知识图谱。 早期关系抽取的方法包括基于传统规则和模板的方法和基于传统机器学习的方法。基于传统规则和模板的方法依赖于手写规则和模板，通过使用触发词和依存关系来匹配文本。基于传统机器学习的方法通过特征工程从文本中提取语法、词法等信息，构造特征向量，然后使用分类器来识别实体对之间的语义关系。近年来，深度学习方法成为了关系抽取的主流方法。关系抽取可以通过各种流行的神经网络架构来实现。CNN 和 RNN 是较早用于关系抽取的深度学习方法，总体而言，CNN 擅长捕捉句子中的局部特征，RNN 设计用于处理序列数据，使其比 CNN 更适合捕捉文本中的长距离依赖关系。基于注意力机制的神经网络增强了关系表示与文本表示之间的相关性，突出了关系抽取的重要信息。注意力机制允许模型在预测关系时关注文本的相关部分，有效地克服了 CNN 和 RNN 在处理长距离依赖关系方面的局限性。它们可以捕捉复杂的句子结构和实体之间的关系，无论它们在文本中的位置如何。GNN 通过构建语义图来尝试捕捉输入序列的非线性结构，使关系抽取模型具有图上的关系推理能力。GNN 可以捕捉实体和关系的相互关联性，这对于纯粹的序列模型来说是困难的。预训练语言模型通过在大规模未标注文本数据上进行训练，学习到文本中所包含的语法和语义知识。随后，通过对预训练模型进行微调，可以直接用来进行关系抽取等下游子任务[138]。

3.4.2.1.3 事件抽取

事件抽取旨在识别和抽取样本源中的事件及其相关信息，事件可以看作是知识图谱中的特定子图。事件抽取不仅涉及识别事件本身，还包括确定事件的触发词、分类事件类型、识别事件的论元以及确定论元在事件中的角色。在例句“特朗普于 2017 年 1 月 20 日在美国国会大厦宣誓就职”中，事件抽取任务具体为检测触发词“就职”，判断事件类型为“任职”，确定“特朗普”“2017 年 1 月 20 日”和“美国国会大厦”为事件论元，并确定它们对应的角色分别为“人物”“时间”和“地点”。 事件抽取技术经历了从基于模式匹配方法到现代深度学习方法的演变。早期的方法依赖于专家知识和预定义的模板，通过模式匹配来识别事件。随着数据和计算能力的提升，机器学习得到了发展。这些方法基于特征来构建分类器，从而进行事件类型和论元的分类。然而，这些传统方法在捕捉深层语义特征方面存在局限。深度学习的兴起显著提升了事件抽取的效果。RNN 用于建模序列信息以提取事件中的论元。JRNN[135]提出了一种基于双向 RNN 的联合事件抽取模型。该模型包括使用 RNN 总结上下文信息的编码阶段以及利用编码信息预测触发词和论元角色的预测阶段。JMEE[136]采用层次注意力机制来实现信息的全局聚合，JMEE 主要由四个模块组成，分别是词表示模块、句法图卷积网络模块、自注意力触发词分类模块和论元分类模块，该模型利用基于注意力的图卷积网络进行联合建模图信息，以提取多个事件触发词和论元。GAIL[137]是一

种使用生成对抗网络（Generative Adversarial Network,GAN）帮助模型关注难以检测的事件的模型。预训练语言模型的出现，为事件抽取带来了新的突破。在 BERT 模型出现之前，主流方法是从文本中识别出触发词，然后根据这些触发词来判断事件类型。随着 BERT 被引入到事件抽取模型中，基于全文识别事件类型的方法逐渐成为主流。这是因为 BERT 在上下文表示能力上表现出色，在文本分类任务中表现良好，尤其是在数据量较少的情况下[134]。

3.4.2.1.4 其他抽取任务

属性抽取用于识别实体或者关系的属性及其值，这些属性丰富了知识图谱中实体和关系的描述。属性可以看作属性值和实体或者关系之间的一种关系，因而可以通过关系抽取的思路来解决。

三元组抽取可以视为一种综合性的知识抽取任务，它包含了实体抽取、关系抽取和属性抽取的内容。具体来说，三元组抽取的目标是从源样本中抽取形如（subject, predicate, object）三元组，这些三元组可以同时包含实体、关系和属性信息，这些三元组可以直接用于构建知识图谱。

另外，知识抽取还包括观点抽取、关键词抽取、主题抽取和情感抽取等，这些虽然不是直接用于构建知识图谱的核心元素，但它们可以丰富和增强知识图谱的内容和功能。通过结合这些抽取任务，可以构建一个更加全面和智能的知识图谱。

3.4.2.2 知识抽取流程

3.4.2.2.1 本体建模

无论是开放域的知识图谱还是包括专业领域的各行业的知识图谱，都需要收集大量的数据，这些数据的收集是有选择性的，这个选择的依据就是本体模型，也称 Schema 设计或本体设计。本体建模解决知识图谱如何组织数据的问题，是数据的底层架构，是一个知识体系框架，能够涵盖住知识图谱所有的数据，决定了数据收集的范围。

本体模型作为知识表达模型，定义了实体类型、实体对应的属性、以及实体和实体之间的关系，通常应根据实际应用需求和数据情况以及业务知识来综合设计。下图为一个装备维保知识图谱的本体模型示例。

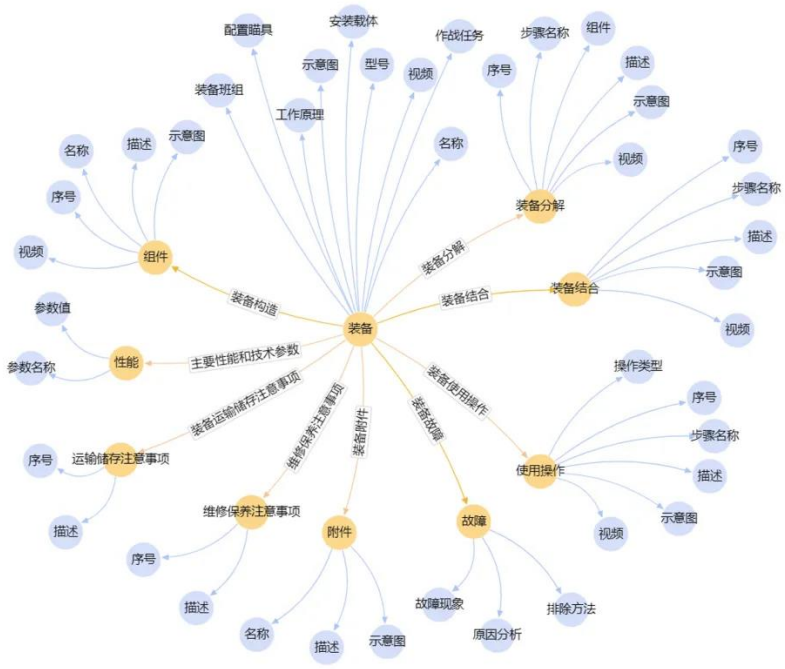

图 3.6 装备维保知识图谱的本体模型示例

3.4.2.2.2 知识抽取

知识抽取是针对结构化数据、非结构化数据，利用大数据、深度学习、机器学习、自然语言处理等技术，将数据转化为 RDF 三元组数据，并统一存储的过程。基本流程如下图所示。

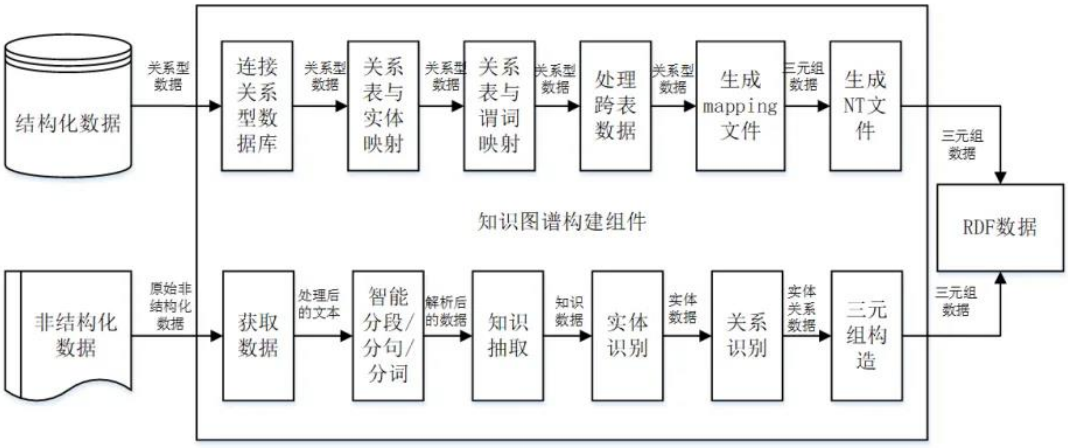

图 3.7 三元组抽取流程

1）结构化数据抽取

用于构建知识图谱的原始数据可能是结构化数据或者非结构化数据。结构化数据通常存储于关系型数据库或 excel 二维表中，有明确的字段定义，数据格式非常规范，通过字段与知识图谱实体、属性、关系的映射，即可自动进行三元组的抽取。

原始数据大多存储在像 MySQL 这样的关系数据库中，并以不同的表格形式区分，而用于构建知识图谱的数据通常以三元组格式存储，因此需要进行这种转换。D2RQ[1]是一个用于将关系数据库内容转换为 RDF 三元组的工具。D2RQ 主要包括 D2R Server、D2RQ Engine 和 D2RQ Mapping 语言。

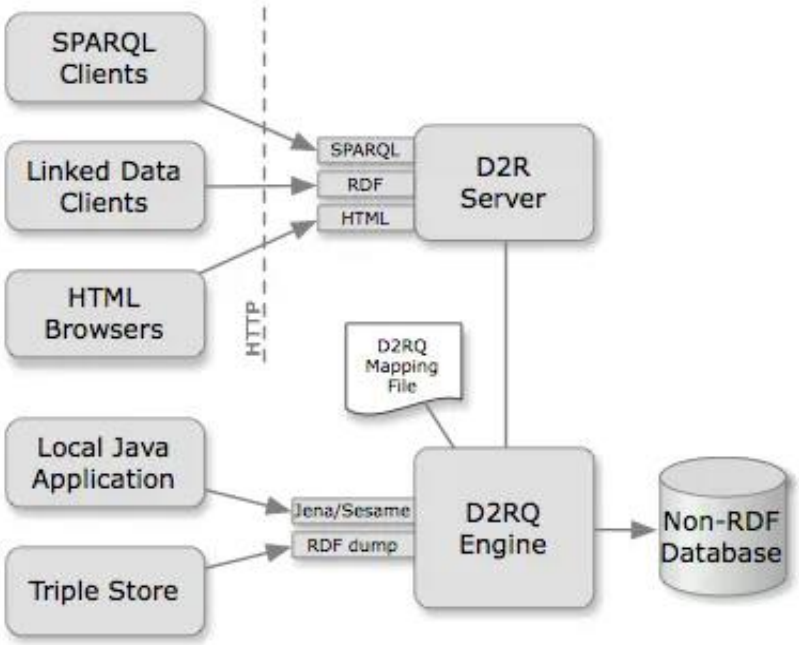

图 3.8 基于 D2RQ 的结构化数据知识抽取框架[134]

2）非结构化数据抽取

非结构化数据多为篇章级的 PDF、Word 数据，不同种类文档的结构具有一定的规律，根据行文方式规范程度，采用相关抽取工具，通过“规则+机器学习”相结合的方式进行知识抽取，主要分为下面几个步骤：

1. **数据获取**：通过知识图谱自动化构建平台，与存储的文档的数据库进行连接，从而获取文档，同时也可在平台上上传当前的 PDF 和 Word 文档。

2. **智能分段**：首先将篇章级的文档进行分段，可根据分段标识来进行分段处理。

3. **智能分句**：主要是对分段后的文档进行分句，采用中文依存句法分析工具，分析句子中词与词之间的依存关系（如主谓关系指主语与谓语间的关系），并根据依存关系以及标点符号进行自动切分。切分后，还要再判断句子中是否存在并列关系或连谓结构，这样的句子一般是在同一事项当中，所以再对相应句子进行合并。

[1] http://d2rq.org

4. **中文分词**：应用中文分词工具进行分词，一方面可以实现中文分词（包括停用词），另一方面可以对词进行词性和语义标注。在实践中有时会将句子中的词分的十分细碎，可以再进行词与词之间的结合，如紧邻的名词，名词间存在代词的情况。这样可以更准确的提取主语。

5. **知识抽取**：最后知识图谱三元组构建，可以采用相关构建工具进行知识抽取。例如应用北京大学 gBuilder[2]工具，首先通过整体抽取流程的流水线构建，然后再进行实体抽取、关系抽取和三元组构建，从而将数据转化为知识。gBuilder 中内置了众多非结构化抽取算法和模型，可通过构建非结构化数据抽取流水线来进行数据的抽取，将数据转化为知识。

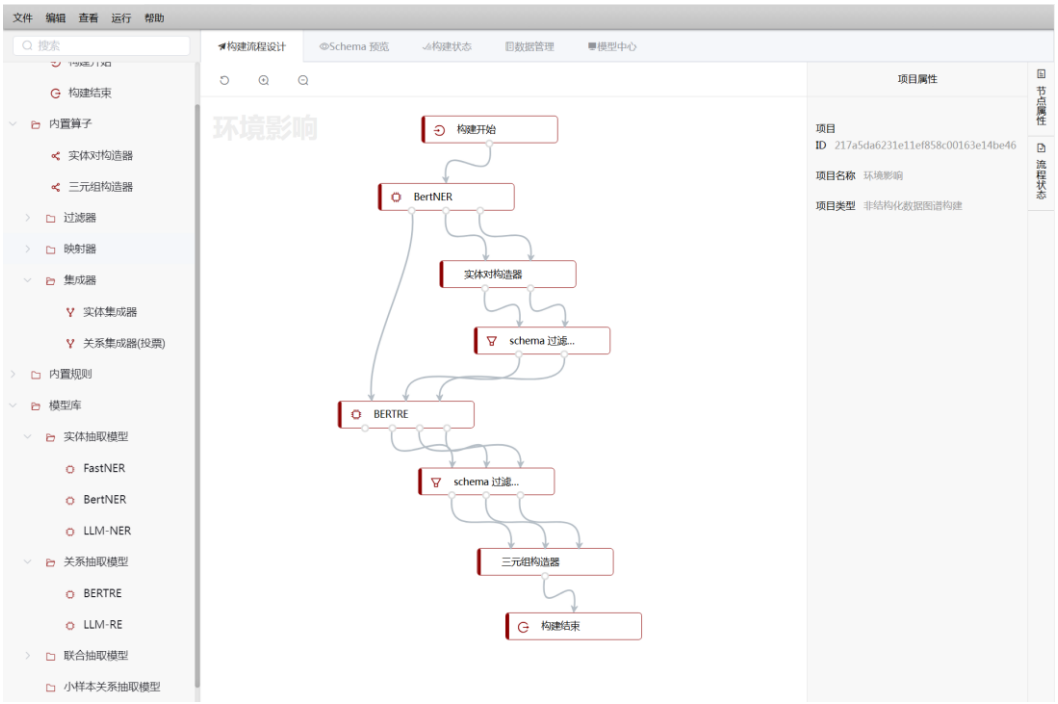

图 3.9 gBuilder 非结构化数据知识抽取流水线设计

在上述过程中，最重要的也是最难的是实体和关系的抽取。在非结构化数据抽取过程中，传统的方法如 Bert 等模型，需要较多的语料标注，才能够达到可接受的抽取效果。大语言模型（Large Language Model,LLM）兴起之后，因其强大的一般知识、文本理解、泛化能力，可实现小样本的非结构化文本数据的知识抽取，通常标注少量的数据，就可以在实体、关系、属性抽取的准确率方面得到显著提升。

3.4.3 知识补全

知识图谱（Knowledge Graph,KG）作为一种重要的数据组织形式，以图形化的方式展示实体及其关系，广泛应用于搜索引擎优化、推荐系统、智能问答等领域。它在这些应用中发挥了关键作

[2] https://www.gstore.cn.

用，但实际应用中的知识图谱常常存在信息不完整的问题。这些缺失的信息可能影响系统的智能化水平和用户体验。

知识图谱补全（Knowledge Graph Completion,KGC）技术旨在通过推测和填补这些缺失的知识，从而提升知识图谱的全面性和准确性。这项技术不仅增强了知识图谱的实用性，还提升了信息检索和智能决策的能力，使其能够更好地服务于实际应用。KGC 技术通过技术手段预测和填补知识图谱中缺失的信息，包括未记录的实体、关系和属性，提高其整体质量和应用效果。

3.4.3.1 主要任务

在知识图谱补全中，三元组预测、链接预测和关系预测是三大核心任务，每一个任务都在知识图谱的构建与优化中起着至关重要的作用。随着技术的不断进步，尤其是人工智能技术的发展，这些任务在处理大规模数据和复杂关系时表现出显著的优势。

3.4.3.1.1 三元组预测

三元组预测的核心任务是识别并填补知识图谱中缺失的“实体-关系-实体”三元组。一个典型的三元组由头实体、关系和尾实体构成。例如，在缺少“苹果公司-总部位于-库比蒂诺”的场景下，三元组预测技术通过分析现有数据，推测并补全该缺失三元组，从而完善知识图谱。

早期的三元组预测依赖逻辑规则和路径搜索技术，通过规则推理和沿已知关系进行路径扩展，推测出缺失的三元组信息。然而，随着知识图谱规模扩大，关系复杂性增加，传统方法在泛化能力和灵活性方面逐渐暴露出局限性，难以应对复杂的知识场景。

为应对这些挑战，基于嵌入（embedding）的模型应运而生，并逐渐成为主流方法。TransE、DistMult 等嵌入模型通过将实体与关系映射到低维向量空间，捕捉它们之间的潜在联系，从而高效地进行三元组预测。这种向量化处理方式有效简化了知识图谱的复杂结构。

近年来，图神经网络（Graph Neural Networks, GNN）的发展推动了三元组预测技术的突破。GNN 通过迭代聚合节点及其邻居的信息，能够更精细地捕捉知识图谱的复杂结构。与传统嵌入方法相比，GNN 在处理异构数据和上下文信息时具有明显优势。代表性模型如 R-GCN（Relational Graph Convolutional Network）和 CompGCN（Composition-based Graph Convolutional Network）已经在大规模知识图谱补全任务中展现了卓越的性能。

3.4.3.1.2 链接预测

链接预测的任务是预测两个已知实体之间可能存在的关系，着重于发现实体间的潜在联系，而非具体的三元组。例如，针对知识图谱中“乔布斯”与“苹果公司”之间的缺失关系，链接预测技术将尝试推测两者间可能的联系，如“创始人”。

最初，链接预测主要依赖共现统计和矩阵分解方法，通过计算实体间的相似度或利用结构属性，推测潜在的关系。然而，随着知识图谱复杂性的增加，传统方法在应对远距离关系和复杂结构时显得力不从心。

随着技术的进步，基于图嵌入的技术，如 LINE、DeepWalk 等方法，逐渐成为链接预测的核心工具。这些方法通过将实体映射为低维向量，捕捉图结构中的潜在关系。然而，面对日益复杂的图结构，嵌入方法的表达能力仍有局限。

深度学习技术的发展为链接预测任务提供了新的动力。图卷积网络（Graph Convolutional Networks,GCN）等神经网络模型能够直接在图结构上操作，利用自适应特征学习，显著提升了关系预测的准确性。此外，对比学习（Contrastive Learning）等方法也逐渐在链接预测中崭露头角，通过引入负样本生成机制，进一步增强模型在复杂知识图谱中的泛化和区分能力。这类方法在处理大规模图谱时表现出极高的鲁棒性和精度。

3.4.3.1.3 关系预测

关系预测任务旨在明确两个实体之间的具体关系类型，而不仅仅是预测它们之间是否存在关系。例如，对于“比尔·盖茨”和“微软”两个实体，关系预测的目标是判断两者之间的具体关系，如“创始人”或“首席执行官”。这一任务要求不仅能识别实体间的联系，还要对关系的性质进行准确分类。

关系预测任务的早期方法主要依赖手工定义的规则和基于路径的推理技术。这类方法通过分析实体间的路径信息或共现模式，推测它们之间的潜在关系类型。然而，随着知识图谱规模的扩展和关系多样性增加，规则驱动的方法在处理复杂关系类型及未见数据时逐渐显现出其局限性。

近年来，深度学习技术，尤其是基于注意力机制（Attention Mechanism）的模型，在关系预测任务中表现出强大的潜力。Attention 机制能够根据上下文信息对不同邻居节点和关系赋予不同的权重，从而实现更精确的关系分类。同时，随着语言模型的发展，BERT 等预训练语言模型被引入到关系预测任务中，进一步提升了模型在文本和结构化数据之间的推理能力。通过将关系预测任务转化为序列预测问题，这些 AI 模型能够从大规模文本数据和知识图谱中提取更多隐含关系，实现更为精准的关系识别与分类。

3.4.3.2 关键技术与流程

在知识图谱补全的过程中，涉及多个关键技术和步骤。以下内容将分为四个主要部分：数据预处理、模型学习、候选处理和事实识别。

3.4.3.2.1 数据预处理

数据预处理是知识图谱补全的基础。此阶段主要包括以下两个关键任务：

实体对齐与融合：在处理不同数据源时，必须将相同的实体统一表示，避免信息重复。不同数据源可能使用不同的表示方式，如不同的命名或标识符。通过实体对齐与融合技术，可以确保这些不同的表示都指向同一个实体，从而避免冗余和冲突。例如，对于“Facebook”这个实体，不同的数据源可能使用“FB”或“Meta”来表示，通过对齐与融合，将这些不同名称统一为一个实体。

知识去重与合并：此步骤的目的是清除重复记录并整合相似的信息，以形成一个更为完整的知识记录。不同数据源可能包含关于同一实体的多条记录，通过去重和合并，可以将这些信息整合为一个完整、准确的知识条目。例如，将多个来源中关于“Google”的信息整合，去除重复条目，从而生成一个全面的知识记录。

3.4.3.2.2 模型学习

模型学习阶段在知识图谱补全过程中至关重要，主要包括以下几个步骤：

数据准备：数据准备是模型训练的基础，涉及收集和整理用于训练和验证的数据集。这些数据集包括已知的实体、关系和三元组，同时也包括负样本（即不存在的三元组）。数据的质量直接影响到模型的训练效果和最终性能，因此确保数据的准确性和全面性是关键。通过系统化的数据整理，可以确保模型训练和验证过程中的数据代表性和多样性，进而提升模型的泛化能力。

模型训练：在模型训练阶段，选择合适的模型是核心任务。常见的模型包括图神经网络（GNN）、逻辑回归等。这些模型通过处理训练数据，旨在预测知识图谱中缺失的三元组。训练过程中，需要不断优化模型参数，以提高其预测的准确性和可靠性。通过反复训练和调整，可以使模型逐渐学会识别潜在的缺失知识，并在面对新数据时做出准确的预测。

模型评估：模型评估是确保模型性能的关键步骤。在这一阶段，通过使用验证数据集对模型进行评估，主要的评估指标包括准确率、召回率和 F1 分数等。这些指标能够全面衡量模型在预测缺失三元组时的表现。根据评估结果，对模型参数进行调整，以优化模型的预测能力。有效的评估和调优过程能够显著提高模型的预测准确性，从而提升知识图谱的整体质量。

3.4.3.2.3 候选处理

候选处理阶段主要包括以下步骤，以提升预测准确性和效率：

候选集生成：这一阶段的任务是生成可能的三元组候选集。这涉及创建所有可能的三元组，其中一些可能是正确的，但尚未在知识图谱中记录。这些候选集为后续的筛选和验证提供了基础，并且可以通过算法生成潜在的三元组组合，为知识图谱补全提供广泛的可能性。

候选过滤：在候选集生成之后，需要通过设置过滤规则和应用算法来筛选出最有可能的候选三元组。这一步骤的目的是提高模型预测的准确性。通过规则和算法，可以排除那些不太可能的

候选三元组，从而确保最终选择的三元组具有较高的准确性和有效性。有效的候选过滤可以减少不必要的计算量，并提升知识图谱的补全效率。

3.4.3.2.3 事实识别

事实识别阶段是知识图谱补全的最后步骤，涉及将训练好的模型应用于候选集，识别和确认缺失的三元组。在这一阶段，被确认的三元组将被添加到知识图谱中，从而提升知识图谱的完整性和准确性。事实识别的目标是通过准确的模型应用，补全潜在的缺失信息，使知识图谱更加全面和可信。这一过程不仅增强了知识图谱的实用性，还为下游应用提供了可靠的数据支持。

3.4.4 知识服务

知识图谱构建完毕后，存储了当前应用中重要的概念、实体、属性、和关系等，这些存储了丰富信息的知识图谱可以服务于很多应用，例如在电商应用中构建了商品知识图谱后，商品知识图谱服务于电商应用中的很多任务，包括货场选品、商品分类、同款商品对齐、商品推荐、以及序列推荐等。知识服务的任务包括知识查询问答、复杂逻辑查询、检索增强问答等，知识服务方式涉及数据的存储与查询、知识图谱模糊查询检索、知识图谱预训练等。本小节将从知识服务任务和知识服务方式两个角度，对融合了主流图学习和人工智能方法的知识图谱服务展开介绍，将首先介绍知识服务涉及的一些典型任务和方法，然后介绍典型的知识服务流程。

3.4.4.1 知识服务任务

3.4.4.1.1 知识查询问答

知识图谱查询问答是指基于自然语言问答的方式完成知识图谱中的知识查询。例如针对一个电商知识图谱问询“在平台上售卖的去年下半年上市的国产手机型号有哪些？”，为了回答这个问题，需要根据问题的语义，在知识图谱中找出对应的数据。根据查询问答问题的复杂程度，可以将查询问答分为简单查询问答和复杂查询问答。知识图谱查询方法包括基于语义匹配的方法和基于检索的方法。

基于语义匹配的方法首先将查询问答问题，经过语义匹配转化为逻辑表达式，如 S-表达式，SPARQL 查询语句等。语义匹配的方法又可分为逐步生成方法和序列到序列的方法，逐步生成法将自然语言到逻辑表达式的翻译过程定义为一系列的步骤，例如首先找到问题的核心实体，然后找到以核心实体为起点，以问题答案为终点的路径，再在路径的节点上添加属性约束等。序列到序列的方法将自然语言到逻辑查询语句的映射过程看作一个语言翻译的过程，并根据标注数据训练一个翻译模型例如基于 T5 的模型，实现以自然语言问句为输入，直接生成问句对应的逻辑表达式。逻辑表达式可以被翻译为可进行知识图谱查询的 SPARQL 查询语言，并基于 SPARQL 查询语句得到问题对应的查询结果。

基于检索的方法首先基于问题在知识图谱中检索相关的子图，然后根据子图中包含的信息进行问题回答。因此基于检索的方法通常包含一个检索器和一个推理器，检索器实现的功能是根据当前的问题从知识图谱中检索和当前问题相关的包含答案的子图，推理器实现的功能是根据检索的子图信息推理出问题对应的答案，例如可以采用问题感知的（类）图神经网络模型对子图进行编码，并根据实体的表示计算当前实体作为问题答案的概率。

3.4.4.1.2 复杂逻辑查询

知识图谱复杂逻辑查询是指对知识图谱进行包含复杂逻辑组合的查询，这个任务的复杂性体现在两个方面，一方面是复杂逻辑查询任务通常包括逻辑或、且、非组合以及其他逻辑约束例如存在量词、全称量词等。例如针对一个人物知识图谱，查询有小于 1 个小孩或者有多余 3 个小孩且有一个是女孩的人居住的城市有哪些。另一方面是复杂逻辑查询任务中通常会包含一些无法查询到正确结果的查询步骤，这个是受到知识图谱本身不全的影响。

复杂逻辑查询问答可以分类两类方法，一类是查询嵌入方法，这类方法采用各种表示学习方法，将逻辑查询语句编码到既定的向量空间中，最后计算查询嵌入表示和答案表示的匹配度得到查询的结果，查询嵌入方法可以通过向量计算推理出缺失的事实的概率，并将高概率的事实考虑进查询过程中。另一类方法是，基于大语言模型的方法，这类方法通过利用大语言模型中的参数化知识弥补知识图谱不全的问题，利用大语言模型通用的逻辑推理能力对查询进行拆解，以应对复杂逻辑查询的复杂性。

3.4.4.1.3 检索增强问答

检索增强问答是指利用知识图谱作为外部知识源，辅助基于自然语言的问答。例如利用 WikiData、OneGraph[3]的数据辅助进行一些知识问答、常识问答等。以知识图谱为外部知识源的检索方法通常依赖一个检索器，这个检索器的功能是根据当前的问从知识图谱中检索有助于回答当前问题的知识。

检索器的方法有几种，一种是将知识图谱中的三元组进行序列化，通过检索器的文本编码器将每个三元组序列编码为一个向量，同时用检索器的文本编码器将问句编码为一个向量，通过向量计算得到和当前问题最相似的三元组作为外部检索的知识，将这些三元组经过线性化之后和问题拼接起来输入语言模型中生成答案。另一种方法是从问题中识别出命名实体，将命名实体和知识图谱中的实体进行对齐，以对齐的实体为起点，从知识图谱中检索这些实体的 k 跳子图，将 k 跳子图序列化之后和问题一起输入语言模型中生成答案。

[3] http://onegraph.openkg.cn/

除了以上的单步检索方法，还可以使用多步检索的方法对检索结果进行优化，使回答问题的过程和外部知识图谱反复迭代地进行交互，充分利用知识图谱中的信息辅助问答。

3.4.4.2 知识服务方式

3.4.4.2.1 数据存储与查询

构建好的知识图谱，尤其是大规模的知识图谱，通常会被存储于图数据库中，典型的图数据库有 Neo4j、TuGraph、gStore 等，这些图数据库通常支持包含亿级的节点和关系的知识图谱的存储，并提供对应的可视化查询界面和命令行查询工具，且均开源了社区版本，以便相关人员使用。

这些图数据库通常采用图查询语言进行数据查询，典型的图查询语言有 Cypher、Gremlin、SPARQL、GQL 等，其中 Cypher 是一种申明式查询语言，语法类似 SQL，主要用于 Neo4j 图数据库；Gremlin 适用于 Apache TinkerPop 框架的图数据库，是一种基于遍历的图查询语言，其查询语句可以被看作是图上的遍历过程；SPARQL 是一种查询 RDF 格式的图数据的查询语言，可应用于 Apache Jena 以及 Virtuoso 等；GQL 是 ISO（国际标准化组织）最新发布的图数据库查询语言，旨在为图数据的存储、管理和查询提供一个统一的标准，GQL 的设计不仅考虑了现有图数据库系统的特性，还借鉴了 SQL 等成熟查询语言的优点，以支持复杂的图模式匹配和路径查找等功能。

这些图数据库为检索知识图谱中的信息提供了丰富的功能，可以完成实体 k 跳子图检索、实体的在特定关系下连接的实体的检索、满足特定属性的实体检索、以及蕴含了或、且、非等操作的复杂逻辑组合检索等，为直接的知识图谱的数据使用提供了便利的方式。

3.4.4.2.2 知识图谱模糊检索

基于图数据库的知识查询方式适用于知道图数据库中存储的实体或关系的 id 或名称的情况下的查询，但在部分应用中，例如基于自然语言的问答中，需要将问题中的实体或关系名称映射到知识图谱中的实体或关系上，这个过程可以被称为基于文本的模糊检索，该问题可以抽象为给定一个实体（关系）名称，从知识图谱中找到与其语义最匹配的实体（关系）。模糊检索方法可以分为两种，一种是基于词袋的模糊检索方法，一种是基于向量的模糊检索方法。

基于词袋的模糊检索方法，通过计算两个文本段之间的相似度。典型的方法有 BM25，这是一个广泛应用于信息检索和文档排名的词袋模型算法。BM25 在计算文本相似度的过程中充分考虑了词频、逆文档频率、文本长度、文本平均长度等因素。

基于向量的模糊检索方法，将要计算相似度的两段文本进行向量化，通过计算向量相似度模拟文本的相似度。典型的文本向量编码检索方法有 SentenceBERT、DPR、ColBERT、SimCSE 等。

模糊检索方法为应用数据映射致知识图谱数据提供了可行的兜底的方法，使得任意的应用任务都可以充分利用知识图谱中的数据。

3.4.4.2.3 知识图谱预训练

基于图数据库的知识图谱存储与查询方法，为使用者提供了忠实于原始数据的知识图谱数据获取方式，但众所周知，知识图谱的数据往往存在不完整的特性，有一部分被蕴含但未被显式表示和存储的数据，这部分缺失的数据可能会导致知识图谱数据服务提供的数据不全面不准确。因此在数据存储和查询服务基础上，知识图谱预训练服务被提出。

知识图谱预训练即对大规模的知识图谱进行预训练，通过设计表示学习模型将知识图谱映射致特定的向量空间中，这样知识图谱中的每个实体和关系将获得向量空间的表示，并可以通过这些向量之间的计算获得三元组的真值，包括缺失的三元组的真值。除了提供三元组真值计算方法，知识图谱预训练方法还可以为下游任务提供向量服务，例如提供实体的向量表示，提供某个实体在某个关系下的尾实体的向量表示，提供实体是否具有某种关系的表示等，这些表示向量可以直接当作特征向量输入下游任务的模型中，以向量服务而非数据服务的方式将知识图谱中的知识被下游任务模型所利用，提升下游任务的效果，典型的知识图谱预训练方式有 PKGM[139]等。

知识图谱预训练使得知识图谱可以为下游任务提供超越于被存储的知识的服务，将知识图谱推理能力也提供给下游任务，使得下游任务可以受益于关系推理、类别推理、规则挖掘等知识推理能力。

3.4.5 总结与展望

图技术和人工智能技术的发展，尤其是大语言模型在语言理解方面的突破，为知识图谱的表示、抽取、补全和服务带来的技术的变革。首先，知识表示向着能表示更深度的语义和更广泛的语义发展；其次，知识抽取的泛化性得到进一步提升，知识抽取成本可以进一步降低，使得低成本快速构建大规模知识图谱成为可能；再者，知识补全从依赖图结构的补全向着混合依赖图结构和文本的方向发展，可以更加充分地利用知识图谱中图结构和语义信息；最后，知识图谱服务的方式多样性逐渐增加，除了检索查询这类传统服务方式，还发展出了辅助大模型思维链等方式。总的来说，以大模型为核心的人工智能技术发展，为知识图谱的构建、维护和应用带来了新的技术范式和应用场景，会进一步促进知识图谱技术的应用和发展。

3.5 图应用

3.5.1 自然语言转图查询

现代关系型数据库使用 SQL（Structured Query Language）作为查询语言，由于 SQL 语言本身复杂的特性，只有少数研发工程师和数据分析师能够熟练使用数据库。于是开发者尝试借助大模型微调（Fine Tuning）等技术将自然语言自动翻译为 SQL 语句，即 Text2SQL，来降低数据库的使用门槛。Text2SQL 这一研究领域在科研工作者不断的探索之下，已然发展十分成熟，拥有数量、

种类均十分丰富的语料数据集，以及对应的评测数据，在大模型微调这一方面，也发展出了多种技术，例如 DAIL-SQL + GPT-4 + Self-Consistency 方案已经在 Spider 测试集上达到了 86.6%的准确率。

同样的，在图数据库领域也存在相似的使用门槛过高的问题，甚至更为严峻。相比于 SQL 相对成熟的语法标准（SQL2023），图查询语言标准（ISO/GQL）尚未全面普及，目前是多种查询语法并存的状态（GQL、PGQ、Cypher、Gremlin、GSQL 等），因此更需要借助大语言模型的自然语言理解能力，降低图数据库查询语言的使用门槛，即 Text2GQL。然而，Text2GQL 这一研究方向由于发展较晚，目前仍面临着几方面的困难。首先，Text2GQL 领域并没有如同 Text2SQL 领域那样的海量数据集可供使用，甚至鲜有公开的 Text2GQL 数据集。其次，Text2GQL 领域并没有一个如同 Spider 数据集的评测标准一样工人的评测标准以及对应的评测数据。最后，由于以上数据集和评测标准的欠缺，各种大模型微调方法的效果也很难在 Text2GQL 领域得到验证。

为了提升用户通过自然语言与图数据库交互的体验，无需掌握复杂的 GQL 语法，需优化自然语言到 GQL 的转换准确性和效率。首先，利用词法分析和语义理解等语义分析技术，提取关键信息并构建语义模型。其次，结合用户的历史查询和会话背景进行上下文理解，以消除歧义。接着，应用机器学习算法训练自然语言与 GQL 的映射关系，不断优化模型参数。在生成 GQL 语句时，依据图数据库的特点实施查询优化。此外，通过收集用户反馈，评估并改进转换结果，从而持续提升用户满意度。

3.5.1.1 语料生成

众所周知，要实现模型微调，构建语料是第一步，也是最关键的一步，语料的质量和丰富度会直接决定微调模型的预测效果。但是前面提到，由于图查询语言标准的不够成熟，想要获取现有的 GQL 语料是一件很困难的事情，并且实际业务语料的丰富度更低。SQL+GQL 语法作为一项创新技术，有了“语法制导的语料生成策略”。

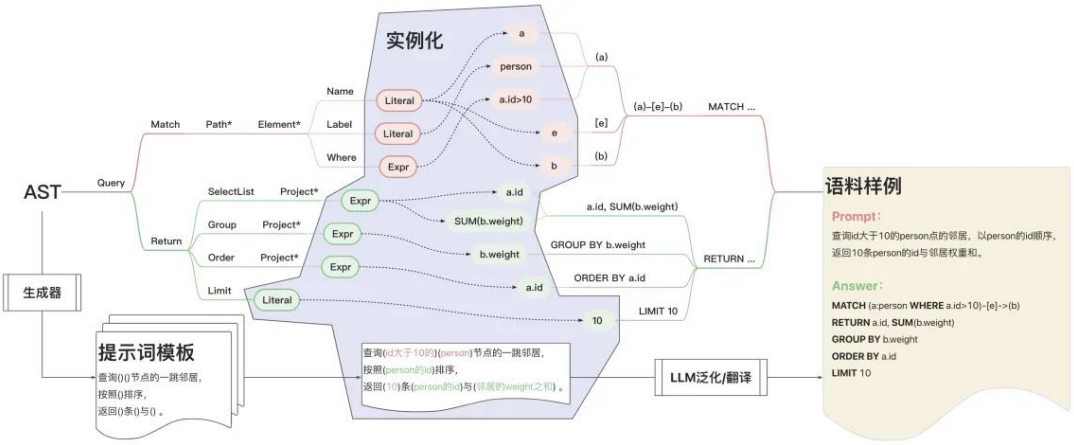

图 3.10 语法制导的语料生成策略

该策略的具体思想如下：

- GQL 抽象语法树（AST）展开后的基本形式就是表达式（Expr），常量（Literal）也是一种特殊的表达式。
- 通过设计表达式实例生成器，批量生成并组合出大量的 AST 实例，得到 GQL 语句样本。
- 特定的 AST 可以通过通用生成器产生对应的提示词模板，提示词模板随着 AST 实例化形成提示词文本。
- 特殊的不适合通过生成器生成的提示词模板可以通过人工构造。
- 初步生成的提示词文本可以借助 LLM 进一步泛化和翻译，生成多样的自然语言提示词文本。

通过该方案，能够将初始语料进行数量级的扩充，以满足后续训练的需要。具体执行流程如下：

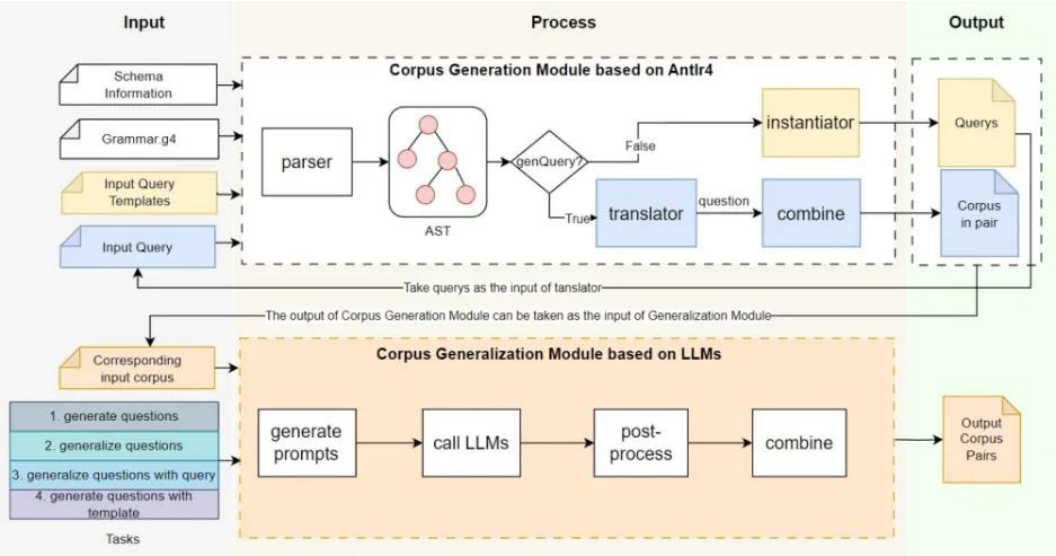

图 3.11 GQL 语料生成核心流程

3.5.1.2 模型微调

大模型中常用的微调方法同样适用于 Text2GQL 任务，如 LoRA 与 QLoRA。

3.5.1.2.1 LoRA 方法

Transformer 的 attention 网络结构中的参数通常是冗余的，它们可以精简到一个低维中完成各种 NLP 任务。低秩分解便是一种将高维稠密参数向量降维分解为稀疏的低维向量的方法。

LoRA[52]的基本原理是在冻结原模型参数的情况下，通过向模型中加入额外的网络层，并只训练这些新增的网络层参数。由于这些新增参数数量较少，这样不仅微调的成本显著下降，还能获得和全模型微调类似的效果，如下图所示：

- Pretrained Weights 部分为预训练好的模型参数，LoRA 在预训练好的模型结构旁边加入了 A 和 B 两个结构，这两个结构的参数分别初始化为高斯分布和 0。
- A 的输入维度和 B 的输出维度分别与原始模型的输入输出维度相同，而 A 的输出维度和 B 的输入维度是一个远小于原始模型输入输出维度的值，这就是 low-rank 的体现，可以极大地减少待训练的参数。
- 在训练时只更新 A、B 的参数，预训练好的模型参数是固定不变的。在推断时利用重参数思想，将 AB 与 W 合并，这样在推断时不会引入额外的计算。而且对于不同的下游任务，只需要在预训练模型基础上重新训练 AB，这样也能加快大模型的训练节奏。

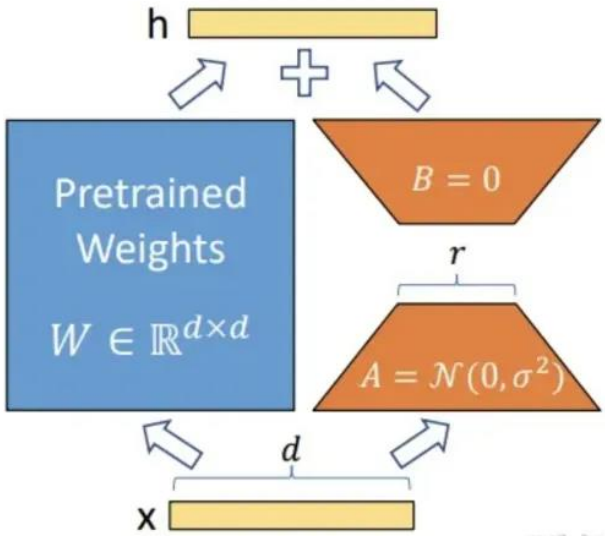

图 3.12 LoRA 算法

LoRA 的优势很明显：

- 预训练模型参数可以共享并保持冻结，因此模型不容易发生灾难性遗忘。
- 秩分解矩阵的参数明显少于原始模型，根据不同的任务可以构建不同的小型 LoRA 模块，移植性很强。我们可以通过替换矩阵 A 和 B 来冻结共享模型并有效地切换任务，从而显著降低存储需求和任务切换开销。
- 当使用 adapter 时，因为我们不需要计算梯度或维护大多数参数的优化器状态，LoRA 使显存开销下降。

- LoRA 简单的线性设计允许我们在输出时将可训练矩阵与冻结权重合并即可，通过构造与完全微调的模型相比，LoRA 不会引入推理延迟。
- LoRA 与许多先前的方法正交，并且可以与其中的许多方法组合，例如 p-tuning。

LoRA 的也有一些缺点：

- LORA 进行低秩分解时候可能会损失一些模型的表达能力和泛化能力。
- LORA 微调方法可能会受到初始化和超参数的影响较大，需要进行适当的调整。

3.5.1.2.2 QLoRA 方法

QLoRA 方法[53]使用一种低精度的存储数据类型（NF4）来压缩预训练的语言模型。通过冻结 LM 参数，将相对少量的可训练参数以 Low-Rank Adapters 的形式添加到模型中，LoRA 层是在训练期间更新的唯一参数，使得模型体量大幅压缩同时推理效果几乎没有受到影响。从 QLoRA 的名字可以看出，QLoRA 实际上是 Quantize+LoRA 技术。

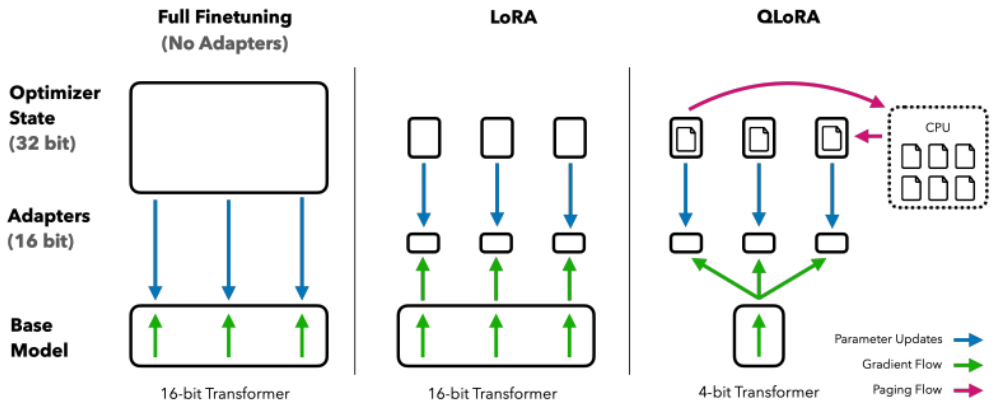

图 3.13 QLoRA 算法[53]

QLoRA 的量化核心技术有三个：4-bit NormalFloat(NF4)量化、二次量化(Double Quantization)和 Paged Optimizers。

- NF4 量化：这种数据类型基于分位数量化技术，并在信息理论上是最优的。由于预训练的神经网络权值通常具有标准差为 0 的正态分布性质，因此我们可以通过缩放系数将所有的权值转换为固定期望值，从而使该分布完全适合我们的数据类型范围。一旦数据类型范围和权重范围匹配，我们就可以像往常一样进行量化。分位数量化技术的主要思想便是将数值尽量落到均值为 0，标准差为[-1,1]的正态分布的固定期望值上。前面我们知道离群值对于模型量化的影响极其重要，而由于分位数估计算法的近似性质，精度量化对于离群值又有很大的误差。分位数量化技术使得每个量化分区中具有相等的期望值，相等的期望值可以避免昂贵的分位数估计和近似误差，使得精确的分位数估计在计算上可行。

- 二次量化：是将额外的量化常数进行二次量化以减小内存开销的过程。例如每 64 个参数块共享一个 32bit 的量化常数， 这样的话相当于每一个参数的量化额外开销为 0.5 bit。这个总体来说也是比较大的一个开销，所以为了进一步优化这个量化开销，我们对其进行二次量化，即把第一次 32bit 量化的输出作为第二次量化的输入，我们采用 256 的块大小对量化常数进行 FP8 量化，这样的话，我们可以把每个参数的量化开销每个参数减少了 0.373bit。
- Paged Optimizers：使用 NVIDIA 统一内存功能，该功能在 CPU 和 GPU 之间进行自动 page 对 page 传输，以便在 GPU 偶尔 OOM 的情况仍然下进行模型训练和微调。 可以理解成显存偶发 OOM 时，QLoRA 会将优化器状态自动的驱逐到 CPU RAM，当在优化器更新步骤中需要内存时，它们会被分页回 GPU 内存，从而保证训练正常训练下去。

3.5.1.3 效果验证

Text2SQL 领域目前比较公认的评价模型预测准确率的方式是执行结果正确性评测，即预测的 SQL 语句只要执行返回结果与标准答案一致，即使预测的 SQL 语句与标准答案的 SQL 语句不完全一致，也认为是正确的。在 Text2SQL 领域，SQL 已经成为了一个通用的标准，测试所使用的数据库和数据可以通过 SQLite 以一种低成本的方式集成在测试系统中。然而在 Text2GQL 领域，由于不同的图数据库使用着不同的图查询语言，数据导入格式也不同，执行结果正确性评测需要针对每一种数据库启动数据库的服务并导入对应的测试数据，开发成本较高。因此针对不同程度的开发成本设计了如下四个级别的评测方式：

- 文本相似度评测（零开发成本）：使用 Jaro-Winkler Distance 等文本相似度算法计算预测的 GQL 语句与标准 GQL 语句的差异，不需要启动数据库服务
- 语法正确性评测（需要语法解析器）：使用.g4 语法文件生成的语法解析器，或者将 JAVA 项目中的语法解析器打包调用，判断预测 GQL 语句的语法正确性，可以与文本相似度评测配合使用
- 执行计划正确性评测（需要数据库服务，无需实际数据）：在不生成实际测试数据的情况下，将数据库解析 GQL 语句后得到的执行计划与标准答案的执行计划进行对比，借此预测执行结果是否一致
- 执行结果正确性评测（需要数据库服务与数据导入）：需要启动数据库并导入与 GQL 语句对应的测试数据，直接对比执行结果，这一方法得到的准确率最具有参考意义，但是开发成本最高。

TuGraph 团队提供的 GQL（tugraph-analytics）语料以及 Cypher（tugraph-db）语料在 CodeLlama-7b-instruct 模型上微调后，文本相似度及语法正确性准确率达到 92%以上。

表 3.1 Text2GQL 微调模型性能

Language	Dataset	Model	Method	Similarity	Grammar
Cypher (tugraph-db)	TuGraph-DB Cypher数据集	CodeLlama-7B-Instruct	base	0.769	0.703
			lora	0.928	0.946
GQL (tugraph-analytics)	TuGraph-Analytics GQL 数据集	CodeLlama-7B-Instruct	base	0.493	0.002
			lora	0.935	0.984

3.5.2 图系统优化

在当今数字化时代，数据量呈现爆炸式增长，数据之间的关联变得日益复杂。图系统作为一种能够有效处理和分析图数据的工具，正逐渐在各个领域发挥重要作用。与此同时，人工智能技术的飞速发展，特别是机器学习和深度学习算法的进步，以及大语言模型在自然语言处理领域的显著进展，为图系统带来了新的机遇和可能性。将图系统与人工智能、大语言模型相结合，可以充分发挥三者的优势，实现更高效的数据处理和分析，为各种应用场景提供更有价值的洞察和决策支持，从而推动各个领域的创新和发展，将三者相结合，可以实现更深入的语义理解、更精准的决策支持、高效的信息检索和个性化交互等。

总的来看，图系统优化的主要目标有：

1、查询性能优化：通过优化图数据库和图处理引擎，减少查询响应时间，提高图数据的访问效率。

2、运维自动化：能够适应动态的数据规模和业务需求，实现系统的扩展伸缩、诊断调优。

3、存储效率优化：合理利用存储空间，降低存储成本，同时确保数据的完整性和一致性。

4、系统安全增强：建立完善的安全机制，防止数据泄露、误操作等安全问题。

5、产品体验优化：提供友好的用户界面和便捷的操作方式，使图系统易于使用和管理。

3.5.2.1 查询性能优化

3.5.2.1.1 任务优化

一）优化执行计划

1） 查询理解与重写

自然语言查询的复杂性使得数据库查询的准确理解变得困难。LLM 可以通过对自然语言查询的分析，理解用户的查询意图，并将其转化为准确的数据库查询语言表达形式。

LLM 还可以对复杂的查询进行重写，以提高查询的效率。例如，将嵌套查询重写为连接查询或子查询组合，减少查询的复杂性和执行时间。通过对查询的理解和重写，LLM 可以帮助数据库管理员和开发人员更快速地构建准确高效的查询。

2） 索引推荐

索引是提高数据库查询性能的重要手段。然而，选择合适的索引并非易事，需要对数据库模式和查询需求有深入的了解。LLM 可以通过分析数据库模式和历史查询记录，推荐可能提高查询性能的索引。

例如，如果历史查询中经常根据某个列进行条件筛选，LLM 可以建议创建该列的索引。对于多表连接查询，LLM 可以推荐创建复合索引，以加快连接操作的速度。同时，LLM 还可以解释创建每个推荐索引的理由，帮助数据库管理员做出决策。

3） 查询计划评估与选择

数据库引擎通常会为一个查询生成多个不同的执行计划，选择最优的执行计划对于提高查询性能至关重要。LLM 可以分析不同的基于规则的优化器（RBO）、基于成本的优化器（CBO）和基于人工智能的优化器（AIBO）等不同策略下的查询计划，并评估它们的潜在性能。

从 RBO 到 CBO 再到 AIBO 的演进体现了查询优化策略的不断进步。RBO 主要依据固定的规则来选择执行计划，其优点是简单快速，但缺乏灵活性，无法适应复杂多变的查询环境。CBO 则是基于成本估算来选择执行计划，考虑了更多的因素如数据分布、索引使用等，比 RBO 更加灵活和准确。AIBO 则是利用人工智能技术，如 LLM，对查询进行更深入的分析和理解，能够更好地适应各种复杂的查询场景。LLM 可以通过考虑查询的特点、数据库的统计信息和硬件资源等因素，预测每个计划的执行时间、资源消耗等性能指标。根据评估结果，LLM 可以为数据库引擎提供建议，选择最优的查询计划。例如，当多个计划在性能上接近时，LLM 可以根据特定的应用场景或性能指标偏好，推荐最合适的计划。

同时，在查询执行过程中，并行计算和异构资源的调度也对查询性能有重要影响。并行计算可以通过同时处理多个任务来提高查询效率，例如将一个大型查询分解为多个子查询并行执行。而异构资源的调度则可以合理利用不同类型的硬件资源，如 CPU、GPU 等。对于一些计算密集型的操作，可以将其分配到 GPU 上进行处理，以提高计算速度。LLM 可以分析查询的性质和硬件资源的情况，为并行计算和异构资源的调度提供建议，使得查询能够在更短的时间内完成。同时，在多节点的分布式数据库环境中，并行计算和异构资源调度还涉及到节点间的通信和协作，LLM 也可以对此提供分析和优化建议，以确保整个查询过程的高效执行。

二）提升运行性能

1） 实时监控与调整

在查询执行过程中，实时监控数据库的性能指标对于及时发现和解决性能问题至关重要。LLM 可以通过与数据库的监控系统集成，实时监控数据库的性能指标，如 CPU 使用率、内存占用、磁盘 I/O 等。

如果发现性能问题，如某个查询导致资源过度消耗或响应时间过长，LLM 可以提出调整建议。例如，建议调整数据库参数、临时增加资源分配（如内存或 CPU 核心），或者重新优化特定的查询。通过实时监控和调整，LLM 可以帮助数据库保持良好的运行时性能。

2）异常检测与处理

查询执行过程中可能会出现各种异常情况，如死锁、长时间等待资源、查询超时等。这些异常情况会严重影响数据库的性能和可用性。LLM 可以通过对数据库日志和性能指标的分析，检测查询执行过程中的异常情况。

一旦发现异常，LLM 可以提供诊断和解决方案。例如，对于死锁情况，LLM 可以分析死锁的原因，并建议采取适当的解锁措施，如回滚某个事务或调整事务的隔离级别。对于超时查询，LLM 可以建议优化查询语句、增加资源或调整查询计划。通过异常检测和处理，LLM 可以提高数据库的稳定性和可靠性。

3）性能预测与资源规划

随着业务的发展和数据量的增长，数据库的性能需求也会不断变化。提前进行性能预测和资源规划可以帮助企业更好地应对未来的挑战。LLM 可以根据历史查询执行数据和当前的数据库负载情况，预测未来的查询性能和资源需求。

例如，预测在特定时间段内的查询流量峰值，并建议提前增加服务器资源或调整数据库配置以应对。同时，LLM 还可以根据预测结果制定长期的性能优化策略，如定期进行数据库维护、优化索引或调整存储布局。通过性能预测和资源规划，LLM 可以帮助企业更好地管理数据库资源，提高数据库的性能和可用性。

3.5.2.1.2 算法优化

为了解决多项式不可解的复杂图问题，结合大语言模型（LLMs）与传统图算法的策略正在逐步显现出优势。首先，可以利用 LLMs 生成更具泛化性的启发式函数，从而提升传统图算法（如图编辑距离、子图匹配）的求解效率。其次，LLMs 的语义推理能力可以与图算法结合，减少搜索空间，使得在复杂任务中的求解过程更加高效。通过这两种方式的结合，传统图算法不仅能够保留其处理结构化数据的优势，还能在语义推理与全局优化方面取得显著提升，从而有效应对复杂图问题。

一）语义优化

图算法在推理复杂语义关系时往往难以有效处理特别是当图中的节点和边涉及大量文本信息时。比如在知识图谱中，节点可能代表特定实体，边则代表实体间的关系。若这些节点和关系包含丰富的语义信息，传统图算法往往无法完全捕捉其复杂的上下文和语义含义。以子图查询为例，严格的执行子图匹配可能会导致遗漏很多虽然结构上与查询图不同构，但是在语义上很接近、有意义的匹配。

大语言模型 LLM 可以通过提供强大的语义理解和上下文推理能力，帮助图算法对复杂文本信息进行更深层次的理解。具体优化方式包括：

语义嵌入增强：LLMs 通过将节点和边的文本信息转化为高维的语义嵌入，使得图算法能够利用语义信息进行更加精准的图结构分析。这种方式不仅能够捕捉节点之间的基本关系，还能识别复杂的语义关联，如同义词、上下位关系等，有效提升图神经网络 GNNs 的表示能力。

上下文感知推理：在知识图谱补全任务中，LLM 能够根据已有的图结构和上下文信息推断出潜在的关系或新节点。例如，面对缺失或不明确的边关系，LLMs 可以利用文本信息补充推理，极大增强了图算法的推理深度与广度。

多模态数据的联合处理：LLMs 能将不同模态的数据，例如文本、图像、图结构等，映射到统一的表示空间，使得图算法可以同时考虑结构化与非结构化信息。这种方式尤其适用于需要综合分析多种数据来源的任务，如推荐系统、情感分析等

特征丰富化与降维：通过 LLMs 生成的特征嵌入，可以丰富图节点的特征维度，从而帮助图算法在高维空间中更好地进行聚类或分类，同时使用图算法降维手段优化特征，减轻模型的计算负担。

二）效率优化

在 LLM 出现之前，已有许多基于学习的（learning-based）方法被提出，以加速复杂的图算法。通过结合人工智能技术与传统图算法，这些方法在处理复杂图问题时展现出了显著的效率提升。例如，在子图匹配问题中，研究者使用强化学习技术来优化匹配顺序，通过学习选择更高效的匹配路径，减少计算开销。另一个典型方法是基于图神经网络，通过对路径的向量表示进行检查，可以有效识别图中的重要候选节点，从而在图匹配过程中实现剪枝操作，显著减少不必要的计算。这些 AI 驱动的技术为图算法加速奠定了基础，使得处理复杂图结构问题时能够更加高效。

在 LLM 问世之后，大语言模型的灵活性和通用性为图算法的加速带来了新的机遇。例如，FunSearch 框架利用了 LLM 的生成能力，为图算法设计和优化提供了创新方案。首先，该框架将特定问题进行抽象并编写具体的算法模板，保留需要大语言模型优化的部分作为提示词输入 LLM。接下来，通过多次采样大语言模型生成的不同算法，并将它们送入评估函数进行评分，保留得分较高的算法并存入算法仓库。之后，从算法仓库中随机选择一个已有的算法作为新一轮提示，继续

输入大语言模型进行迭代生成。通过反复迭代和评估，最终可以得到经过大语言模型优化的全新算法。

尽管 FunSearch 框架展现出强大的生成和优化能力，但由于问题模板需要手动设计，尤其是对于复杂的图问题，如何合理设计提示模板对优化效果至关重要。此外，这种方法依赖大量的大模型推理调用，虽然在算法设计方面已显著提升了效率，但在成本控制上仍有改进空间，特别是在大规模图问题中的应用探索。

3.5.2.2 运维自动化

3.5.2.2.1 系统扩展优化

随着大语言模型（LLM）技术的不断发展，其在图数据库及相关领域的潜在应用前景广阔。首先，LLM 可以极大地增强工作负载预测的准确性，通过分析用户查询模式和历史数据，实时生成精确的预测模型，帮助系统管理员提前识别高负载情景，从而优化资源配置。其次，LLM 能够在资源分配优化中发挥关键作用，自动学习和适应不同的查询需求，动态调整计算和存储资源，以确保系统在高峰期的高效运行。此外，LLM 在自动扩展方面的应用前景同样显著。通过实时监测系统状态和负载变化，LLM 可以智能化地推荐扩展策略，使得系统能够快速响应突发流量，保持稳定性能。结合这些优势，LLM 的应用将推动图数据库管理的智能化转型，提升系统的整体扩展性和灵活性，为未来的数据库架构提供新的可能性。

一）工作负载预测

工作负载预测 （Workload Forecast）在图数据库系统中扮演着至关重要的角色，因为图数据库的独特结构和数据关系使得工作负载的变化往往具有复杂性。

在现实应用中，某些操作（如图更新、图查询）可能在特定时间段内显著增加，尤其是在社交网络分析、推荐系统或实时数据处理等场景中。例如，在大型社交平台上，当用户活动增加时，系统需要快速响应频繁的查询和更新请求。同时，由于图数据的特点，热点操作往往集中在图的特定局部区域，例如，由于热点事件往往与地区和社群高度关联，热点操作往往某一社群的节点中。当某个节点因为用户互动而频繁更新时，传统方法可能未能识别出这一热点区域，导致该部分的性能瓶颈，而其他节点却处于闲置状态。

传统的工作负载预测预测方法利用历史数据和专家经验对工作负载进行预测。随着工作负载的动态变化，历史数据可能无法有效预测未来的工作负载波动，导致资源配置滞后，系统在高峰期无法及时响应用户需求，造成性能瓶颈。此外，传统方法通常需要大量的人工干预与调整，这不仅增加了管理成本，还可能导致人为错误，进一步降低系统的可靠性和效率。

在这种情况下，大模型（如图神经网络）展现出了显著的优势和机遇。一方面，它们能够通过学习节点及其关系的复杂模式，实时捕捉局部负载变化。这种能力使得管理员可以及时调整资

源配置，确保系统在高负载时仍能保持良好的性能，避免服务中断或性能下降；另一方面，由于大模型具有强大的语义解析能力，可以阅读和理解图数据库中出现的语义信息，例如社交网络上的推文和评论。这种能力使得模型能够识别出潜在的热点话题或用户行为模式，从而为预测未来的负载变化提供更深入的洞察。

二） 资源分配优化

在图数据库中，资源分配优化（Resource Allocation）不仅关乎计算资源的有效利用，还涉及存储和网络带宽的合理分配。

由于图数据库的查询通常需要对大量节点和边进行操作，且这类操作往往发生在整体数据的热点区域。例如，在执行复杂的图算法（如可达性或子图匹配）时，特别是在需要同时服务多个用户和多个算法的情况下，资源的即时调配显得尤为重要，这直接影响了查询的响应时间和系统的整体性能。

在这种情境下，首先需要衡量每个任务的紧急程度。某些查询可能需要实时响应，例如社交网络中的即时消息或动态推荐，而图特征分析等任务耗时较长，对系统时延的要求相对较低。通过识别任务的紧急性，系统可以优先分配资源给高优先级的查询，确保关键操作的及时完成。其次，考虑多任务之间的公共计算也是资源优化的重要方面。许多图算法在处理数据时可能会共享相同的计算资源或数据集。通过智能调度，系统能够识别这些公共计算部分，从而减少冗余计算，提高整体效率。例如，在执行多个最短路径计算时，可以共享中间结果，避免重复处理相同的节点和边。

大模型的引入使得资源分配不再仅依赖于静态的历史数据，而是基于实时的工作负载特征进行动态调整。随着系统检测到某一特定查询模式增多时，模型可以自动增加相应的计算资源，优化查询的响应时间。此外，大模型能够持续学习系统的使用模式，这种自适应能力使其能够识别潜在的负载变化，从而提供智能化的资源配置建议。

三）自动化伸缩

自动化伸缩（Auto-Scaling）是图数据库系统应对动态工作负载的关键机制，尤其在面对突发流量时更为重要。

传统的自动扩展方法往往依赖于预设的阈值和静态策略，这在面对图数据库复杂的查询模式时，可能导致资源配置不当。例如，在数据分析高峰期，用户可能会同时发起多个复杂查询，而传统方法可能无法及时扩展资源，最终导致查询延迟或失败，影响用户体验。

通过引入大模型，系统能够实时监测负载变化，自动调整资源配置。通过在图数据库中部署资源监控 Agent。当资源监控 Agent 检测到用户活动激增时，基于 Agent 的自动扩展机制可以可

以激增活动的特点，对对应的计算资源进行分配，增加计算节点，确保系统能够处理突发的查询请求。这种基于实时数据的动态响应能力，显著提高了系统的灵活性和适应性。

此外，利用大模型的自适应学习能力，使用拓展的历史数据对大模型进行微调，可以使其能够不断优化扩展策略。通过分析历史数据和实时负载，模型可以识别出不同场景下的最佳扩展时机和资源需求，逐步提高系统在各类情况下的响应能力和资源使用效率。例如，在某些情况下，系统可能会发现特定查询模式在高峰期出现的频率，从而预先调整资源，避免潜在的性能瓶颈。这种智能化的自动扩展方法，不仅提升了图数据库的稳定性和可用性，还确保了资源的高效利用。

3.5.2.2.2 自动化任务诊断

在图数据库系统的运行过程中，海量的日志和监控数据源源不断地生成，这些数据包含了系统性能、资源使用、查询响应等关键信息。大语言模型（LLM）在智能诊断方面的应用，能够显著提升故障检测的效率和准确性。

一）实时分析海量日志

LLM 具备强大的自然语言处理能力，能够自动解析和理解图数据库生成的各类日志文件。通过训练，LLM 可以识别不同类型的日志信息，如错误日志、警告日志和信息日志，并根据上下文关系快速定位异常模式。例如，当系统出现查询延迟增加的情况时，LLM 可以扫描相关日志，识别出与延迟相关的具体错误信息或警告信号，及时发出预警。

二）监控数据的动态解读

除了静态的日志分析，LLM 还能够处理和解读实时监控数据。通过集成图数据库的监控系统，LLM 可以持续跟踪 CPU 使用率、内存消耗、磁盘 I/O 等关键指标，并通过时间序列分析识别出潜在的性能瓶颈或资源短缺。例如，在高并发访问场景下，LLM 可以实时监测到系统资源的异常消耗，并迅速分析其与当前查询请求之间的关联，从而辅助系统管理员做出快速反应。

三）根因分析与报告生成

LLM 通过学习系统的正常运行模式和大量的历史故障案例，具备了识别和分析复杂故障原因的能力。当系统出现问题时，LLM 能够综合考虑多个因素，如资源瓶颈、配置错误、数据不一致等，进行多维度的原因分析。例如，若某节点频繁超载，LLM 不仅会指出资源使用的异常，还会进一步分析是否由于特定查询模式导致的负载集中，或者是由于网络延迟引发的数据传输瓶颈。

在此基础上，基于对系统状态和历史数据的全面理解，LLM 能够自动生成详细的根因分析报告。这些报告不仅指出问题的表面现象，还深入挖掘问题背后的本质原因，并提供清晰的逻辑链条。例如，在发现某个查询导致系统性能下降时，LLM 可能会通过分析查询执行计划、数据访问路径和资源使用情况，确定具体是由于某个索引缺失或数据分布不均衡引发的性能问题。通过这种深层

次的原因分析与详尽的报告生成，LLM 能够帮助运维人员迅速定位并解决复杂的系统故障，提升图数据库系统的稳定性和可靠性。

四）智能化排障建议

在明确问题根源后，LLM 能够根据系统的具体情况和最佳实践，生成具体的排障建议。

针对配置错误，LLM 分析当前系统配置与最佳实践的差异，建议调整缓存大小、连接池设置或查询优化参数以优化系统性能；

对于数据不一致或损坏的问题，LLM 提供数据校验、修复或重新同步的步骤，确保数据的完整性和一致性；

针对导致系统故障的查询，LLM 建议重构查询语句、添加必要的索引或优化数据访问路径，以提高查询效率。

此外，LLM 能够根据具体问题推荐或生成自动化脚本和工具，例如日志分析脚本或监控配置，帮助运维人员快速实施解决方案。

最后，LLM 还提出预防性措施，如定期系统审查、优化配置策略、加强权限控制和完善监控报警机制，以减少类似问题的再次发生。

通过这些具体的排障建议，LLM 有效地支持运维人员解决图数据库系统中的各种问题，提升系统的可靠性和可维护性。

3.5.2.2.3 智能化调优

在图数据库系统的运行过程中，优化配置参数对于提升系统性能和稳定性至关重要。LLM 通过分析图数据库的运行数据和性能指标，能够自动识别需要优化的系统参数，并提供相应的调整建议。

首先，LLM 能够对比当前系统配置与行业最佳实践，识别出潜在的配置瓶颈。例如，通过分析查询响应时间、内存使用率和磁盘 I/O 等关键指标，LLM 可以确定哪些参数（如缓存大小、并发连接数或索引策略）需要调整以提升系统性能。

其次，LLM 具备实时学习和适应能力，能够基于持续收集的性能数据，动态调整参数设置，确保系统始终处于最佳运行状态。

此外，LLM 还能够分析不同参数调整对系统性能的影响，帮助运维人员在实施变更前评估其潜在效果，降低调优风险。通过自动化的参数调优，LLM 不仅简化了运维流程，还显著提升了图数据库的响应速度和资源利用效率，确保系统能够高效处理复杂查询和大规模数据操作，增强整体系统的可靠性和可维护性。

3.5.2.3 存储效率优化

3.5.2.3.1 数据预取与缓存

传统数据库通常采用几种数据预取与缓存策略来优化 I/O 效率，包括基于 LRU（最近最少使用）算法的缓存管理、预取策略如顺序预取和基于访问模式的预取。这些策略依赖固定的规则和启发式算法，通常根据用户的历史访问记录来决定哪些数据应被缓存。然而，这些方法往往无法灵活适应用户行为的变化，且对复杂的访问模式识别能力较弱。

相比之下，LLM 在数据预取与缓存策略方面展现出显著优势。首先，LLM 通过深度学习和自然语言处理，可以更深入地理解用户的查询意图和上下文，能够识别出更复杂的访问模式。这使得 LLM 能够预测用户未来的需求，从而更准确地决定哪些数据应该被预取并加载到缓存中。其次，LLM 的自适应能力使其能够实时调整策略，快速响应用户行为的变化，而传统数据库往往需要手动调节或重新配置。最后，LLM 能够综合考虑多种因素（如时间、用户行为和数据特性）来优化缓存策略，提升数据访问的效率和准确性。

3.5.2.3.2 存储索引设计

一）存储结构

传统图数据库采用多种存储结构来高效管理节点和边的数据，例如压缩稀疏列（CSC）、压缩稀疏行（CSR）、链表（Linked List）和键值对（KV Pair）等。这些存储结构在不同的应用场景下各具优势。例如，CSC 和 CSR 适用于高效的矩阵运算和快速的邻接访问，链表结构便于动态插入和删除操作，而 KV Pair 则在处理稀疏数据和灵活的模式匹配方面表现出色。然而，这些传统方法通常依赖于固定的存储策略，难以根据实时数据和查询模式的变化进行动态优化。引入大语言模型（LLM）后，系统可以通过深度学习和数据分析，自动识别最适合当前数据特性和访问模式的存储结构。LLM 能够根据节点和边的属性、查询频次以及数据分布情况，智能选择或组合不同的存储结构，从而提升数据访问效率和存储利用率。

二）索引设计

索引设计是提升图数据库查询性能的关键因素，传统方法通常依赖预定义的索引策略，如基于节点属性、关系类型或路径模式的索引。这些方法在处理静态的查询模式时效果显著，但在面对动态和复杂的查询需求时，往往难以保持高效性。大语言模型（LLM）的引入为索引设计带来了智能化的提升。LLM 通过分析海量的查询日志和数据模式，能够深入理解复杂的查询意图和数据关系，自动识别最具价值的索引结构。基于实时的查询频次、数据访问路径和节点关系，LLM 可以生成最优的索引方案，如多维索引、组合索引或分层索引，显著提升查询响应速度。此外，LLM 具备自适应学习能力，能够持续监测系统的查询负载和数据变化，动态调整索引策略，确保索引结构始终与实际需求高度匹配。

3.5.2.4 系统安全增强

随着图系统在各个领域的广泛应用，如金融、医疗、社交网络等，其存储和处理的数据变得越来越敏感和重要。此外，频繁发生的黑客攻击和数据泄露等事件给企业和用户造成了巨大的损失。因此，加强图系统的安全防护，提高其安全性能，成为了当务之急。

3.5.2.4.1 数据加密

在图数据库中，数据的敏感性各异，因此自动加密策略显得尤为重要。大语言模型（LLM）能够分析数据的敏感度，自动生成相应的加密规则，以保护关键数据不被未授权访问。这种自动化的加密策略不仅提高了数据安全性，还减少了人为错误的风险。

密钥管理是安全防护的核心。借助 LLM 的语义理解能力，系统可以提高密钥分配和管理的智能化水平，确保密钥的安全性与有效性。通过深入分析密钥使用情况和访问模式，LLM 能够优化密钥生命周期管理，确保密钥的及时更新和安全存储。

当人工设置的防护级别滞后或不匹配时，LLM 可以自动调整加密策略或向人类专家发出提示。这种动态调整机制确保系统始终处于最佳防护状态，及时响应变化的安全需求。通过结合 LLM 的自动化能力与人类专家的判断，数据加密和密钥管理的整体安全性得以提升，确保敏感数据在图数据库环境中的安全性和完整性。

3.5.2.4.2 漏洞管理

编译阶段，LLM 对图数据库的源代码进行静态分析，检查代码中的潜在漏洞和不符合安全规范的部分。通过高级的自然语言处理技术，LLM 能够识别代码中的安全缺陷，如未处理的输入验证、权限管理漏洞等，确保在代码进入测试阶段前，尽可能地减少已知的安全问题。

测试阶段，LLM 辅助自动化生成和执行测试用例，包括常规测试和极端情况（Edge Case）测试。通过模拟各种可能的攻击场景和异常行为，LLM 能够有效地验证代码的安全性和稳定性。自动化的测试流程不仅提高了测试的覆盖率，还确保了在面对复杂和罕见的情况时，系统能够保持稳健，减少潜在漏洞被利用的风险。

部署阶段，LLM 负责实时监控图数据库的运行状态，通过分析日志和网络流量，及时发现异常模式和潜在的安全威胁。LLM 能够主动识别并评估新出现的风险，提供实时的风险预警和响应建议。此外，LLM 还可以根据监控数据动态调整安全策略，确保图数据库在实际运行环境中的持续安全与稳定，预防数据泄露或损坏。

通过“编译->测试->部署”三个节点的全面覆盖流程，充分发挥了 LLM 在漏洞管理各个环节中的优势，从源代码的静态检查，到测试阶段的全面验证，再到部署后的实时监控，全面提升了图数据库的安全性和可靠性。

3.5.2.4.3 安全监控

实时安全监控是保障图数据库安全的重要环节。LLM 能够学习正常的行为模式，并实时检测异常活动，从而及时发现潜在的安全威胁。通过分析用户行为、访问模式和数据流量，LLM 能够识别出与正常行为偏离的活动。在安全事件发生时，LLM 结合强大的理解和分析能力，可以快速响应，提高事件处理的效率，减轻安全团队的负担。其自动化的响应机制能够在第一时间采取防护措施，阻止进一步的损害。

3.5.2.4.4 智能决策

在图数据库安全防护中，大语言模型（LLM）通过自动化安全流程与智能化决策的紧密结合，显著提升了整体的安全性和响应效率。

首先，LLM 在自动化安全流程中发挥了关键作用，它能够自动解析和分析海量的安全日志，识别异常行为和潜在威胁，从而大幅减少人工审查的工作量。同时，LLM 能够自动扫描系统漏洞，生成修复脚本和建议，加快漏洞修复的进程。通过持续监控用户活动，LLM 还能及时识别不合规操作并发出警报，有效防范内部威胁。这样的自动化流程不仅释放了人力资源，还使安全团队能够专注于处理更复杂的问题，提升了整体的安全防护能力。

在自动化流程的基础上，LLM 进一步通过其卓越的文本分析和推理能力，整合了来自多个来源的安全知识和最佳实践，形成了庞大的知识库。这使得安全专家能够随时进行即时查询和分析，快速获取所需信息，从而大幅提高工作效率。此外，LLM 还能够模拟不同的安全场景，帮助专家评估各种策略的效果，优化应急预案，增强决策的针对性和科学性。这一过程不仅提升了安全团队的响应速度，也确保了决策的准确性和有效性。

随着 LLM 对大量数据的深度语义理解和模式识别能力的发挥，它能够识别出潜在的安全模式和趋势，理解事件之间的关联性。通过结合历史数据和实时信息，LLM 构建了动态的风险评估模型，能够预测潜在威胁并量化其影响。这使得 LLM 能够主动识别风险，发现异常模式，为安全团队提供全面的风险视角，并提前预警潜在的安全事件，确保图数据库系统的持续安全。

在 LLM 整合安全知识库，并对潜在的风险进行识别后，LLM 能够为安全专家提供自动化的只能决策。基于动态风险评估模型，LLM 预测不同威胁的可能性及其影响，帮助专家制定修复策略的优先级，确保关键漏洞得到迅速解决，降低潜在损失。同时，LLM 还可以根据实时监控数据动态调整安全策略，优化整体防护措施，确保决策的及时性和有效性。

3.5.2.5 产品体验优化

3.5.2.5.1 自然语言交互

大语言模型（LLM）的核心优势在于其对自然语言的高度理解和生成能力，它能够模拟人类的语言交流，提供更智能的互动体验。在图数据库系统中，智能化的交互可以通过 LLM 快速回应用户的问题，减少对人工客服的依赖。例如，当用户在查询数据时遇到问题，LLM 能够快速理解请求，并提供有效的解决方案，减少等待时间。此外，LLM 还可以与语音识别技术结合，实现语音交互，使得用户可以通过语音完成复杂的数据查询和操作。这种自然语言交互适应不同用户需求，大大提升了服务的便捷性和用户体验的流畅度。

3.5.2.5.2 用户体验定制

个性化是现代用户体验的重要组成部分，LLM 能够根据用户的行为数据、历史偏好以及当前上下文生成个性化的推荐和内容。在图数据库系统中，LLM 可以分析用户的查询历史，推荐相关的数据集或信息，提升用户的使用体验。同时，LLM 还能够为用户生成个性化的对话，例如智能助手可以根据用户的日常习惯提供定制化的查询建议。这种个性化体验不仅增加了用户的参与度，还能使用户感受到系统的智能与贴心。

3.5.2.5.3 查询意图识别

查询意图识别是 LLMs 在优化用户体验中的一个关键功能，尤其适用于处理模糊或不完整的用户问题。当用户提出的问题有多个可能的匹配时，LLMs 能够通过智能反问，快速引导用户澄清需求，并缩小范围。特别低，是在图数据库系统中，一个模糊的问题可能有非常多的匹配，需要 LLMs 对用户的查询意图进行反问和识别。例如，用户在电影查询中提出“我想找一个爱情电影，他的导演同时也是演员”这样一个模糊的问题时，图数据库中可能存在大量满足条件的结果，如果讲所有结果同时返回，用户不得不逐个检查，造成用户体验不佳。此时，LLMs 可以进一步询问：“这部电影是哪个国家的”“这部电影是何时上映的”通过这样的问题澄清，引导用户澄清需求，并缩小范围，减少用户的困惑和等待时间。这种高效的查询意图识别提升了问题解决的准确性，优化了用户体验。

3.5.2.5.4 多语言与全球化

在全球化的市场中，提供多语言支持对于吸引和留住国际用户至关重要。LLM 具备强大的多语言处理能力，可以理解并生成多种语言的内容，从而实现跨语言的无缝沟通。例如，图数据库系统在为不同地区的用户提供服务时，LLM 能够自动翻译用户的请求，并生成相应语言的回应，减少因语言差异带来的沟通障碍。这不仅能够提高用户满意度，还能增强系统在全球市场中的竞争力。

3.5.2.5.5 用户反馈优化

LLM 还可以帮助图数据库系统更好地分析和处理用户反馈，推动产品和服务的持续改进。借助 LLM，系统能够快速处理海量的用户评论和反馈数据，并从中提取有价值的建议。例如，LLM 可以根据用户的评价数据，发现系统中的常见问题，并生成改进方案的建议。这种自动化分析不仅节省了时间和人力，还提高了反馈处理的效率，帮助系统快速迭代，精准满足用户需求。

3.5.2.5.6 数据隐私保护

在优化用户体验的过程中，数据安全和隐私保护是不可忽视的重要因素。LLM 能够帮助图数据库系统实现更好的数据管理和隐私保护。例如，LLM 可以通过自动化的数据脱敏技术，在用户数据的传输和分析过程中去除敏感信息，确保用户隐私得到保护。此外，LLM 还能够通过分析用户的语言模式和行为，识别潜在的安全威胁，如网络欺诈和数据泄露。在金融或医疗领域，LLM 可以实时监控用户的查询记录，识别异常行为，提供反欺诈预警。这种对安全性的提升，不仅增加用户对系统的信任，也能有效降低风险。

通过智能化的语言交互、个性化推荐、情感分析、多语言支持、反馈优化与数据安全保障，大语言模型在优化图数据库系统用户体验上展现了广泛的应用前景。系统通过合理应用 LLM，能够大幅提升服务的智能化和个性化水平，从而在竞争激烈的市场中脱颖而出。未来，随着 LLM 技术的不断发展，它在优化用户体验方面的作用将会更加显著，图数据库系统应积极探索并整合这一技术，打造更为出色的用户体验。

3.5.3 GraphRAG

3.5.3.1 背景

3.5.3.1.1 LLM 发展及其挑战

近年来，大规模语言模型（LLM）在自然语言处理领域取得了显著进展。模型如 GPT-3 和 GPT-4，通过学习大量的文本数据，能够生成自然流畅的文本，进行自动翻译，并且处理对话和文本总结等任务。这些技术突破使得计算机能够更好地理解和生成语言，从而提升了许多应用场景的智能水平。

尽管 LLM 在很多任务中表现出色，它们仍面临一些挑战。首先，在处理长文本时，LLM 有时会丢失上下文信息，导致生成的内容前后不一致。其次，由于这些模型基于静态的训练数据，对于新信息的适应能力较弱，可能会生成过时或不准确的内容，特别是在涉及专业领域时。具体而言，LLM 存在以下几个主要问题：

1、生成幻觉（Hallucination）：LLM 有时会生成不真实或虚假的信息，这种现象被称为“生成幻觉”。例如，当遇到特定或冷门的问题时，模型可能会创建出看似合理但实际上并不存在的

答案。这是因为模型在生成内容时，主要依赖于从训练数据中学习到的模式，而不是准确区分真实和虚假的信息。

2、专业知识不足（Domain Knowledge Deficiency）：虽然 LLM 在处理通用话题时表现良好，但在特定专业领域（如医学、法律等）的知识深度和准确性可能不足。这些模型的训练数据通常涵盖的是通用内容，导致在这些领域的生成结果可能缺乏足够的专业性和可靠性。

3、信息时效性低（Low Temporal Relevance）：LLM 的训练数据是静态的，这意味着模型在训练完成后不会自动更新。因此，模型在面对最新信息或近期事件时，可能无法提供最新的回答。例如，对于最近的科技进展或新闻事件，模型可能无法及时反映，从而生成过时的信息。

4、计算成本高（High Computational Cost）：训练和运行 LLM 需要大量的计算资源，这使得其成本相对较高。大规模的计算和存储需求不仅增加了经济成本，也对环境造成了负担。许多组织可能难以承担这些高昂的费用。

5、黑箱特性（Black Box Nature）：LLM 的内部决策过程复杂且不透明，这种特性被称为“黑箱”问题。模型的工作机制对用户来说是不可见的，这使得在模型出现问题时，难以追踪具体原因并进行改进。这种缺乏可解释性的问题限制了对模型的优化和调整，影响了其长期的可靠性和改进能力。

3.5.3.1.2 检索增强生成 RAG 及其局限性

检索增强生成 RAG（Retrieval Augmented Generation，RAG）是一种将检索和生成相结合的技术框架。在生成答案时，RAG 不仅依赖于预训练模型的内部知识，还通过从外部知识库中检索相关信息来增强生成过程。具体而言，RAG 包含以下两个主要阶段：

- **检索阶段（Retrieve）：**从知识库中检索与用户查询相关的文档或信息。
- **生成阶段（Generate）：**利用检索到的外部信息和用户输入，通过生成模型生成答案。

这种方法能够弥补生成模型的知识盲点，提供更加准确和可靠的回答。

RAG 的目标是通过知识库增强内容生成的质量，通常做法是将检索出来的文档作为提示词的上下文，一并提供给大模型让其生成更可靠的答案。更进一步地，RAG 的整体链路还可以与提示词工程（Prompt Engineering）、模型微调（Fine Tuning）、知识图谱（Knowledge Graph）等技术结合，构成更广义的 RAG 问答链路。

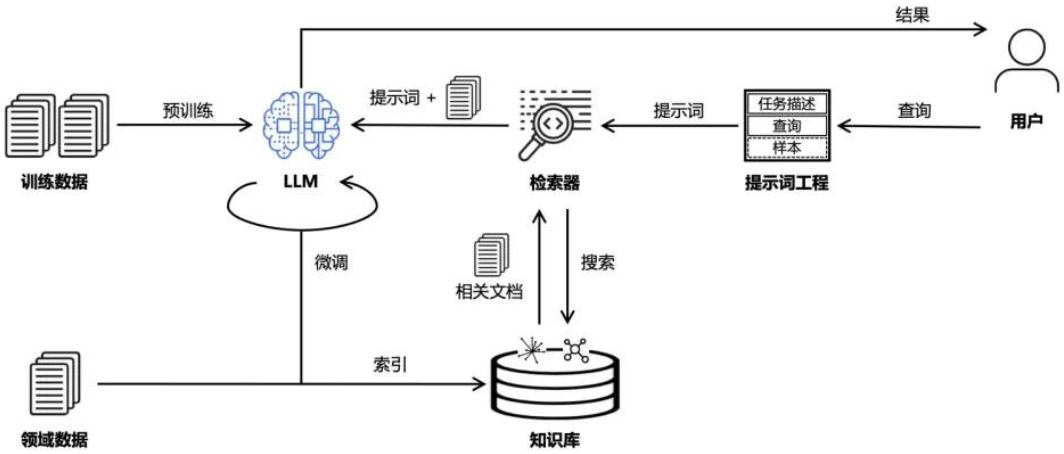

图 3.14 广义的 RAG 问答链路

除了增强内容生成，RAG 的理念还可以进一步泛化到链路的其他阶段：

- **增强训练**：引入知识检索器增强大模型预训练，以改进大模型的问答质量和可解释性。
- **增强微调**：实现对大模型和检索器的双指令微调，RAFT 通过微调让大模型可以识别干扰文档。
- **增强语料**：支持多模态数据的检索，提升了大模型在文本/图像混合检索场景下的推理质量。
- **增强知识**：使用图社区摘要解决总结性查询任务的问题，将知识图谱技术应用到 RAG。
- **增强检索**：通过对检索到的文档置信度进行评估，提升问答上下文的质量。
- **增强推理**：在推理阶段将 RAG 与 CoT 相结合，以改进长期推理和生成任务的效果。

知识库作为 RAG 链路的核心组件，直接影响了知识的存储与召回。支持融合索引（Converged Index）的知识库，可以更好地应对多样化的应用场景，因此设计通用的 RAG 架构应该兼容多种知识索引格式，包括 GraphRAG。

RAG 也有一些局限性。例如，RAG 在处理事务关联时的能力有限。想象一下，如果你需要回答一个关于公司内部部门之间合作关系的问题，RAG 可能无法有效整合涉及不同部门的复杂关系。比如，问到“研发部门如何与市场部门协作以推出新产品？”，RAG 可能会从知识库中提取关于研发和市场的一般信息，但难以综合它们之间的具体关系和交互细节，从而提供一个准确的答案。

3.5.3.1.3 基于图的新型检索增强生成技术 GraphRAG

GraphRAG 在 RAG 模型的基础上进行了改进，引入了图结构来处理信息。与传统的 RAG 模型不同，GraphRAG 将知识表示为图，并利用图中节点和边的关系来改进信息检索和生成。这种图结构能够捕捉和处理复杂的关系和事务关联，从而提供更准确、更全面的结果。

例如，在处理公司部门协作的问题时，GraphRAG 可以通过图结构明确表示研发部门和市场部门之间的关系、沟通渠道和合作历史，从而生成一个更为详细和精准的回答。这使得 GraphRAG 特别适合那些涉及复杂数据和多层次关系的领域，如知识图谱、电子商务和医疗等。GraphRAG 通过有效利用图结构中的信息，提升了检索和生成的质量，使得生成模型不仅能够处理传统的文本数据，还能更好地整合和利用复杂的关系数据，从而提供更智能、更高效的解决方案。

相比于传统 RAG，GraphRAG 从增强知识确定性角度做了进一步的改进，也就是知识内容增强的思路。

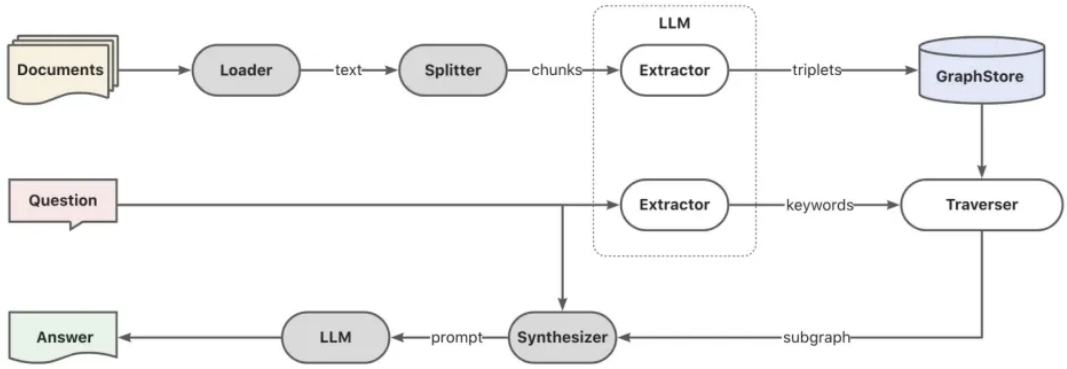

图 3.15 一般的 GraphRAG 链路

3.5.3.2 GraphRAG 概述

本节将探讨 GraphRAG 的技术原理，阐述其核心概念以及在检索增强生成（Retrieval Augmented Generation,RAG）中的应用方式。通过理解 RAG 和 GraphRAG 的定义，以及图索引和检索策略，读者将更好地掌握如何利用图结构和算法提升知识索引和召回的效率和效果。

3.5.3.2.1 基本概念

GraphRAG 是在 RAG 框架中引入图结构、图查询和图算法的一系列方法的总称。它通过在知识索引和召回阶段利用图形数据结构，增强了知识的表示和检索能力。GraphRAG 的核心思想包括：

- **知识表示**：将原始知识抽取并组织成图状结构，如知识三元组、属性图等。
- **图索引**：利用图索引技术高效地存储和检索图形化的知识数据。
- **图检索**：在检索阶段，使用图查询和图算法，从图结构中获取相关的知识。

在 GraphRAG 中，图结构的作用尤为重要。传统的 RAG 模型主要依赖于线性的检索方式，这意味着模型从知识库中逐条检索信息。而图结构则通过节点和边的关系，提供了更为丰富的上下文信息。在信息检索中，图结构能帮助模型识别和利用信息之间的复杂联系。例如，处理涉及多个实体和其相互关系的问题时，图结构能够清晰地展示这些实体和关系，从而提高检索的准确性和相关性。

在生成文本的过程中，图结构通过提供多层次的关系信息，帮助模型生成更连贯和上下文一致的内容。图中的节点代表具体的知识点，边则表示这些知识点之间的关系。这种结构化的表示方式帮助生成模块更好地理解和整合信息，从而避免了信息遗漏或前后不一致的问题。

GraphRAG 还引入了多模态信息融合的概念。这意味着模型不仅依赖于文本信息，还能够整合来自图结构的多种信息源。具体来说，GraphRAG 结合了图中的结构信息和文本信息来生成内容。这种融合方式可以显著提升生成任务的质量，使得模型能够提供更丰富、更深入的回答。

例如，在处理医学领域的问题时，GraphRAG 能够结合医学知识图谱中的疾病、症状、治疗方法等信息与生成模型的文本生成能力，从而提供更全面和准确的医学建议。通过多模态信息融合，GraphRAG 能够更好地利用图结构中的关系和上下文信息，从而生成具有更高深度和广度的内容。

3.5.3.2.2 主要组件

GraphRAG 模型由三个核心组件构成：图索引组件、图检索组件和增强生成组件。这些组件协同工作，提高信息检索和生成的效果。

一）图索引组件

图索引组件负责通过自然语言处理或大型语言模型将外部知识库中的信息抽取并组合成图状结构。具体包括：

- **知识三元组（Subject-Predicate-Object）**：从文本中抽取实体（如人名、地点）及其关系（如“属于”），形成基本的知识单元。
- **属性图（Property Graph）**：在节点和边上附加属性信息，丰富图的语义。
- **图谱状知识与原文片段关联**：将图结构与原始文本片段关联，保留上下文信息。
- **关联原始知识生成的问答对和摘要**：通过关联原始知识，生成相关的问答对和摘要，丰富图的内容。

此外，GraphRAG 还支持利用现有的图状数据或知识图谱作为知识来源。这些现有的数据可能包括：

- **公共知识图谱**：如维基百科、DBpedia 等公共资源，包含大量实体和关系。
- **企业内部知识图谱**：由企业构建的专有领域知识图谱，涵盖特定行业或业务领域的信息。
- **多表 ETL 导入的图数据**：从多个数据表经过 ETL（抽取、转换、加载）流程导入的图形数据。

通过直接利用这些现有的图状数据，GraphRAG 可以避免从头开始构建图结构，充分发挥已有知识资源的价值。

通过图索引组件，GraphRAG 将知识以结构化方式组织，并定期更新图，保持其准确性和时效性，准备好供后续检索和生成使用。

二）图检索组件

图检索组件从图结构中提取与用户问题相关的信息，在 RAG 的检索阶段，利用图结构和特性，可以采用多种召回策略，以充分发挥图形化知识的优势。

1、关键实体的提取与图检索

在查询时，从用户的问题中提取关键实体，然后在图中进行检索：

- **实体匹配**：找到与关键实体对应的节点。
- **关系扩展**：沿着图中的边，探索与关键实体相关的节点和关系。
- **子图提取**：提取与查询相关的子图，作为知识检索的结果。

这种方法利用了图结构中实体和关系的显式表示，能够高效地获取相关知识，并通过大型语言模型的上下文学习合成答案。

2、图算法的应用

利用图算法，可以进一步优化检索结果和答案生成过程：

- **节点重要性评估**：使用节点中心性等算法，评估节点的重要性，优先检索关键节点的信息。
- **聚类分析**：通过社区发现等方法，将图中的节点分组，获取宏观的知识结构。
- **路径搜索**：寻找两个实体之间的最短路径，揭示它们之间的关系链条。

这些算法有助于处理知识的权重和宏观总结信息，提高答案的准确性和相关性。

3、利用现有的图状数据和知识图谱

在存在现有图状数据或知识图谱的场景下，GraphRAG 可以直接利用这些数据作为知识来源，增强检索和生成能力。

对于数值型的图数据（如社交网络、物流网络等），系统可以：

- **文本到查询转换（Text-to-Query）**：利用大型语言模型，将用户的自然语言需求转换为图查询（如路径计算、最短路径搜索）。
- **代理工具调用（Agentic Tools）**：通过代理式工具，执行相应的图计算和数据检索。
- **结果解释与呈现**：将计算结果以易于理解的形式返回给用户。

对于知识型的图数据（如公共知识图谱、企业内部知识图谱）：

- **本地搜索召回**：直接从知识图谱中检索相关的实体和关系。
- **知识扩展**：利用图谱的连接性，发现与查询相关的更多信息。
- **答案生成**：结合检索结果，生成准确且丰富的回答。

通过利用现有的图状数据，GraphRAG 可以充分发挥这些资源的价值，提高系统的效率和效果。

4、全局与局部问题的处理

- **全局性宏观问题**：如“哪些文章的观点比较独特”，系统会从图上的所有知识聚类的总结中获取信息，作为 RAG 上下文，用于回答全局性问题。
- **局部性问题**：从图上的关键知识点出发，找到相关的知识链条与原始知识块，回答具体的问题。

三）增强生成组件

增强生成组件利用图检索模块提供的信息生成最终的回答或文本，功能包括：

- **上下文融合**：将检索到的信息与用户的问题结合，形成完整上下文。
- **文本生成**：使用生成模型生成自然流畅的回答。
- **质量评估**：检查生成的文本，确保其准确性和一致性。

增强生成组件通过结合图信息，生成更精确和丰富的回答。

3.5.3.3 GraphRAG 的优势

GraphRAG 在检索增强生成（RAG）框架中引入图结构和算法，具有以下显著优势：

1、细粒度知识点的提取：通过构建图状结构，GraphRAG 能够从原始知识中提取细粒度的知识点，如实体、属性和关系。这使得系统在回答具体问题时，能够提供更加精准和详细的答案。

2、深层次关联的挖掘：利用图结构固有的连接性，GraphRAG 可以深入挖掘知识中的深层次关联，发现隐藏的关系链条。这有助于提供更全面的知识视角，支持复杂问题的解答和推理。

3、最佳利用现有图状知识：对于已有的图状数据或知识图谱，GraphRAG 提供了最优的利用方式。无需重新构建，直接将现有的图数据纳入系统，实现资源的高效利用和价值最大化。

4、全局性宏观问题的领先解决方案：在回答全局性宏观问题时，GraphRAG 通过对图结构的全局分析和聚类，总结出整体性的知识概览。相比传统方法，GraphRAG 提供了当前最先进（state-of-the-art）的解决方案。

5、对传统 RAG 的有效补充：传统 RAG 通常采用分块方式处理知识，而 GraphRAG 作为一种自然且有效的补充方法，利用图结构丰富的关联信息，弥补了分块处理的局限性，提升了知识检索和答案生成的质量。

GraphRAG 提供了一种强大的知识索引和检索方法。通过利用图形化的知识表示和检索策略，GraphRAG 能够更好地捕获知识之间的复杂关系，提升检索效率和答案质量。在存在现有图状数据或知识图谱的场景下，GraphRAG 能够充分利用这些资源，为构建智能化、可扩展的知识应用平台奠定了坚实的基础。

3.5.3.4 GraphRAG 的改进策略

最简单的 GraphRAG 方案存在知识抽取困难、知识表示不全、知识召回不准等问题，因此需要综合多种手段改进 GraphRAG 链路，如引入文档结构、图社区摘要、混合索引等。

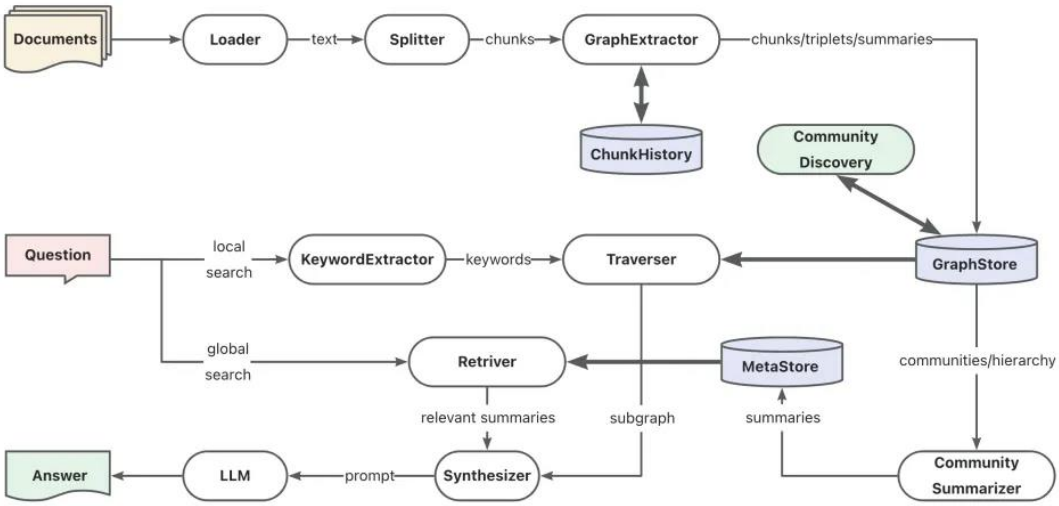

图 3.16 一种改进的 GraphRAG 链路

总的来看，可以从以下几个方面对 GraphRAG 链路进行改进：

1、增强索引：扩充信息来源、增强知识抽取能力，从数量和质量上提升文档索引效果。

2、增强存储：优化知识图谱结构，提升知识库的存储质量和效率。

3、增强检索：支持多样化的知识库信息检索与召回，应对多样化的问答场景。

3.5.3.4.1 增强索引

一）引入文档结构信息

一般的 GraphRAG 链路在处理语料时，首先将文档拆分为文本块，并抽取每块文本的实体和关系信息。然而这种处理方式会导致实体与文档结构之间的关联信息丢失。文档结构本身蕴含了重

要的层级关系，可以为知识图谱检索提供重要的上下文信息。另外，保留文档结构有助于数据的溯源，为问题答案提供更为可靠的依据。

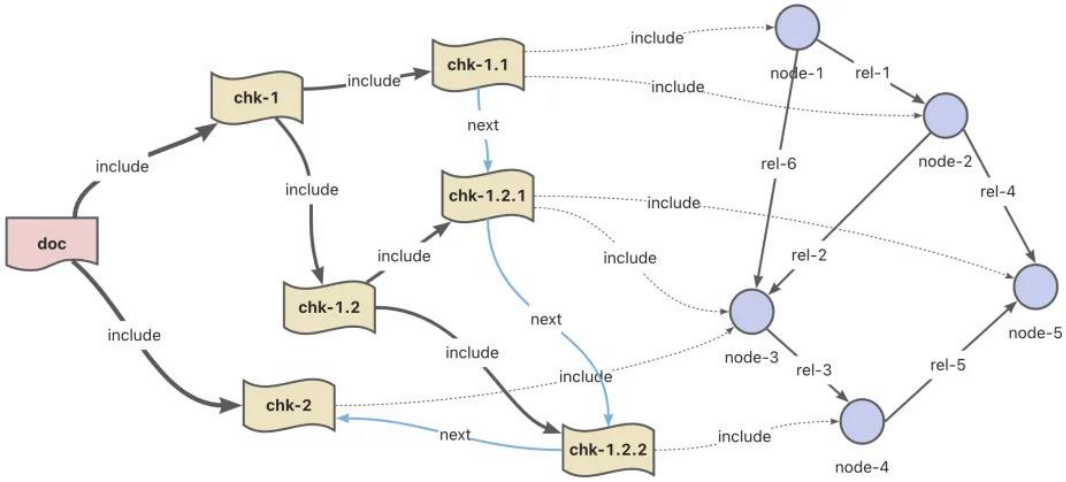

图 3.17 带文档结构信息的知识图谱

二）上下文关联能力

使用大模型进行知识抽取时，完整的上下文对结果有显著影响。在抽取过程中，存储已处理的文本块信息，在调用大模型时将关联度较高的历史文本块作为上下文，连同要抽取的文本块一起提供给大模型。尽管这种方法可能导致一些 token 的浪费，但保留完整的上下文有助于提升了抽取结果的质量。

三）优化知识抽取提示词

当下通用大模型对图数据结构的理解能力还有很大的改进空间，借助大模型工程技术改进大模型对图的理解能力，可以有效地提升知识抽取结果的质量。

通过优化提示词，增强大模型对知识的理解力。从提示词的基本结构出发，可以从以下方面提升知识抽取效果。

- 角色：给大模型设定“知识图谱工程专家”的角色，可以收获意想不到的效果，让输出更加专业、稳定。
- 指令：指示大模型需要完成的任务，如三元组抽取、元素总结等。
- 上下文：向大模型提供任务相关的背景、技能列表等，也可以通过思维链引导大模型做出更细致的处理，比如如何抽取实体、关系，如何使用跨文档关联信息。
- 约束：对大模型的行为设定限制条件，避免不恰当的处理和幻觉，保证稳定性输出。
- 输入：大模型要执行任务的输入，即知识抽取的原始信息，包括待抽取文本和关联性段落。

- 输出格式：指定大模型输出的特定格式，方便后续的解析处理。
- 样本：提供案例样本供大模型参考，提高输出的准确度。

四）知识抽取微调模型

借助于专有的知识抽取微调模型，让特定领域的知识抽取更加高效。比如由蚂蚁和浙大联合研发的大模型知识抽取框架 OneKE 在零样本泛化性能上全面超过了现有模型。

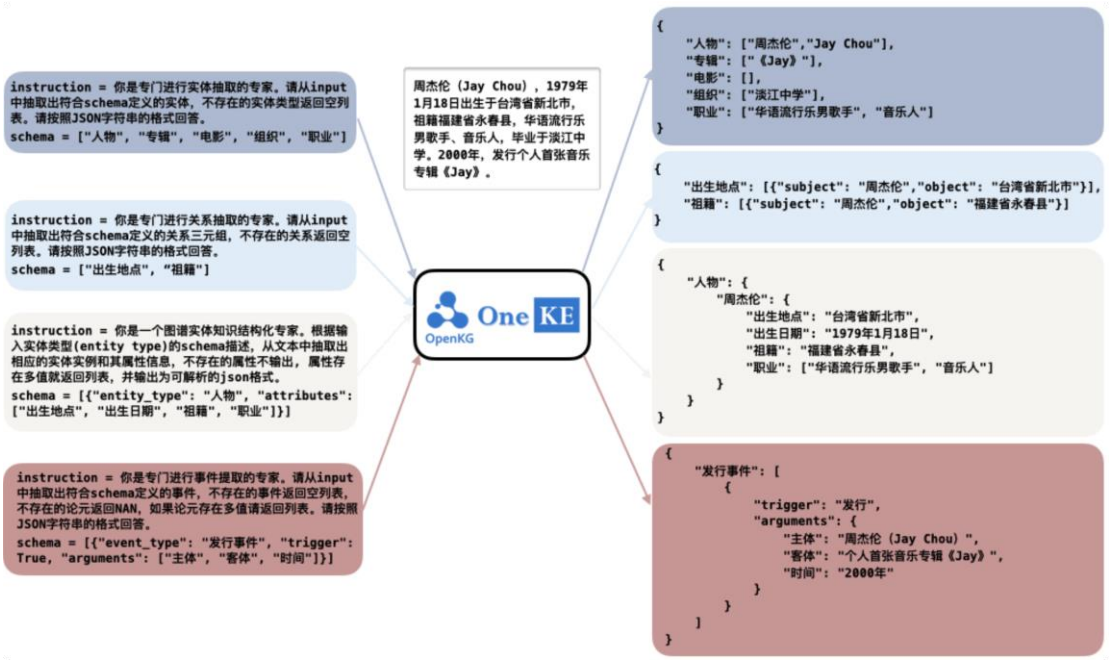

图 3.18 知识抽取模型示例

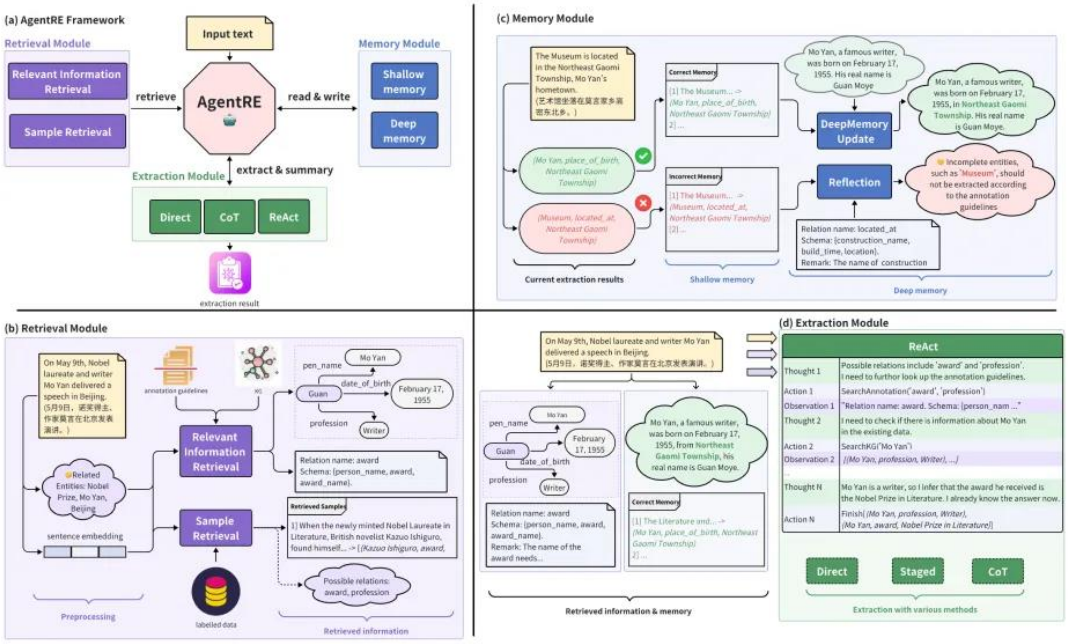

图 3.19 知识抽取智能体[49]

五）知识抽取智能体

借助于智能体引入记忆和反思机制，可以进一步提升知识抽取的准确性。如 AgentRE 框架可以解决在复杂场景中关系抽取面临的关系类型多样、实体间关系模糊等问题。

3.5.3.4.2 增强存储

一）引入高维图特征

受限于大模型本身对图谱的理解能力，直接基于抽取后知识图谱做问答并不一定能获得可靠的答案。为了让知识图谱的数据可以更好地被大模型所理解，借助于图计算领域的技术，为知识图谱赋予更多样化的高维图特征，协助大模型理解图谱数据，进一步改善问答质量。

具体的手段包括但不仅限于：

- 二跳图特征：最直接的图特征计算方式，提供节点的邻居信息，如节点公共邻居、邻居聚合指标等。
- 路径特征：借助于图上路径算法，描述节点间的连通特征，如最短路径、DFS/BFS、随机游走等。
- 社区特征：聚合相似节点集合，描述节点间的同质特征，进一步提供社区摘要，如 LPA、Luvain、Leiden 等。
- 重要性特征：描述节点的重要程度，辅助提取关键信息，如 PageRank、节点聚集系数等。

二）关联原始文档

前边提到的文档结构信息标记了实体的文档来源，但不能具体甄别关系的具体来源。如果需要进一步细化知识图谱中的数据来源粒度，需要在关系上保留具体的来源文档 ID 和文本块 ID。检索阶段时，可以将知识图谱子图中关系边涉及到的文档和文本块详情一并提供给大模型上下文，避免知识抽取过程导致的文档细节内容丢失的问题。

三）支持融合索引

随着大模型技术的兴起，融合索引（Converged Index）能力支持，已逐步成为数据库和大数据产品的重要技术发展路线，当然作为知识图谱的底座，图数据库也不例外。本质上，融合索引可以有效地打通大数据和大模型场景，基于一套数据存储，提供多样化的查询分析支持。

主流的索引格式包括但不限于：

- 表索引：提供传统的关系型数据查询与分析能力，实现基于表数据的过滤、分析、聚合等能力。
- 图索引：提供关联数据分析能力以及图迭代算法，实现基于图数据的高维分析与洞察。

- 向量索引：提供向量化存储与相似性查询能力，扩展数据检索的多样性。
- 全文索引：提供基于关键词的文档查询能力，扩展数据检索的多样性。
- 其他：例如多模态数据的索引，如图片、音频、视频等。

四）存储格式增强

在图数据库层面支持更多样化的存储格式，可以为知识图谱提供更友好的交互界面和更高的查询存储性能。

首先是“弱 Schema”能力，即无需事先声明图谱结构，允许上层应用随意的修改图谱数据。这样的交互方式，尤其是对事前不能确定图谱结构的知识图谱构建任务来说非常重要，用户可以根据自己的需要抽取三元组中的实体和关系数据，无需做特定的格式转换便能写入图数据库。另外，很多图算法的结果图特征一般需要通过临时字段写回原图，以便对图谱进行信息增强，尤其是对实现无法确定使用哪一类图算法的应用场景，弱 Schema 能力可以提供未声明的字段的及时更新。

然后是“多模态”能力，随着大模型技术的不断演进，多模态大模型、多模态 RAG、多模态知识图谱的场景也相继出现。现有的图数据库大多数还是使用二进制类型存储图片、音频、视频等数据，对查询性能有极大的影响，因此对多模态数据索引格式的支持，也是图数据存储格式亟待改进的方向。

3.5.3.4.3 增强检索

一）支持混合检索

朴素意义的混合检索可以理解为对多种存储系统的并行多路召回，例如同时根据用户的查询进行向量数据库的相似性召回、知识图谱的检索。这样做虽然实现方案比较简单，但是存在多路召回的数据结果不相干，甚至矛盾的情况。这是因为原始数据是通过不同的格式多写到异构的存储系统，天然存在不一致性。而借助于融合索引的混合检索则不会出现类似问题，用户可以基于向量相似度或者关键词直接召回知识图谱中的子图结构，从而保证的数据语义的一致性。

除了并行多路召回，多种存储系统也可以彼此辅助，构建更复杂的检索方案。例如待社区摘要的知识图谱，可以通过图数据库实现知识图谱明细子图的召回，同时通过向量数据库提供社区摘要的相似性召回，为问答提供更完备的上下文信息。

二）自然语言查询

基于自然语言查询中关键词的知识图谱召回，只能做粗粒度的检索，无法精确地利用查询文本中的条件、聚合维度等信息做精确检索，也无法回答不包含具体关键词的泛化查询问题，因此正确地理解用户问题意图，并生成准确的图查询语句就十分有必要（参考章节 3.8 的内容）。而对

用户问题的意图识别和图查询生成，最终都离不开智能体解决方案。大多数情况下，我们需要结合对话的环境和上下文信息，甚至需要调用外部工具，执行多步推理，以辅助决策生成最理想的图查询语句。

三）多跳推理能力

多跳推理能够更好地应对指导手册类的文档问答需求。Graph 尽管相较于 Embedding Vector 能够更好地提取和存储文档中的思维链（CoT）和行动链（CoA），但是目前还没有完全适配检索思维链和行动链的算法。为了解决这个问题，可以借助多跳推理及相关的算法满足我们的需求。

1、起始点选择：

- 分析查询，找出相关的知识图谱节点（可能多个）作为起点。

2、路径探索：

- 从起点开始，在图中进行有限深度的搜索（通常为 N 跳）。

3、记录探索过程中发现的路径。

4、路径评分：

- 对每条路径进行评分，考虑因素包括：
  - 与查询的相关性
  - 路径长度
  - 节点的重要性

5、最佳路径选择：

- 选择评分较高的少数几条路径。

6、上下文生成：

- 将选中的路径转换为自然语言描述。

7、与语言模型集成：

- 将生成的描述作为上下文提供给大语言模型。
- 引导模型使用这些信息来回答原始查询。

总之这个算法解释了，如何利用图结构来增强大语言模型的理解和推理能力，特别是在处理需要多步思考的复杂问题时。

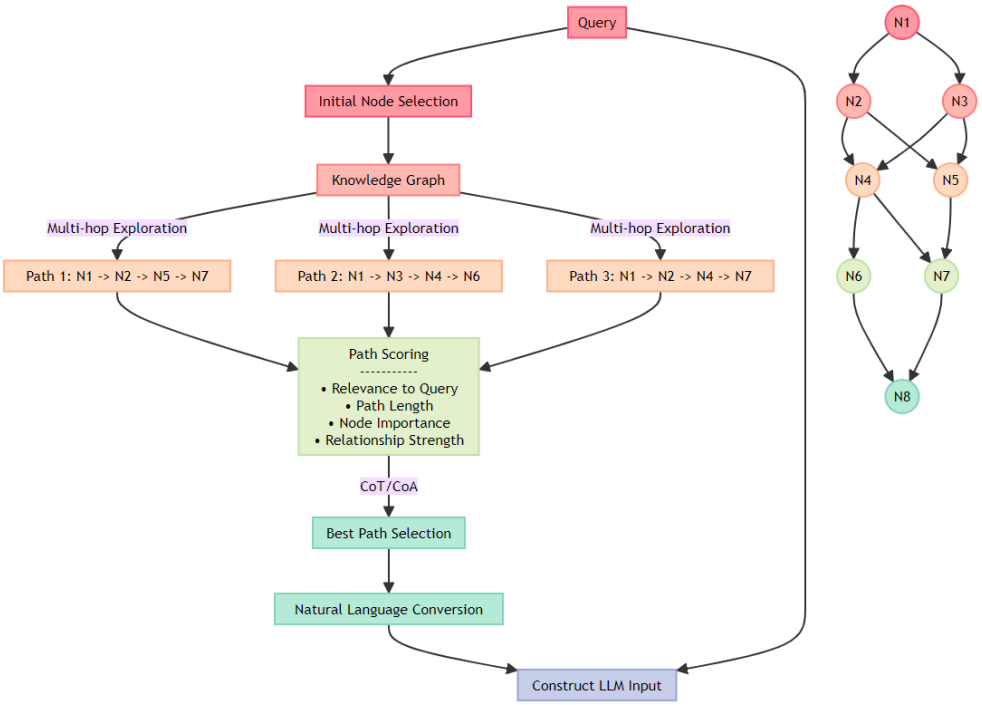

图 3.20 多跳推理检索

四）性能测试

传统的 RAG 有一些比较成熟的性能测试方案，如 RAGAS、ARES、RECALL、RGB、CRUD-RAG 等。截至目前为止，GraphRAG 仍缺乏合理有效地性能测试方案。为了让 GraphRAG 的优化方案更有的放矢，构建完备的 GraphRAG Benchmark 测试是当下亟待解决的问题。

3.5.3.5 未来展望

3.5.3.5.1 技术挑战

1、当前 GraphRAG 面临的技术瓶颈

随着现代社会中信息的爆炸式增长，各类新兴实体、关系与事件不断涌现，现有知识可能面临着过时的问题。在这种背景下，知识图谱的准确性与及时性受到了挑战，亟需更加健壮的知识图谱动态管理与维护策略。

当知识图谱的规模逐渐扩大时，从图谱中检索不同粒度信息（如实体、路径等）的难度也随之增加，针对大规模图数据的高效检索算法仍有待研究与开发。

图数据的质量和一致性也是一个持续的挑战。图中可能存在错误数据或不一致的关系，这会直接影响模型生成内容的准确性和可靠性。

此外，在工业生产中，知识源的表现形式往往并非单一，其通常涉及文档、三元组、表格乃至音频、视频等不同模态的数据。无法有效地整合与利用异质的知识类型可能成为制约 GraphRAG 性能上限的一大瓶颈。

2、潜在的解决方案与研究方向

为了更有效地管理动态的知识，可以从知识图谱的架构入手。例如，时序知识图谱为每个事实三元组标记了生效的时间范围，能反映事件随时间的演化性，这一特性使其适合作为 GraphRAG 的底层模式。

而对于涌现的大量知识，可以尝试知识过滤、去重等方式减小新增知识图的规模，更新过时的知识。

为缓解知识数量增长引起的检索困难，需要设计分块、并行、混合的检索策略，以稳定整体性能。

进一步地，对检索得到的不同格式、模态信息，可以针对下游目标为模型设计合适的指令微调任务，以此缩小不同知识源之间的语义鸿沟。

3.5.3.5.2 发展方向

1、GraphRAG 在更广泛领域的应用前景

GraphRAG 已在通用领域中的多个任务（如开放域问答、推荐系统等）上证明了自身的潜力。近期的工作在医疗、金融等场景上取得了值得关注的进展。构建领域相关知识库，使算法赋能各类细分领域将成为未来的研究与落地热点。

2、图数据和生成模型的进一步融合

结构化的图数据与输入生成模型的非结构化文本在语义空间中存在一定的偏差。为缓解这一困难，一些工作侧重于将图数据转化为非结构化形式，如三元组、描述文本、代码等，使之与生成模型兼容。另一些工作则利用图神经网络对图数据进行编码，通过注意力、前缀调优等方式融合文本特征与图特征。针对图数据与生成模型探索更优的融合策略是提升 GraphRAG 整体性能的重要方向。

3、下一代图增强生成系统的展望

下一代图增强生成系统应具有安全、透明、可解释的特点。以问答场景为例，对检索结果中可能存在的知识冲突或无关内容，系统需要正确地识别检索返回的有用信息，并在知识匮乏时要求重新检索或拒绝回答。当系统做出回答时，应当能同时给出相应的推理路径与思考过程，以此提升回答的可信度。

3.5.4 智能体

在大模型出现之前，智能体的研究主要用作策略函数，解决一些具体场景中的问题，一般都是针对某个具体的任务在隔离环境中进行。直到大语言模型发布，智能体具备自主思考与决策的能力，智能体的研究与发展出现井喷式爆发发展。尤其是 AutoGPT、MetaGPT、AutoGen、ChatDev 等项目与框架的出现，智能体的研究与应用浪潮被推到一个崭新的高度。

尽管大语言模型已经具备了一定的思考与决策的能力，但将其与现实世界打通，具备与跟实际环境交互的能力，初步具备类人的自主工作的能力，还需要很多工作要做，包括角色、记忆、思考规划以及行动等。

为了弥补 LLM 和自主智能体之间的差距，需要围绕模型构建一套自主智能的架构，而这里关键的步骤智能体的架构的设计。基于以上的背景，我们在 DB-GPT 框架当中，设计了一套结合 TuGraph 的数据驱动的多智能体协作框架，可以更好的结合多种数据进行智能体的构建与协作，主要包含一下核心特性:

- 丰富的记忆支持: 包括感知记忆、短期记忆、长期记忆、混合记忆等。
- 支持多种协作模式: 支持固定编排、动态规划、预编排等。
- 易集成: 兼容主流的开源智能体协议，可以快速被其他智能体框架集成
- 数据驱动: 智能体的思考、规划、行动等环境都受到 DataFrame 上下文的驱动与约束。

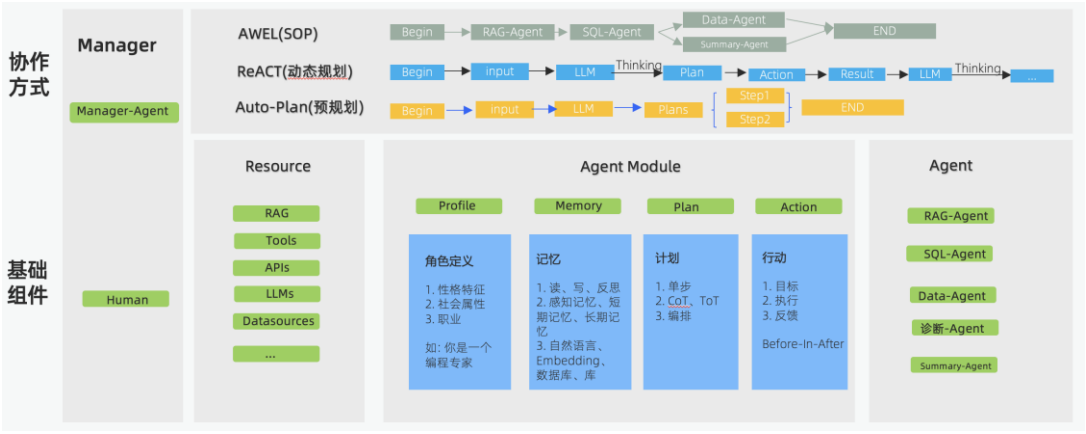

图 3.21 多智能体协作框架

3.5.4.1 Profile

一个完整的智能体至少包含以下四个核心模块: 1. Profile、2. Memory 3. Plan 4. Action。Profile 模块在智能体中主要用来做角色认定，通常通过 Prompt 方式来进行指定，通过角色认定可以影响大语言模型的行为，智能体的 Profile 通常会说明其年龄、职业、心理、社会关系等。角色认定是一种重要的社会和组织机制，通过明确个体在特定系统或环境中的地位和职责，有助于维护秩序、

提升效率和促进合作。在实际应用中，描述智能体信息取决于业务场景。一个完整的 Profile 设定需要具备以下方面的信息:

a）命名: 即每个智能体的名称代号，如 dbgpt

b）角色: 设定智能体的角色定义，如 Reporter

c）目标: 设定智能体的目标

d）性格、社会关系等约束条件设定, 如“你只负责收集和整理历史消息中已经存在的分析 GQL，不自行生成任何分析 GQL”

3.5.4.2 Memory

Memory 即智能体记忆模块，主要用来存储、获取、检索信息。在记忆存储格式上，不仅支持自然语言、Embedding、关系存储这样简单的格式，还支持复杂的图存储。基于丰富的存储格式，进一步支持了多种记忆结构，如感知记忆、短期记忆、长期记忆、混合记忆等。基于记忆模块，智能体可以具备长久的记忆，在降低幻觉的同时，可以进一步完成自我进化，完成更复杂的任务。同时基于图的记忆，在复杂关系识别，反思等方面相比简单记忆有更好的表现。

图 3.22 智能体记忆

3.5.4.3 Plan

人类在面临复杂任务时，人类倾向于将其构造为简单的子任务并且独立进行解决。Plan 模块的目的是通过类人的能力，让智能体的行为更具逻辑性、更强大、更可信。除了 CoT、ToT 之外，我们进一步发现基于图的计划更符合复杂任务的拆分逻辑，且更容易表达清楚任务之间的协作关系。

3.5.4.4 Action

Action 模块负责将智能体的决策转化为具体的结果。此模块一般直接与环境进行交互，同时受 Profile、Memory 和 Plan 模块的影响。

3.5.4.5 智能体工作流

我们基于 DAG（有向无环图）设计了一套智能体工作流表达式语言 AWEL（Agentic Workflow Expression Language）。基于 DAG 的智能体工作流可以确保智能体之间依赖关系与协作的正确性，同时基于 DAG 高效、强拓展、容错等能力，让智能体的编排更灵活。

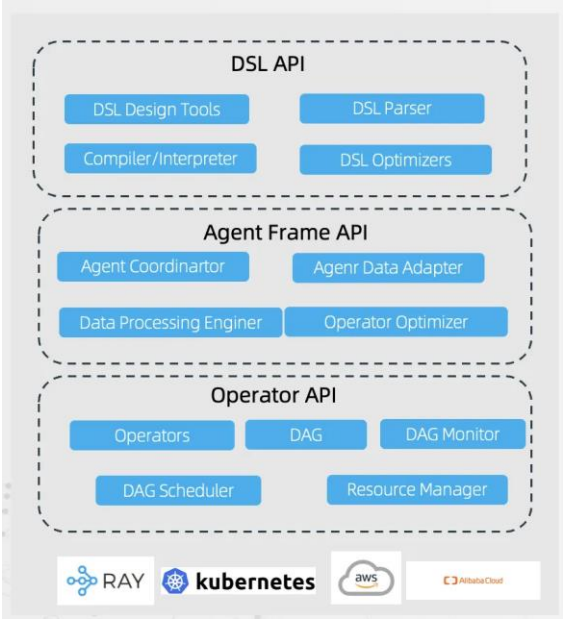

图 3.23 智能体工作流架构

DAG Tree

```
HttpTrigger(node_id=b71f0ccc-4539-445e-8d87-1ec5ef6e7e83)
 -> MapOperator(node_id=23d11ff8-a1d4-4541-a1ff-adfb0eadd614)
   -> ChatHistoryPromptComposerOperator(node_id=3ee4c2af-d625-445b-a0f3-60863404d82e)
     -> LLMBranchOperator(node_id=bdc20002-b9e9-4ce8-9211-5aba83f4c16e)
       -> LLMOperator(node_id=b4c3fe48-a9d8-4968-a128-df5a6b6edd4c, node_name=llm_task)
       |  -> MapOperator(node_id=ce768e92-5299-43ab-ac04-09755b5f19e2)
       |    -> JoinOperator(node_id=f33f239e-90dd-4dd4-bc12-30d6d0912cfb)
       -> StreamingLLMOperator(node_id=ca529d57-1bd7-479c-9047-5cb5512ade8e, node_name=streaming_llm_task)
         -> OpenAIStreamingOutputOperator(node_id=1952037e-775a-4c9c-a8c3-7ca678f15871)
           -> JoinOperator(node_id=f33f239e-90dd-4dd4-bc12-30d6d0912cfb)
```

DAG graph

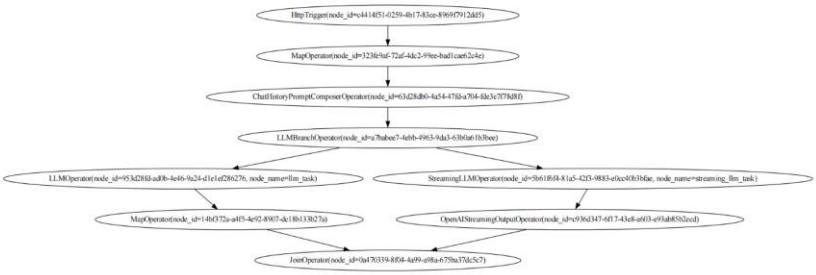

图 3.24 智能体工作流样例

3.5.4.6 多智能体协同

多智能体协同目前主要有以下几种方案 1. 基于 SOP 编排的协同 2. 基于 ReACT 动态规划协同 3. 基于大语言模型 Auto-Plan 的预编排的协同。 针对不同的场景可以选择适合的协同方案，如在确

定性的严谨场景下，基于 SOP 的协同可以获得更准确、更严谨的结果。 在开放性场景下，Auto-Plan 与 ReACT 可以充分发挥模型的思考能力，在解决泛化能力上也更具优势。 我们结合 AWEL、Auto-Plan、图实现的多智能体协同，可以满足各类场景的诉求。在满足更复杂的协同关系的同时，智能体之间的系统效率也有出色的表现。

3.5.4.7 图方案生成

图数据库作为近年来崛起的新型技术，尽管在处理关联关系查询时展现出卓越的性能，但由于缺乏丰富的案例和方案，使用门槛较高，导致其无法像关系型数据库那样广泛普及。这一挑战使得许多用户在上手图数据库时面临困难。

随着大模型的出现，我们可以借助其强大的智能能力来降低图数据库的使用门槛。在不熟悉的领域，它能为我们提供构图建议；在进行图数据分析时，它可以协助编写查询语句；同时，在实际使用中，帮助快速查找具体操作方法。这些功能有效地降低了用户对图数据库的上手难度。

然而，目前大模型仍存在一定的随机性，其在处理多步骤任务时的理解与执行，常常存在不确定性。为了解决这个问题，我们需要将任务分解得更加精确，以帮助大模型更好地理解我们的意图。这也是单智能体概念的提出初衷：我们为单个智能体设定独立且明确的任务，从而提高其执行效率。例如，我们可以构建一个智能建图模型的智能体，其输入为用户需求，输出为标准的图模型构建建议。

但是，完整的使用流程通常包括多个子任务，因此需要多个智能体的协调配合才能完成。在这一过程中，多智能体的协作成为关键。通过不同智能体间的协作，例如构图智能体、数据模拟智能体、数据分析智能体、方案总结智能体和文档解答智能体的串联，我们能够快速针对图数据库的上手难题提供有效解决方案。

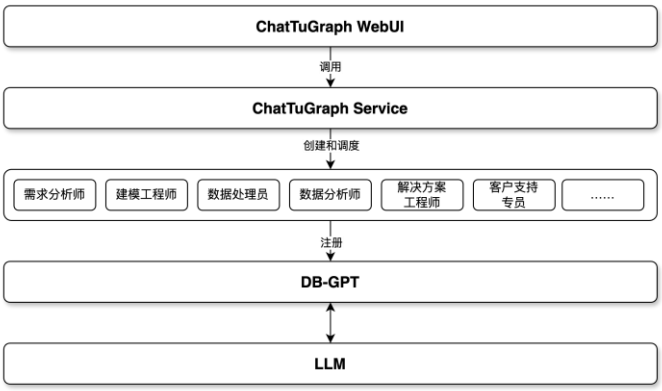

图 3.25 图数据库智能体设计

通过这样的多智能体应用，我们不仅能提升图数据库的使用效率，更能在实际操作中沉淀出针对性的解决方案，帮助用户在图数据库的应用中取得更好的成果。

第 4 章 解决方案

4.1 基于图数据库+AI 的申请反欺诈解决方案

对于零售信贷的反欺诈来说，传统方式有两种：

- 一种是常用的反欺诈规则，通过历史案件总结下来的专家知识，对异常欺诈行为建立规则模型，用规则模型进行欺诈行为的特征描述，帮助业务将欺诈行为和正常借贷行为区别开来。
- 另一种方式则是通过历史的欺诈申请进件和正常的申请进件数据进行机器学习模型建模，利用数据挖掘的手段从高维空间中筛选出异常欺诈申请进件。

无论是专家规则还是机器学习建模，都是基于个体特征的分析，随着欺诈黑产技术的演变，有组织的团伙欺诈行为越来越多，通过对风险的分散，传统方式难以识别，存在以下问题：

- 欺诈手段呈现多样化、专业化、团体化等特征，传统的专家规则和机器学习模型对于通过多层关系进行掩饰的复杂欺诈手段或者团伙欺诈难以识别。
- 统计模型或者机器学习模型更多的是针对独立个体的分析挖掘，忽略了在欺诈行为中复杂的关联关系导致难以发现行为相对稀疏的个体。
- 信息割裂，没有统一的框架和视图描述客户的全生命周期。各个业务环节的数据之间缺少必要的逻辑视图和交叉校验。

基于图数据库+AI，可以从以下几个方面提升申请反欺诈效果，从事后分析提升到事中分析，从个体分析到复杂关联分析：

- 基于一张图的多源数据融合：通过统一数据语义视图，打破信息割裂，为风控进行更多维度的背景信息真实性核验。
- 基于社群的风控策略：通过各类图算法，加强对可疑黑产中介团伙的分析与识别。
- 补充关系特征维度：通过图特征弥补传统机器学习模型只能学习个体统计特征的短板，加入实体之间的关联关系，提高模型的泛化能力，增加复杂欺诈案件的识别。
- 可视化增强交互式探索：关系网络可视化探索功能和图智能分析算法相结合加强对复杂案件下的风险主体筛查能力，帮助反欺诈人员深入分析主体之间复杂的深度关系。

基于以上方案背景，提出以下图数据库+AI 的技术架构解决方案：

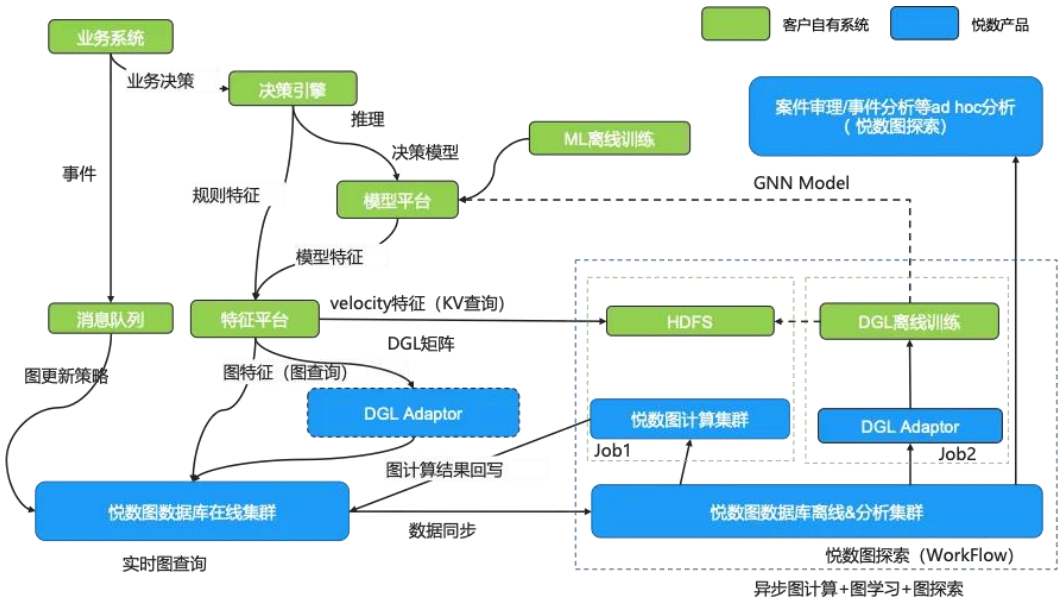

图 4.1 基于图数据库+AI 的解决方案

该架构可将实时/准实时/离线多种技术场景融合，支撑在申请反欺诈场景中的各类业务需求：

- 在线集群：处理高并发写入，高并发查询业务。
- 离线&分析集群：处理异步图计算等复杂计算业务；支撑全图计算和图学习；支持分布式图计算，针对全图的计算可以在闲时按需拉起。
- 两套引擎，一套存储，同时支持实时业务和离线分析业务。
- 通过集群同步功能将数据从在线集群实时同步到离线&分析集群。

业务流程如下：

- 当一笔进件产生时，通过消息队列将数据实时更新到悦数图数据库在线集群中。
- 决策引擎调用原有模型平台，机器学习模型可离线训练，同时调用模型特征、图特征、velocity 特征。
- 特征平台发起实时图特征查询，调用悦数图数据库在线集群，毫秒级时延返回结果。
- 悦数图数据库在线集群将数据实时同步至悦数图数据库离线&分析集群。
- 悦数图数据库离线&分析集群可发起多个子图/全图图计算任务，运行图算法，秒级/分钟级时延返回结果，后续作为实时图特征返回。
- DGL 离线训练按需调取悦数图数据库离线&分析集群数据。
- 业务人员按需使用悦数图探索进行案件审理/事件分析等 ad hoc 分析。

4.2 基于关联分析的企业决策智能化解决方案

基于关联分析的企业决策智能化解决方案旨在为企业提供智能化的数据关联分析决策支持，通过先进的图数据处理方法和持续的模型管理能力，帮助企业在复杂的商业环境中迅速作出明智决策。通过标准化的技术路径和框架化的模型构建，确保数据驱动的决策过程高效、准确且具备可操控性。

目前，许多企业在进行数据分析时面临多重挑战：

1、模型的维护难度较大：传统模型难以进行持续监控和管理，可能导致性能下降，无法灵活应对业务变化。

2、结果难以解释：模型的输出缺乏充分的可解释性，使得决策者在理解和应用结果时遇到障碍。

3、数据源的多样性：在复杂多样的数据源环境中，统筹管理和运行现有模型的技术难题显著增加。

针对这些挑战，本方案提出如下的关键特性与功能：

- 基础模型服务能力：通过统一的范式和框架，支撑不同场景下的图模型应用。
  - 全流程梳理：基于 TuGraph 系统，涵盖样本数据的管理、图数据处理、IDE 实现建模流程的标准化；
  - Graph AI 框架：利用图结构进行建模，适用范围广泛，包括：
    - 属性传导：应用于高净值客户识别、反洗钱和反欺诈等场景。
    - 向量化匹配：有效进行产品及信息推荐。
    - 因果推理：帮助定位业务系统告警的根因及进行舆情预警。
    - 时序图分析：用于债基风险预警以及客户社交网络的分析。
  - 标准建模流程：流程包括梳理业务问题、进行数据清洗、构建图结构及搭建图模型，旨在降低成本并提高可预测性。
- 模型持续管理能力：应对传统模型在持续管理方面的缺陷，解决模型衰减、难以解释以及技术环境多样化等问题，借助 TuGraph 的数据模型，统一运行环境、整合数据源。
  - 预测数据源的有效管理：确保数据源的质量和可靠性，通过定期审核和更新来优化数据收集和处理流程，以提高模型预测的准确性。
  - 图模型的持续监控：实时监测图模型的表现和趋势，通过指标跟踪及时发现问题，确保模型的有效运行和及时调整。

- 完备的图模型预测服务：构建全面的预测系统，集成多个图模型，并提供一致的输出接口，便于用户进行多维度的决策支持。
- 预测结果的可解释性：采用可视化和说明性工具，使用户能够理解模型的决策过程，提升透明度，增加用户对模型结果的信任。
- 模型的干预及权限管理：建立权限控制机制，确保只有授权者可以进行模型调整和干预，保护模型的安全性和稳定性。
- 优化模型资源的配置：通过资源监测和分析，动态调整计算和存储资源分配，提高模型运行的效率和效果。

- 互联互通能力：解决方案的核心能力与其他业务系统或平台无缝对接，确保智能化应用（如业务系统和数据）和基础数据服务之间的流畅互通，促进数据资产高效流通和利用。

通过上述的技术框架和管理能力，本方案可以为企业提供可靠的分析决策支持，包括：

- 提升决策效率：通过智能化的数据关联分析，帮助企业快速获取、解读和运用数据信息，做出更准确的决策。
- 降低运营风险：通过有效的模型管理与监控机制，提高模型的可持续性，降低因模型老化带来的潜在风险。
- 增强业务洞察力：基于图数据的多维关联分析，帮助企业挖掘潜在客户及市场机会，为企业战略决策提供有力支持。

4.3 基于图算法分析的安全风控解决方案

为解决安全风控中常见的图数据处理、模型训练评估和风险分析算法应用等问题，蚂蚁集团建设了基于图算法分析的安全风控解决方案，该方案通过关系视角来描述风险，并利用全面的风险数据构建风险关系网络，从而形成风控知识图谱，实现了风控全链路的图数据应用。依托蚂蚁集团的 TuGraph 图数据管理平台，集成了图特征、图算法和图组件，打造出一体化的图运营平台。通过图数据来描绘复杂的风险模式，利用图计算进行实时的风险防控，并通过图应用实现大规模的风险管理。

图 4.2 全图风控产品整体架构

其中，图算法部分已沉淀成为一套算法框架 GeaSec（Graph extended analysis for Alipay Security），可以快速、有效、准确地使用图算法解决业务问题。GeaSec 由大规模图上的异常检测工具 GAD tools（Graph Abnormal Detection tools）和图风控神经网络算法库 GREAT（GNN based Risk Exploration Algorithms using Torch）组成。

- 大规模图上的异常检测工具 GAD tools：帮助用户更好地发现潜在的风险节点，提高风险控制的准确性和效率。该工具包含节点异常、链路异常和群组异常等类型的异常检测方法，可以帮助用户快速准确地发现潜在的风险节点、识别资金销赃交易并发现潜在的风险链路、更好地发现和控制潜在的风险团伙和社区。作为一种针对大规模图的异常检测算法工具，具有非常强大的功能和应用价值。它能够帮助用户快速准确地发现潜在的风险节点、链路和群组，提高风险控制的准确性和效率，在金融、社交等领域的风险控制业务中有着广阔的应用前景。

- 图风控神经网络算法库 GREAT：是一个基于 Torch 的风险图算法库，它提供了一整套流程，包括数据准备、特征预处理、环境配置、模型训练评估和线上部署，使得图算法的使用门槛降低，自研算法的开发效率提高。为了满足安全风控场景的需求，GREAT 打造了一系列具有风控特征的图算法，包括对比学习、流式动态图、子网络发现、图拓扑结构表征、预训练、少样本的图异常检测、图预计算等方案。这些算法可以解决风险交易方向的重要性问题、资金流转与上下文交易关系发现、图模型预测可解释要求与可信 AI 建设；GREAT 还提出了图对比学习方法、图预训练方法以及图上异常检测等解决方案，解决高分团伙提纯与低浓度灰团伙召回需求、消息传播机制对图拓扑结构特征提取能力较弱的问题、图近线部署推理成本较高以及子图点边冗余等。它可以帮助用户更好地发现和控制潜在的风险，提高风险控制的准确性和效率。同时，GREAT 还支持快速复用，可以帮助用户快速开发自己的图算法解决方案。

4.4 图异常检测智能化解决方案

本节介绍一种基于创邻科技 Galaxybase 图数据库的图异常检测智能化解决方案。在现代数据驱动的商业环境中，随着数据量的快速增长和关系复杂性的提升，异常行为往往隐藏在复杂的关联网络中。这在金融反欺诈、网络安全和供应链管理等关键领域尤为明显。当异常行为发生时，可能带来以下问题：

- 经济损失：在金融领域，异常交易可能导致大规模资金损失，直接影响企业盈利能力。
- 数据泄露：网络安全中的异常活动可能导致敏感信息被窃取，进而危害用户隐私和企业机密。
- 运营中断：在供应链管理中，异常行为可能引发物流延误，导致生产停滞和客户订单无法及时交付。

针对这些问题，创邻科技基于 Galaxybase 分布式图数据库强大的 HTAP 能力，设计了一套异常检测解决方案。通过对数据中的节点、边及其关系进行深度分析，该方案能够识别复杂的异常模式。以下是该方案的具体步骤：

- 多维关联数据融合：通过图计算引擎，将多源数据整合为全局图谱，构建完整关联网络。全局视图不仅包含节点特征，还反映节点间复杂连接，系统能在更大范围识别异常。
- 图算法驱动的异常检测：利用图算法识别网络中的异常节点和连接。例如，PageRank 算法衡量节点重要性，识别异常高或低连接节点；标签传播算法通过已知异常样本传播风险标签，识别潜在异常节点；社群检测算法识别不符合常规行为的孤立社群，适用于团伙欺诈检测。
- 时序图分析与动态监控：时序图分析应对动态数据，捕捉节点关系随时间变化趋势，实时检测异常行为。结合动态更新机制，系统可持续监控并动态调整模型参数，提高适应能力。
- 结合 Graph AI 模型异常检测：系统通过整合图模型的分析结果与异常检测模型，充分利用图数据中的关系和结构信息。通过扩展输入数据的特征维度，系统能够更全面地捕捉复杂模式和潜在异常，从而提高检测的准确性和稳定性。
- 可视化与交互式分析：提供图谱可视化功能，通过交互式图展示异常节点、社群及其关系，帮助决策者直观理解异常来源和影响，提高检测可解释性，助力业务人员深入分析潜在风险。

该解决方案可适用于多种场景：

- 金融反欺诈：分析资金流动、设备共享等复杂关联，检测团伙欺诈。

- 网络安全：基于设备通信网络、用户行为图分析，识别网络攻击和恶意节点。
- 供应链管理：在多方关系网络中，检测异常物流路径和可疑供应商行为，降低风险。

4.5 Graph 驱动的检索增强生成技术解决方案

本节介绍一种基于悦数 Graph 驱动的检索增强生成（Retrieval Augmented Generation，RAG）系统的技术方案。该方案旨在构建一个代理式（agentic）的 RAG 知识库管理与知识推理应用平台，利用悦数 Graph 的分布式、云原生特性，以及其对千亿点、万亿边规模图数据的支持，实现高性能、多租户和多索引能力。系统支持用户根据不同用途和特征来管理知识，并针对不同类型的知识文档采用不同的索引方式，提升了系统的通用性和可扩展性。

4.5.1 系统概述

传统的 RAG 系统主要依赖于文本检索和向量检索技术，而本方案通过引入悦数 Graph，利用其强大的图数据库能力，支持更复杂的知识表示和推理。悦数 Graph 是一款高性能的分布式云原生图数据库，支持千亿点、万亿边的存储和查询，具备多模型支持和高度可扩展性。通过整合 BM25、向量索引和悦数 Graph 的图索引等多种索引方式，系统能够有效地组织和检索多类型的知识文档，为多样化的应用场景提供支持。

4.5.2 知识管理与索引策略

4.5.2.1 按用途和特征管理知识

系统允许用户根据知识的不同用途（如故障排除、产品推荐、知识问答、研发辅助）和特征（如结构化、非结构化、半结构化）来分类和管理知识。得益于悦数 Graph 的多租户支持，不同的用户和应用可以在同一平台上独立管理各自的知识库，确保数据的隔离和安全。这种灵活的管理方式有助于提高知识的组织效率和检索准确性。

4.5.2.2 多类型索引方式

针对不同类型的知识文档，系统采用以下索引方式：

- **BM25 索引**：适用于传统的文本检索，主要针对结构化和半结构化的数据。悦数 Graph 支持 BM25 索引，提供高效的关键词匹配检索。
- **向量索引**：将文本转换为向量形式，适用于语义检索，能够捕获文本的深层语义关系。悦数 Graph 支持向量索引，能够高效地进行相似度计算和近似最近邻搜索。

- **图索引**：利用悦数 Graph 的分布式存储和查询能力，构建大规模的知识图谱，实现超大规模图数据的高效索引和检索。同时，悦数 Graph 基于 ISO-GQL 扩展的算法支持，使得各种 Graph RAG 的索引和召回非常高效、灵活。

4.5.3 知识索引的调用与处理

当代理（agent）调用知识索引时，会根据需求访问不同的知识及其对应的索引方式。对于同一知识的多个索引，系统采取多路并行召回的策略，最终通过合并或重新排序（rerank）对答案或上下文进行后处理。具体而言：

- **多路并行召回**：同时从 BM25 索引、向量索引和悦数 Graph 图索引中获取相关结果。悦数 Graph 的高性能查询和高并发支持，确保了检索过程的效率。
- **结果合并与重新排序**：利用特定的算法或策略，将多种索引方式的结果进行融合，提升答案的准确性和相关性。

这种方法充分利用了不同索引方式的优势，确保了检索结果的全面性和精确性。

4.5.4 图状知识的召回策略

针对图状知识的召回，系统根据不同的召回策略和用户意图进行区分，主要分为全局性问题（Global Question）和局部性问题（Local Question）。此外，系统还支持利用现有的图状数据作为 RAG 的知识来源。

4.5.4.1 全局性问题

对于全局性问题（Global Question），如“哪些文章的观点比较独特”，系统会：

- **知识聚合**：利用悦数 Graph 的图聚类（Graph Cluster）功能，从知识图谱中的所有知识聚类提取总结信息。
- **上下文构建**：将这些总结作为 RAG 的上下文，提供宏观层面的答案。
- **全局分析**：利用大型语言模型对汇总的内容进行分析，生成综合性的回答。

4.5.4.2 局部性问题

对于局部性问题（Local Question），系统会：

- **关键节点定位**：从用户的问题出发，利用悦数 Graph 的高效查询，定位知识图谱中的关键知识点。
- **知识链条构建**：沿着图谱关系，利用悦数 Graph 基于 ISO-GQL 扩展的算法，找到相关的知识链条和原始知识块。

- **深入回答**：提供针对特定主题的详细答案，满足用户的细粒度需求。

4.5.4.3 利用现有图状数据

数值型图数据

对于数值型的图数据（如社交网络、物流网络等），系统可以：

- **文本到查询转换**：利用大型语言模型，将用户的自然语言需求转换为悦数 Graph 的查询语言（nGQL）语句。
- **代理工具调用**：通过代理式工具（agentic tools），在悦数 Graph 中执行相应的图计算和数据检索。
- **结果解释与呈现**：将计算结果以易于理解的形式返回给用户，得益于悦数 Graph 的高性能和高并发支持，确保了实时性。

知识型图数据

对于知识型的图数据（如公共知识图谱）：

- **本地搜索召回**：直接在悦数 Graph 中检索相关的实体和关系，利用其高效的索引和查询能力。
- **知识扩展**：利用图谱的连接性，发现与查询相关的更多信息。
- **答案生成**：结合检索结果，生成准确且丰富的回答。

4.5.5 知识应用与工具集成

被索引的知识作为代理式知识应用的查询来源，支持在用户描述的应用场景下进行工具调用和组合。具体表现为：

- **故障排除**：结合悦数 Graph 的图计算能力和多索引召回，快速定位问题根源，提供精确的解决方案。
- **产品推荐**：利用用户偏好和行为数据，结合悦数 Graph 的关联分析功能和向量索引，推荐最适合的产品或服务。
- **知识问答**：通过多索引的并行召回和上下文合并，借助悦数 Graph 的高并发支持，提供准确且全面的答案。
- **研发辅助**：为研发人员提供相关技术资料，利用悦数 Graph 的大规模数据处理能力，支持创新和问题解决。

4.5.6 悦数 Graph RAG 的优势

通过引入悦数 Graph，系统具备以下独特优势：

1、超大规模数据处理：悦数 Graph 支持千亿点、万亿边的存储与查询，满足超大规模图数据的需求，适用于处理海量的知识数据。

2、高性能与高并发：得益于高度优化的存储和索引机制，以及分布式架构，悦数 Graph 能够在毫秒级响应复杂的查询，支持高并发的访问，提升系统的整体性能。

3、企业级多租户支持：悦数 Graph 作为分布式云原生图数据库，天然适合企业级多租户 RAG 系统。不同用户和应用可以在同一平台上独立运行，确保数据的安全性和隔离性，同时方便资源的统一管理。

4、多索引支持：悦数 Graph 支持 BM25 和向量索引，满足不同类型知识的存储和检索需求。多种索引方式的整合，使得系统在知识检索时更加灵活和高效。

5、ISO-GQL 扩展的算法支持：悦数 Graph 基于 ISO-GQL 扩展了多种图算法支持，使得各种 Graph RAG 的索引和召回非常高效、灵活。在索引和检索阶段，可以灵活地应用各种图算法，如节点重要性评估、聚类分析等，提升知识检索的准确性和深度。

6、灵活的扩展性：悦数 Graph 的水平扩展能力使得系统可以根据业务需求灵活调整资源配置，保持高可用性和稳定性。

4.5.7 结论

本方案通过引入悦数 Graph，充分利用其分布式、云原生、多模和高性能的优势，构建了一个功能强大且可扩展的 RAG 系统。多类型的索引方式和多路并行召回策略，满足了不同知识文档的检索需求，提升了检索的准确性和效率。针对不同用户意图的召回策略，以及对现有图状数据的有效利用，使系统能够灵活地适应全局性和局部性的问题。悦数 Graph RAG 的优势体现在对超大规模数据的处理、高并发性能、多租户支持，以及对现有图状数据的最佳利用，为各类应用场景下的知识管理和利用奠定了坚实的基础。

4.6 面向专业领域的知识增强生成 (KAG) 解决方案

4.6.1 大模型垂直领域应用的关键问题

经过近两年的研究与实践，业界已普遍认识到大语言模型的优势与局限性，以及其在特定行业应用中的挑战。虽然大语言模型展现了强大的理解与生成能力，但在专业领域中仍存在缺乏领域知识、难以进行复杂决策及可靠性不足等问题。

4.6.1.1 LLM 不具备严谨的思考能力

首先，大语言模型无法提供严谨的推理能力。例如，对于“《1989 一念间》和《极品绝配》共同的主演是谁？”这个问题，国内几款大型语言模型结果显示回复的准确性和一致性都较低。即便某些模型能给出答案，也存在逻辑错误或问题拆解不当的情况。随着条件变的复杂，如变换条件为“男主演”“女主演”或添加时间约束，准确率和稳定性会不断下降。

为解决这些问题，行业内进行了诸多探索。比如，通过构建 Chain-of-Thought （COT） 模型，定义 Multiple/Tree/Graph 思维链模版，引导 LLM 合理拆解问题。今年以来，越来越多的研究聚焦于将 RAG 技术融入到大语言模型中，以弥补其在事实信息上的不足。进一步的发展则涉及 GraphRAG，即采用图结构来优化检索机制。

目前，引入外部知识库的方法被广泛应用，但即使是在引入了如 RAG 这样的技术，将特定领域的知识库或事实文档提供给大型语言模型进行重新生成时，仍不能完全保证生成答案的准确性。

4.6.1.2 事实、逻辑、精准性错误

下图左侧展示的是用大模型，对政府报告某个指标的解读示例，尽管业务人员已经提前做了标注，大模型仍然会加入自己的理解，导致信息失真或缺乏依据的错误。

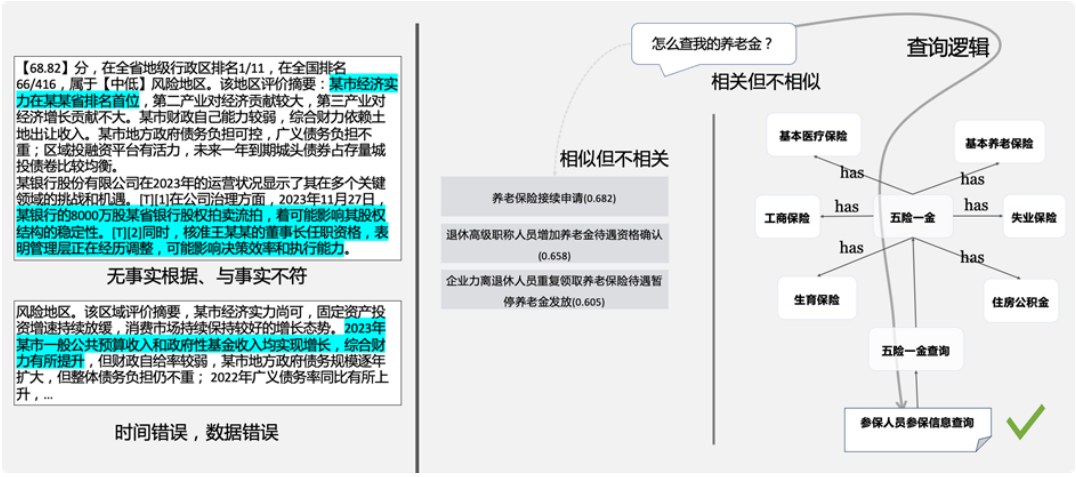

图 4.3 大模型存在事实、逻辑、精准性错误

即使提供了外部知识库，召回过程中的不准确问题依然存在。右侧示例说明了基于向量计算的 RAG 方法存在的缺陷。例如，在查询如何查找养老金时，直接利用向量计算召回的文档，与业务专家定义的知识并不相关。

在垂直领域内，许多知识虽然在表面上看似不相似，但实际上紧密相关。例如，“养老金”属于“五险一金”的范畴，与国家政策密切相关，大模型不能对此类信息进行随意生成。因此，需要预定义的领域知识结构来约束模型的行为，并提供有效的知识输入。

4.6.1.3 通用 RAG 也难以解决 LLM 幻觉问题

通常人们认为，引入 RAG 和外部知识库后，就能有效避免大模型的幻觉问题。其实不然，这种方式产生的幻觉问题甚至更为隐蔽。

图 4.4 通用 RAG 也难以解决 LLM 幻觉问题

近期蚂蚁集团发布了一项关于 RAG 引发幻觉现象的测评报告，根据评估结果显示，即便加入了 RAG 技术，大型语言模型仍然存在 30%-40% 的幻觉率，这是一个相当高的比例。因此，在垂直领域应用大型语言模型时，除了文本的检索增强，还需要更加知识化的表达，降低幻觉。

4.6.1.4 专业知识服务的挑战和要求

在真实的业务决策场景中，无论是生成研究报告还是处理车险理赔，解决复杂问题时都需要经过严格的步骤，包括问题规划、数据收集、执行决策以及生成和反馈等流程。在将大语言模型应用到专业领域时，也必须有一个严格且可控的决策过程。

所以，在基于大模型提供专业知识服务时，为了更好地服务于社会和特定领域，必须满足以下几个条件：

- 首先，必须确保知识的准确性，包括知识边界的完整性、知识结构和语义的清晰性；
- 其次，需要具备逻辑严谨性、时间敏感性和数字敏感性；
- 最后，还需要完备的上下文信息，以方便在知识决策时获取完备的支持信息。

以上也是当前多数大模型所欠缺的能力。

4.6.2 KAG：面向专业领域的知识增强生成技术框架

针对以上大模型垂直领域应用的关键问题，蚂蚁集团经过了大量探索，构建了面向专业领域的知识增强生成技术框架 KAG（Knowledge-Enhanced Generation），并于 2024 年 9 月在外滩大会进行了发布。

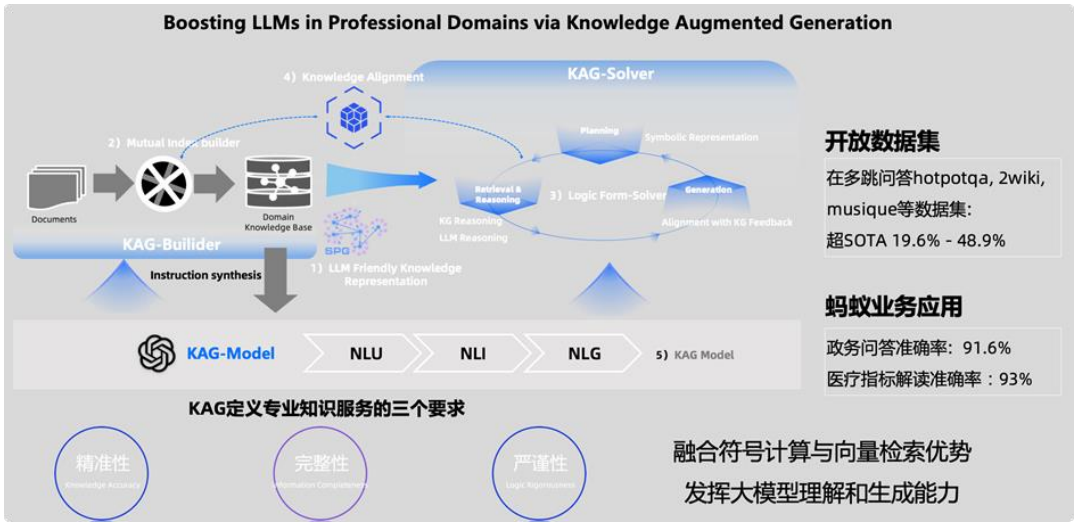

图 4.5 知识增强生成技术框架

上图展示了知识增强生成框架 KAG 的整体原理，该框架是在 OpenSPG 开源项目（蚂蚁集团 23 年开源的语义增强可编程知识图谱项目）基础上的升级。KAG 针对当前大语言模型与知识图谱结合的四个方面进行了增强：

首先，在知识表示上进行了增强。原有知识图谱受到强 Schema 约束，导致应用门槛较高且数据较为稀疏，使得在回答垂直领域问题时经常无解。为此，KAG 对知识表示进行了面向大语言模型的优化升级，使知识图谱能够更好地支持大型语言模型的应用。

其次，图作为一个优秀的集成工具，可以更好地连接各类知识，无论是严谨的学术知识还是文本中的信息。因此，KAG 创建了互索引结构，将原来的 term-based 倒排索引，升级成 graph-based 倒排索引。这样不仅能够有效地索引文档，还能保持文档间的语义关联性和实体间的连贯性。

第三，在推理过程中，KAG 采用了符号化拆解方式，以确保逻辑严谨性。语言模型生成的语言很难保证逻辑一致性，因此 KAG 引入了 LogicForm 驱动的 Solver 和 Reasoning，来进行基于符号的拆解。

第四，为了弥合知识图谱构建成本与实际应用效率之间的差距，KAG 借鉴了开放信息抽取（open information extraction） 的方法来构建知识图谱，这种方法大大降低了构建成本，但也引入了更多噪声。因此，KAG 同时引入了知识对齐 （knowledge alignment） 机制，利用概念知识完成开放信息与领域知识之间的对齐，旨在平衡开放信息抽取与语义对齐的需求。

4.6.2.1 LLMs 友好的知识表示

首先，KAG 对语义表示进行了升级。这是继 23 年 OpenSPG 项目开源后的进一步发展。OpenSPG 项目的初衷之一，就是将知识图谱从二元静态结构升级为多元动态结构。24 年，基于在深度上下文感知方面取得的进展，KAG 增强了对文本上下文的理解，可以提供更丰富的上下文信息，更好地服务于语言模型。

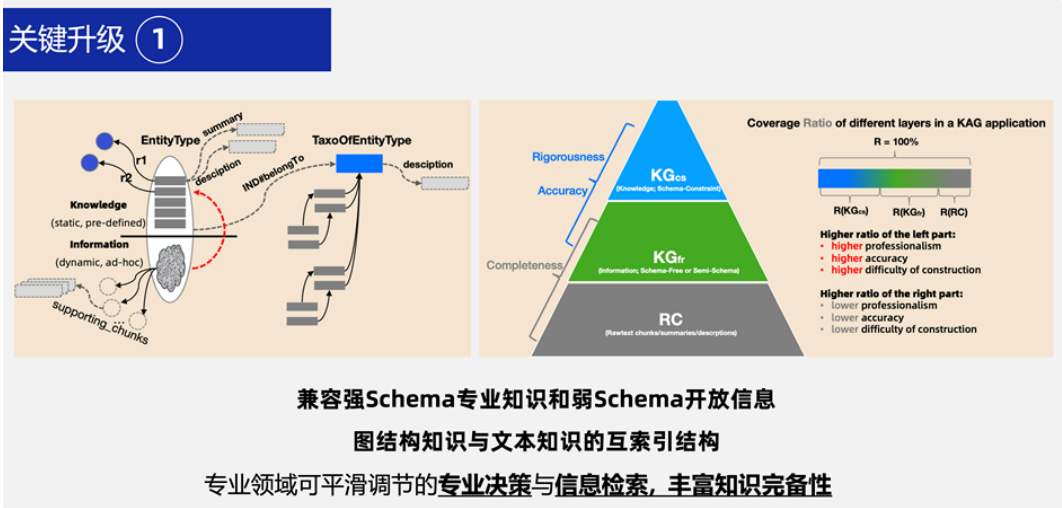

图 4.6 LLMs 友好的知识表示

其次，KAG 对图结构进行了升级。当前的图结构主要分为两大流派：LPG (Labeled Property Graph) 和 RDF (Resource Description Framework)。LPG 能更好地兼容数据库，因为它具有 Schema 模式，而 RDF 则相对开放。为了更好地支持大型语言模型，并实现从数据 (data) 到信息 (information) 再到知识 (knowledge) 的转变，KAG 参考了 DIKW 层次范式来实现统一的融合，使得在同一实体空间中，既能进行 Schema 约束的建模也能进行无模式 (schema-free) 的建模。

4.6.2.2 互索引：结构化知识与文本数据互索引结构

关键升级二，是从原有的 term-based 倒排索引升级到了 graph-based 倒排索引。通过基于实体和关系对文档进行倒排索引，这样既可以在同一空间内完成图计算中的图遍历，也可以关联到相应的文档片段 (chunk)，进行相关性的召回。

目前大火的 GraphRAG 范式的两种主要做法：一种是微软的 GraphRAG，实际上微软的 GraphRAG 并不算是严格意义上的 GraphRAG，它只是用知识图谱的方式组织了文档结构，做了分层摘要，并且最终用摘要来回答用户问题。这种方式反而会引入更多幻觉，这种做法在评估生成答案时，只考虑了流畅性、问题支持度和全面性，而没有从事实性角度进行评价。我们的测评显示，微软 GraphRAG 在事实回答准确率方面表现并不佳。

另一种以 HippoRAG 为代表，它采用图的方式构建倒排索引，并用图的方式召回文档来回答问题。在多跳信息问答上，相比传统的 Naive RAG 方法，HippoRAG 表现出了显著提升。

当获取到原始文档后，首先进行开放信息抽取。关于结构化构建的部分就不展开讲了，传统知识图谱中及开源的 OpenSPG 中都已经有了较为成熟的解决方案。KAG 会逐步抽取文档中的关键元素及描述性信息，并对文本块 (chunk) 进行有效的语义切分，最终形成的图结构将包含三部分：具体业务实体、通用概念知识以及文本块。这样一来，既可以在同一空间内完成图计算中的图遍

历，也可以关联到相应的文档片段 (chunk)，进行相关性的召回。如下图所示，通过图结构可以更有效地组织文档间的关联。

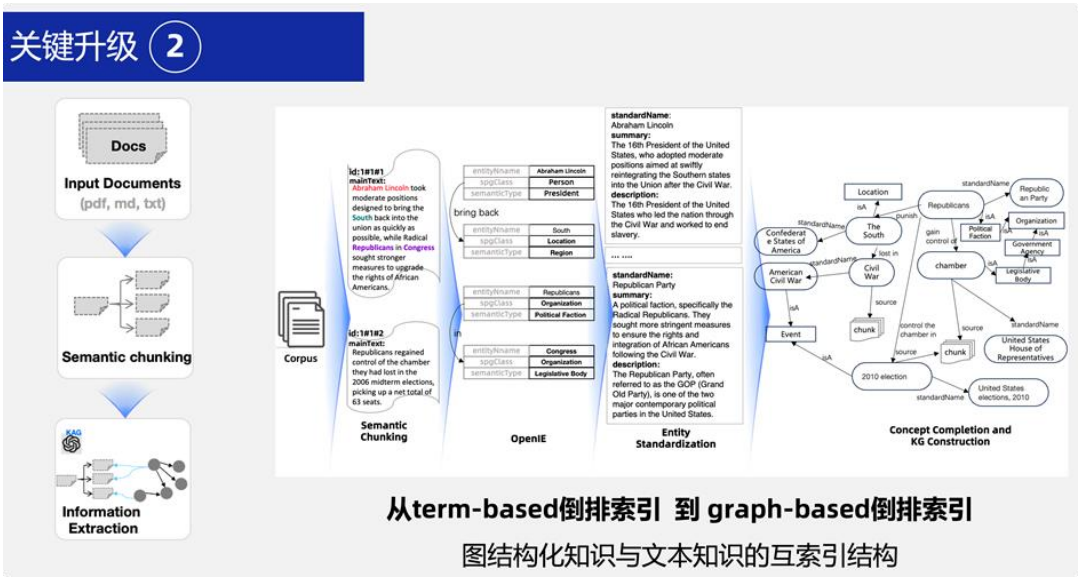

图 4.7 结构化知识与文本数据互索引结构

以上开放知识抽取可基于 OneKE 进行，OneKE 是蚂蚁集团与浙江大学合作于 24 年 5 月发布的大模型知识抽取框架，主要支持结构化信息抽取，使得在较小参数量的大模型上也能取得与更大参数量大模型相媲美的效果。它在实体识别、关系抽取和事件抽取等方面表现出色。最近，OneKE 又做了进一步升级，使其能够同时支持开放信息抽取。

4.6.2.3 混合推理：符号决策、向量检索与大模型混合推理

第三个升级是构建一个混合推理引擎。就像人类在回答问题前，要经过思考和规划一样，KAG 开发了一套技术范式，基于知识图谱来支撑严谨决策的问题。采用混合互索引的方式，既支持时间、数值、逻辑敏感的复杂决策执行，又能通过信息检索补充知识图谱的稀疏性和知识不足之处。

我们希望能够在垂直领域实现更准确的事实性回答，同时尽量不破坏知识的分层结构。这意味着在同一领域内，既有专业且严谨的 Schema 约束知识，也有通过文档提取出的图结构信息或知识，以及原始文档。如果能够实现这些不同层级知识的融合，就可以构建一个从严格到相对宽松的决策范式。

近期 OpenAI 发布的 o1 模型也是在长链条逻辑推理上有了重大进步，但出于竞争优势的考虑，不向用户展示原始思维链。

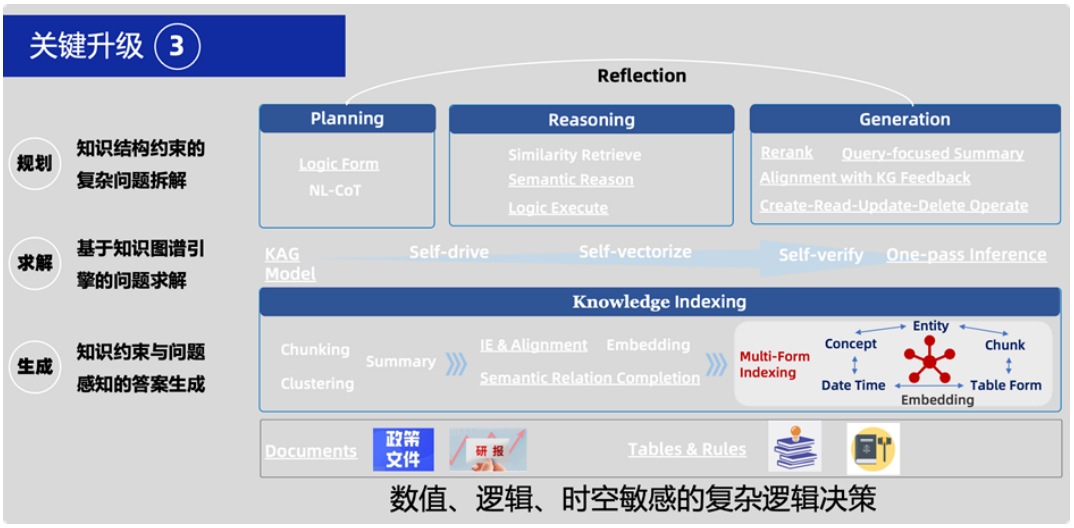

图 4.8 符号决策、向量检索与大模型混合推理

在 KAG 框架中，采用符号驱动的问题求解方法，能够生成逻辑上可执行的 Logic form 表达式，即 Logic Query 作为中间态的逻辑执行计划。获得 Logic Query 后，由于所有数据均基于图结构构建，就可以在图空间中进行操作。图空间内部存在分层结构，首先是逻辑严谨的知识，其次是开放的信息知识。这使得 KAG 可以分层决策，首先在逻辑严谨的知识层进行决策，如果没有找到答案，则在开放信息层继续决策，如果仍未找到答案，则在 chunk 空间进行关联检索，从而显著提高召回率和回答的准确性。

最后的生成阶段，目前沿用了业界一些主流方法，例如 query-focused summary，这种方法能根据 Query 结构来提取答案。传统知识图谱或索引的一个主要问题是索引构建与用户查询相分离，容易导致知识粒度不匹配，而通过 query-focused 总结方式可以更好地弥补这一差距。

下图展示了 KAG 的整体混合推理架构图及具体示例。例如，当询问“美国内战后，主张对南方各州实行严厉惩罚的政党在 2010 年控制了哪个机构？”时，系统会将其拆解成逻辑符号表达形式。这种表达方式可以直接转化为 KGDSL，但考虑到自然语言生成的函数表达的准确率和简洁性，KAG 选择采用自然语言生成的函数表达来表示逻辑执行计划。

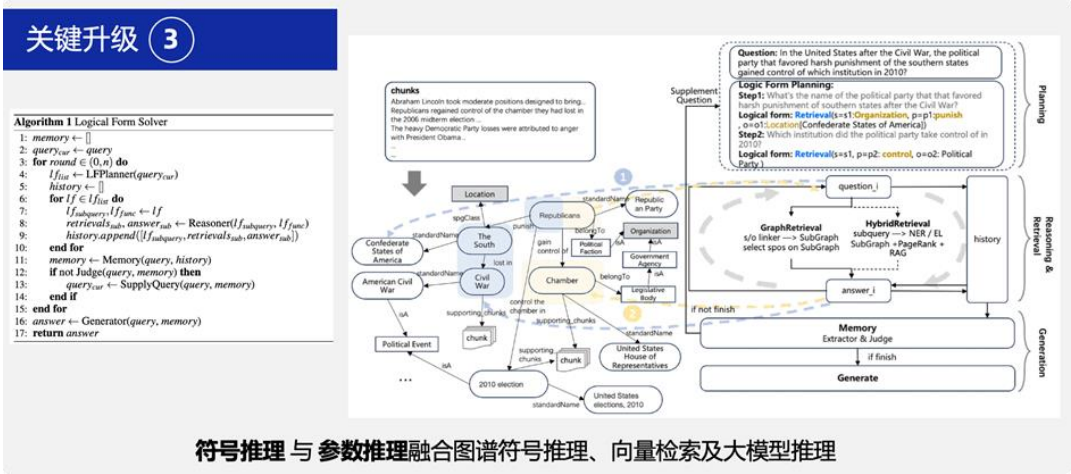

图 4.9 符号决策与大模型混合推理

目前，KAG 采取了三阶段推理，在生成之前，首先在图谱存储空间中进行 exact match，接着进行 SPO 的 Retrieval，然后是 chunk 的 Retrieval，实现分层检索与推理。

在生成阶段，通过引入图谱知识来缓解或抑制大语言模型生成时产生的幻觉。在前面文本中抽取知识图谱时，文本与知识图谱之间形成了良好的结构化数据与文本映射关系。首先，从文本到结构化，可以提取出关键要素信息；其次，结构化的图谱使大语言模型熟悉基于此类图结构生成文本的任务形式。因此，蚂蚁设计了文本到 SPO 及 SPO 到文本之间的双向映射任务，前者用于知识抽取，后者用于生成过程。通过这种方式合成语料，无论是用于 SFT 阶段还是强化对齐阶段，都能较好减少大型语言模型的幻觉。

通过原始文本可以抽取多个三元组，通过微调和强化对齐，将这些信息注入语言模型中，在生成时更好地遵循结构范式。蚂蚁将这一能力应用到内部业务中，例如区域风险报告生成场景。相较于原有的归档模型生成，幻觉率有了明显下降。

4.6.2.4 语义对齐：平衡信息检索与专业决策

第四个关键升级在于平衡专业决策与信息检索。信息检索本质上是对搜索引擎的升级，允许一定程度的错误率，但专业决策，错误的容忍度是很低。在统一的知识服务框架下，同时进行信息检索和专业决策是一项挑战。因此，KAG 对这一能力进行了升级，在顶层通过开放信息抽取获得结构化要素，在底层通过 Schema 约束构建更为严谨的知识。

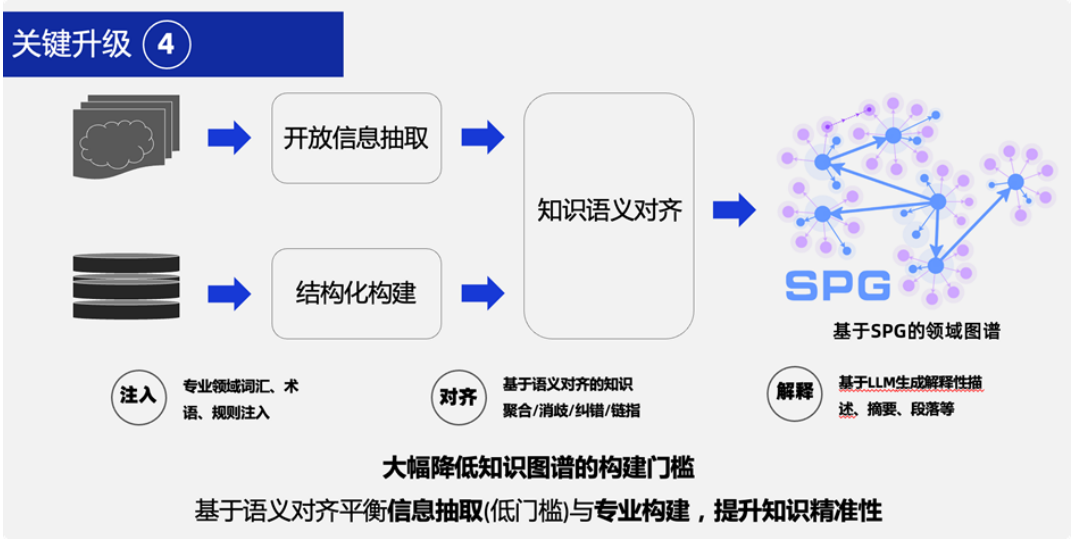

图 4.10 平衡专业决策与信息检索

最终，通过基于概念的语义对齐，构建一个基于 SPG 的领域知识图谱，能更好地兼容信息检索所需的开放抽取能力，和专业决策所需的 Schema 约束构建能力。

下图是一个简单的示例，展示了 KAG 如何基于开放信息抽取构建一个语义对齐后的图谱。从左侧的原始文本开始，对其进行语义切分，再进一步信息抽取，即可建立实体之间的关联，此时图谱仍包含大量噪声。当前业界主流的 GraphRAG 解决方案仅达到 information extraction 阶段，即生成三元组图后直接写入图数据库。然而，语义对齐才是知识图谱构建最困难的部分。为此，KAG 进行了大量探索，比如在提取的信息中运用图谱的传统方法，如实体链接、实体融合、概念与事实分层等，最终整个图结构的密度和语义完备性得到了显著改善。

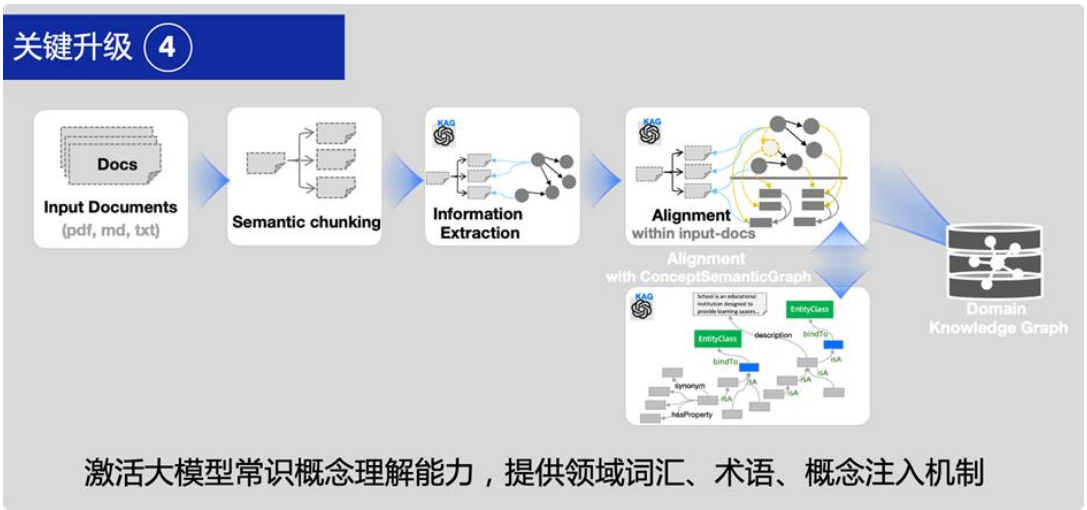

图 4.11 开放信息抽取语义自动对齐

除了开放域外，KAG 在垂直领域也进行了应用。垂直领域包含大量术语库、词汇库和概念库，例如医疗术语、法律术语和经济术语，这些术语对于大型语言模型来说可能难以完全掌握。因此，

KGA 通过在开放抽取过程中尝试实现领域词汇和概念的注入，以提高抽取效率和与领域知识对齐的准确性。

Framework	Model	HotpotQA		2WikiMultiHopQA		MuSiQue	
		EM	F1	EM	F1	EM	F1
NativeRAG [26, 25]	ChatGPT-3.5	43.4	57.7	33.4	43.3	15.5	26.4
HippoRAG [6, 25]	ChatGPT-3.5	41.8	55.0	46.6	59.2	19.2	29.8
IRCoT+NativeRAG	ChatGPT-3.5	45.5	58.4	35.4	45.1	19.1	30.5
IRCoT+HippoRAG	ChatGPT-3.5	45.7	59.2	47.7	62.7	21.9	33.3
IRCoT+HippoRAG	DeepSeek-V2	51.0	63.7	48.0	57.1	26.2	36.5
KAG *w/* LFS_{ref_3}	DeepSeek-V2	59.8	74.0	66.3	76.1	35.4	48.2
KAG *w/* $LFSH_{ref_3}$	DeepSeek-V2	**62.5**	**76.2**	**67.8**	**76.2**	**36.7**	**48.7**

Table 8: The end-to-end generation performance of different RAG models on three multi-hop Q&A datasets. The values in **bold** and underline are the best and second best indicators respectively.

Graph Index	Reasoning	HotpotQA		2Wiki		MuSiQue	
		EM	F1	EM	F1	EM	F1
M_Indexing	CR_{ref_3}	52.4	65.4	48.2	56.0	24.6	36.6
K_Alignment	CR_{ref_3}	54.7	69.5	62.7	72.5	29.6	41.1
	LFS_{ref_1}	59.1	73.4	65.2	74.4	31.3	43.4
	LFS_{ref_3}	59.8	74.0	66.3	76.1	35.4	48.2
	$LFSH_{ref_1}$	61.5	76.0	66.0	75.0	33.5	44.3
	$LFSH_{ref_3}$	**62.5**	**76.2**	**67.8**	**76.2**	**36.7**	**48.7**

Table 10: The end-to-end generation performance of different model methods on three multi-hop Q&A datasets. The backbone model is DeepSeek-V2 API. As is described in Algorithm 17, ref_3 represents a maximum of 3 rounds of reflection, and ref_1 represents a maximum of 1 round, which means that no reflection is introduced.

图 4.12 KAG 在通用数据集上的效果

经过优化，不仅验证了 KAG 在垂直领域的适应性，在通用数据集多跳问答中与现有 RAG 方法进行比较，发现它明显优于 SOTA 方法，在 2wiki 上 F1 相对提升 33.5%，在 hotpotQA 上相对提高 19.6% 。

4.6.3 KAG 在垂直领域的应用效果

今年以来，KAG 在蚂蚁 AI 生活管家“支小宝”、AI 健康管家等多个业务中进行了应用。在政务问答场景中，相较于传统的 Naive RAG 方法，准确率从 66% 提升到了 91%。在医疗问答方面，目前的准确率超过 80%，在更垂直的指标解读任务上，已达到 90% 以上的准确率。这些场景证明了这套方法不仅适用于通用领域的信息检索，也适用于垂直领域的专业决策。

图 4.13 KAG 在垂直领域中的应用

4.7 中英双语大模型知识抽取框架 OneKE

4.7.1 概述

大语言模型目前已显著提升了人工智能系统处理世界知识的能力，然而，以大语言模型为代表的生成式人工智能依然存在推理能力不足、事实知识匮乏、生成结果不稳定等问题，这些都极大的阻碍了大语言模型的产业化落地。

基于非结构化文档的知识构建一直是知识图谱大规模落地的关键难题之一，因为真实世界的信息高度碎片化、非结构化，大语言模型在处理信息抽取任务时仍因抽取内容与自然语言表述之间的巨大差异导致效果不佳，自然语言文本信息表达中因隐式、长距离上下文关联存在较多的歧义、多义、隐喻等，给知识抽取任务带来较大的挑战。

针对上述问题，蚂蚁集团与浙江大学依托多年积累的知识图谱与自然语言处理技术，联合构建和升级蚂蚁百灵大模型在知识抽取领域的能力，并发布中英双语大模型知识抽取框架 OneKE，同时开源基于 Chinese-Alpaca-2-13B 全参数微调的版本。测评指标显示，OneKE 在多个全监督及零样本实体/关系/事件抽取任务上取得了相对较好的效果。

4.7.2 OneKE 简介

OneKE 主要聚焦基于 Schema 的可泛化信息抽取，采用了基于 Schema 的轮询指令构造技术，专门针对提升大模型在结构化信息抽取的泛化能力进行了优化，旨在通过提供中英双语、可泛化的大模型知识抽取，OneKE 在一定程度上具备统一、通用、可泛化的知识抽取能力。同时，配套开源 OpenSPG 及 DeepKE 开源框架的支持，支持开箱即用。帮助研究人员和开发者更好地处理信息抽取、数据结构化、知识图谱构建等问题。

OneKE 的典型特点：

1、多领域多任务泛化性。支持金融、常识、医疗等领域实体多属性、事件多论元的抽取，不限制属性数量；

2、中英文双语支持。支持中文和英文两种语言文本的知识抽取任务；

3、完善的工具链支持。OneKE 依托 OpenSPG 及 DeepKE 开源库提供了完善的 SFT 及抽取工具支持，开箱即用。

4.7.3 OneKE 训练方法

4.7.3.1 整体方案

在 OneKE 的构建过程中，采用了 3 类任务 15 个领域 33 个数据集，通过对数据进行归一化和清洗提升数据质量，并在质量微调阶段采用了“基于 Schema 的轮询指令构造”技术，有效提升了模型的泛化能力。OneKE 的整体构建框架如图 4.14 所示。

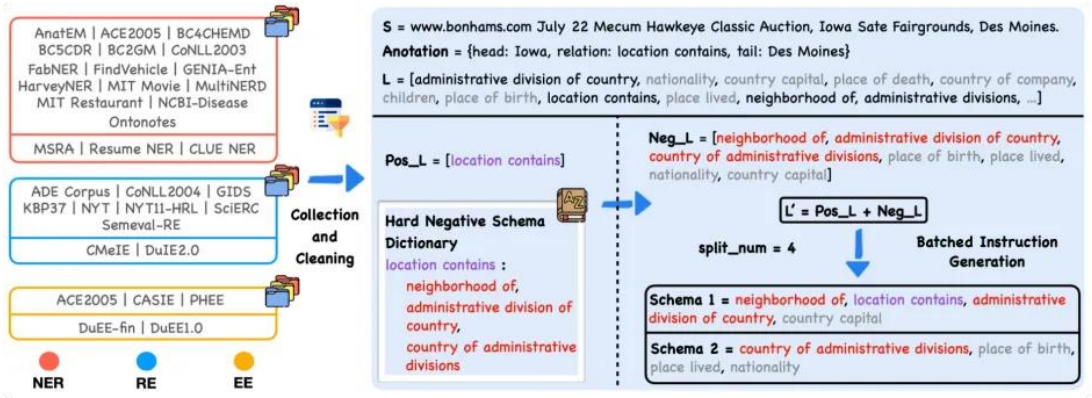

图 4.14 OneKE 整体构建框架

4.7.3.2 数据处理

数据的形式和质量是提升大模型能力的关键。针对不同领域、任务、语言的数据格式不统一问题，OneKE 在训练前进行了数据的归一化与清洗。首先计算每个数据集的训练集、验证集和测试集内的文本重叠情况。如果发现一个文本实例在同一个文件中多次出现，并且伴随着不一致的标签，则移除该实例。其次，设计启发式规则以过滤低质量和无意义的数据：1）非字母字符占文本总量超过 80%；2）文本长度不足五个字符且没有任何标签；3）高频出现的停用词，如‘the’、‘to’、‘of’等，超过 80%。

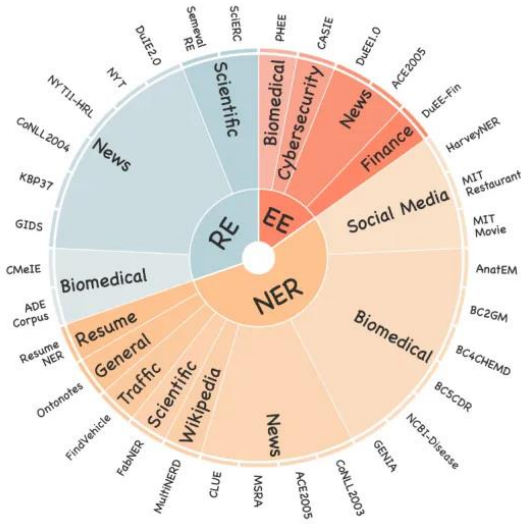

图 4.15 OneKE 中不同任务、领域及数据源分布情况

4.7.3.3 指令微调

根据前人工作以及错误案例分析发现，抽取模型误判的主要原因可能在于 Schema 语义相似导致的混淆。因此，OneKE 在指令微调训练过程中采用了“基于 Schema 的轮询指令构造”技术。

Algorithm 1 Schema-Based Instruction Generation

Require: Text S, Full label set L, Positive label set Pos_L, Number of schemas to split $split\_num$
Ensure: Negative label set Neg_L, Set of instructions $Instructions$
Step 1: Initialize Hard Negative Schema Dictionary $\mathcal{K}$
for all schema in L **do**
  $\mathcal{K}[schema] \leftarrow$ SEMANTIC-SIMILAR$(schema, L)$
end for
Step 2: Obtain Hard Negative Schemas
$Hard\_L \leftarrow \emptyset$
for all schema in Pos_L **do**
  $Hard\_L \leftarrow Hard\_L \cup \mathcal{K}[schema]$
end for
$Other\_L \leftarrow L - Pos\_L - Hard\_L$
$Other\_L \leftarrow$ RANDOM-SELECT$(Other_L, split_num)$
$Neg\_L \leftarrow Hard\_L \cup Other\_L$
$L' = Neg\_L \cup Pos\_L$
Shuffle L' to obtain a randomized sequence
Step 3: Batched Instruction Generation
$Instructions \leftarrow []$
$num\_batches \leftarrow \lceil \frac{|L'|}{split\_num} \rceil$
for $i \leftarrow 1$ **to** $num\_batches$ **do**
  $Batch \leftarrow$ SEQUENTIAL-SELECT$(L', split_num, i)$
  $Instructions \leftarrow Instructions \cup$ GENERATE-INSTRUCTION$(Batch)$
end for

图 4.16 OneKE 指令微调算法

具体如图 4.14 右侧所示，先构建一个困难负样本字典，其键值对应的是 Schema 及其语义上相近的 Schema 集。难负样本的构建旨在促进语义近似的 Schema 更频繁地出现在指令中，同时也能在不牺牲性能的情况下减少训练样本量。然后，采取一种批次化指令生成方法，动态限制每条指令中询问的模式数量为 N（其范围在 4 到 6 之间）。

即使在评估阶段询问的 Schema 数目与训练时不同，通过轮询机制可以将询问数量平均分散至 N 个，从而缓解泛化性能下降的问题。具体算法如下图所示，详细技术细节可参阅论文“IEPile: Unearthing Large-Scale Schema-Based Information Extraction Corpus”。

通过“基于 Schema 的轮询指令构造”技术，并融合开源及蚂蚁业务相关 NER、RE、EE 等近 50 个数据集可得到约 0.4B tokens 的大规模高质量抽取指令微调数据，其中部分数据已通过 IEPile 开源。OneKE 模型是通过在 LLaMA 上进行全参数微调得到的，这一过程利用了以上大规模高质量的抽取指令数据。

4.7.4 OneKE 效果

如下图所示，OneKE 具备相对较好的中英双语可泛化的知识抽取能力，其中在中文 NER 命名实体识别类任务、RE 关系抽取类任务、EE 事件抽取类任务上取得了相对较好的效果。

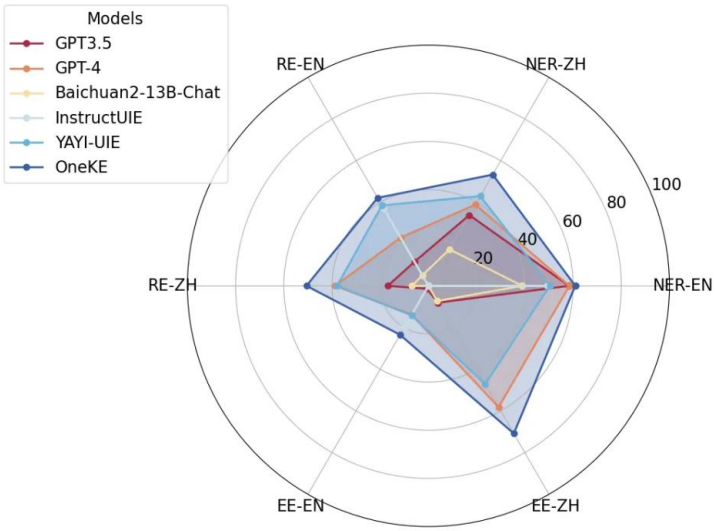

图 4.17 OneKE 在零样本泛化性上与其他大模型的对比结果

4.7.5 OneKE 使用方法与案例

OneKE 中指令的格式采用了类 JSON 字符串的结构，本质上是一种字典类型的字符串。它由以下三个字段构成：（1） 'instruction'，即任务描述，以自然语言指定模型扮演的角色以及需要完成的任务；（2） 'schema'，这是一份需提取的标签列表，明确指出了待抽取信息的关键字段，反应

用户的需求，这是动态可变的；（3） 'input'，指的是用于信息抽取的源文本。目前可通过 DeepKE-LLM 或 OpenSPG 来直接使用 OneKE，高级用户可自行转换和构造指令使用 OneKE。

1）基于 DeepKE-LLM 使用 OneKE

用户可以按照 DeepKE-LLM 项目指引完成环境配置、模型权重获取、数据转换后直接使用 OneKE，DeepKE-LLM 也支持对 OneKE 进行量化（如 4bit 量化）以实现在低功耗设备上运行 OneKE。

DeepKE-LLM 项目：

https://github.com/zjunlp/DeepKE/blob/main/example/llm/OneKE.md

2）基于 OpenSPG 使用 OneKE

用户可以按照 OpenSPG 项目指引完成环境配置、模型权重获取、Schema 定义、数据转换后直接使用 OneKE。

OpenSPG 项目：

https://openspg.yuque.com/ndx6g9/nmwkzz/dht0wtgycuw032gd

基于 OpenSPG kNext 编程框架，用户可以按照 SPG Schema 的定义，提交端到端的图谱构建任务，实现文本到知识的自动转换，同时实现属性标化、实体链指，更新并写入到图谱存储，同时还可以使用 SPG KGDSL 查询构建好的结果。用户也可以添加领域指令数据后提交本地的 SFT 任务。

4.7.6 局限与不足

OneKE 在全监督及多领域泛化性上有比较出色的表现，统一的指令结构也能让业务通过增加更多领域标注数据以获取更好的模型能力。通过 OneKE 框架，证明了基于大模型统一知识构建框架的可行性。

然而，在实际的工业应用中，业务对知识要素的覆盖率、准确率要求非常高，统一 Schema 指令结构难以覆盖所有的知识表示形式，因此 OneKE 依然存在抽不全、抽不准以及难以处理较长文本的问题。

由于模型的规模有限，模型输出极大地依赖于输入的提示词（Prompt）。因此，不同的尝试可能会产生不一致的结果，且可能存在幻觉输出。蚂蚁与浙江大学也在并行探索开放知识抽取，联动图谱自动构建系统，持续优化和提升 OneKE 新领域及新类型上的适应性。

第 5 章 应用案例

5.1 产业落地

5.1.1 能源电力

在某电力调度控制中心，悦数图数据库的应用极大地提升了调度控制的效率和准确性。通过构造融合调度、配网、暂态和市场等多类业务的统一时空立体图模型，该中心实现了设备间关联关系从全量秒级计算到增量毫秒级读取的根本性改变。这一模型不仅揭示了新型电力系统的特征，还为电力调度提供了更全面和准确的数据支持，使得调度决策更加科学和高效。

在图存算方面，悦数图数据库构建了超高速图存算引擎，成功解决了电力系统分析计算平台的可扩展性不足、计算结构与逻辑复杂、计算效率低下等难题。该引擎的引入显著提升了系统的运行稳定性和效率，使得大规模电力调度和控制任务能够在短时间内高效完成，确保了电力系统的平稳运行。

通过图应用，悦数图数据库进一步提高了电力调度控制中心的风险控制水平。结合静态图中心度、动态负荷波动性和短路电流要求，该中心能够有效发现新能源控制断面，提高风险预控水平。层层筛选减小计算规模，使得 SCUC 计算效率平均提升近 3 倍，从而能够应对大规模市场出清需求。通过这些应用，电力调度控制中心不仅提高了风险预判能力，降低了停电风险，还能自动识别风险预警断面超过 5000 次。

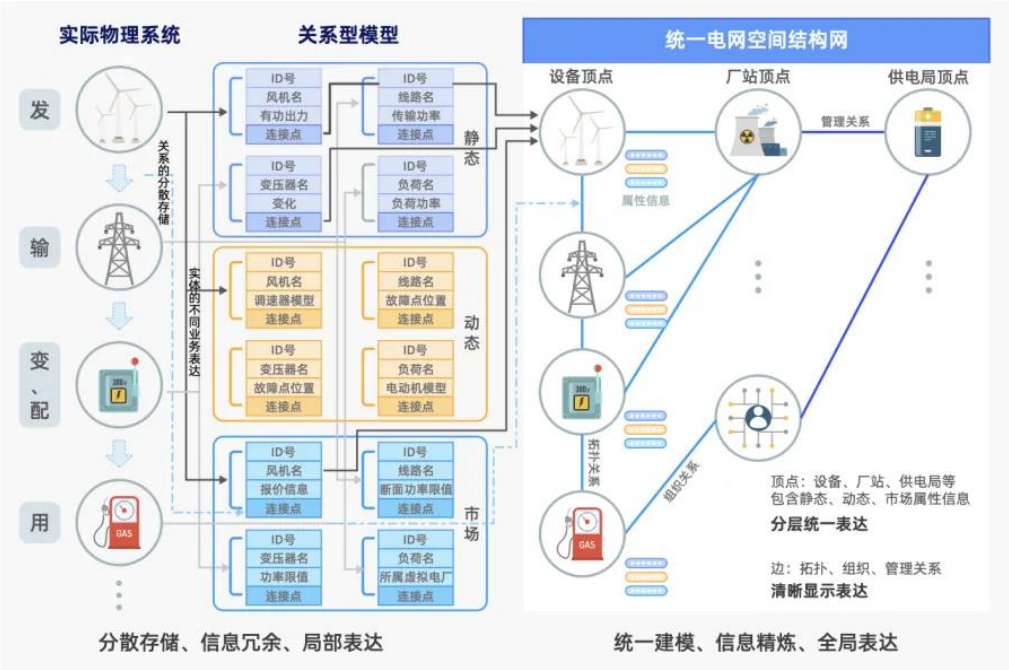

图 5.1 电网图模型

在经济效益方面，悦数图数据库实现了自动化系统的高效分析计算，节约了超过 600 万元的服务器投入成本。同时，实现了轻量化运维，节约了系统运维人力资源成本超过 700 万元。更为重要的是，激发了市场主体的活力，间接产生了发电侧现货市场结算电费收益超过百亿元。

在社会效益方面，悦数图数据库的应用显著提高了电力行业图模构建效率与交互规范性，增强了系统风险预判能力，降低了停电风险。现货市场优化计算效率的提升，支持了现货市场出清超过 5 万次，为电力市场的稳定运行提供了强有力的技术支持。

通过以上多方面的实践应用，某电力调度控制中心充分展示了悦数图数据库在新能源行业中的巨大潜力和广泛应用前景。该解决方案不仅提升了电力系统的运行效率和安全性，还为整个行业带来了显著的社会和经济效益。

5.1.2 金融

5.1.2.1 信用卡反欺诈

在当前金融业务迅速扩展的背景下，信用卡中心正面临越来越复杂的风控挑战，特别是在应对新型信用卡欺诈团伙时，传统风控手段已显得力不从心。欺诈行为通常具备规模化、隐蔽性、团伙作案、少样本和动态演进等特征，使得传统的专家规则和机器学习模型难以应对这些复杂的风险。

为了解决这一问题，某银行信用卡中心使用创邻科技的 Galaxybase 图数据库提供的 HTAP 能力，构建了实时图数据读写与图计算技术相结合的解决方案。这个系统整合了来自银行内部和外部的多种数据，建立了一个覆盖数十亿条记录的信用卡申请网络图谱。这张图谱揭示了申请人之间的复杂关联，比如共享电话号码和设备号等信息，从而帮助识别看似独立的申请背后的潜在联系，特别是在团伙欺诈的场景下。

在这一方案中，图计算技术的核心在于应用了一系列基于关联关系进行模式检测的算法。首先，标签传播算法被用于在图谱中标记风险。系统通过对少量已知欺诈样本进行标记，能够将这些风险标签在整个网络中传播，从而有效识别潜在的欺诈节点。这种方法能够发现与已知欺诈样本相关的其他节点，提升了检测的全面性。PageRank 算法则用于评估节点在网络中的重要性，通过分析节点的连接情况，识别出对欺诈网络至关重要的节点，帮助集中资源处理高风险区域，从而提高了风控的效率。此外，系统基于图计算结果构建了欺诈检测模型，考虑了节点的多层次关系，深入理解了图谱中的复杂结构，从而提高了对潜在欺诈模式的识别能力。为了应对欺诈行为的快速变化，系统自动根据实时图计算结果调整欺诈检测模型，保持对新风险的高度敏感。

Galaxybase 图数据库的高效分布式并行处理能力和 HTAP 特性，使得系统能够在每个申请生成时实时进行关联分析，并在毫秒级别完成风险评估。这种实时处理能力帮助信用卡中心迅速反应，有效降低客户群体性风险。

通过上述实践，Galaxybase 的图计算技术在风险检测方面展现了显著的优势，相比传统方法，它不仅提升了检测的精准度和效率，还大幅减少了人工审核的工作量，实现了风控流程的自动化。系统成功识别了多个重大欺诈案件，揭示了涉及金额超过亿元的欺诈团伙，从而显著增强了信用卡中心的风控能力。该系统的应用有效降低了资金损失，带来了显著的社会和经济效益。

5.1.2.2 存款流失预警

对于商业银行，客户的存款流失问题一直备受关注。尤其是高价值客户的存款如果不断减少，银行的资金流动性将造成较大压力。存款流失预警系统有助于精准识别高风险客户并采取有效措施。经过分析，银行发现高价值客户的存款变动情况与其流失风险高度相关。如果客户的活期存款急剧减少，且不再进行频繁交易，银行便迫切需要识别出来，为其提供个性化服务以降低流失风险。然而，传统的数据分析方法无法提供及时且准确的预警，造成了潜在流失客户的巨大损失。

基于 TuGraph 图系统建设的分析系统，可以通过账户的交易和联系，挖掘更多特征，用于学习预警模型。除个体维度的特征（如个人年龄、性别、账户规模、变动频率等）外，增加账户交易的特征，例如一段时间内的交易、交易渠道等，捕捉账户间联系，自动学习拓扑模式，建立更加精准的分类模型。

最终构建的流失客户预测模型，在对客户进行风险评估时，其流失概率排名前 2000 的客户，命中率近 80%；前 10000 名客户的流失金额占总流失金额的比例约为 72%。

在预测模型的基础上进行预警系统和机制的建设，银行可以降低高价值客户的流失率，恢复了客户的存款信心。流失金额显著减少，整体客户满意度提升，促进了客户与银行之间的坚实关系，最终推动了银行业务的持续增长。

5.1.2.3 交易风控

1、团伙挖掘与反欺诈场景

在洗钱操作中，多个账户通常会通过多个中间账户进行转账，以规避单一账户的可疑交易监测。传统的基于单一账户或简单规则的检测方法只能捕捉个别异常交易，团伙成员可能通过复杂的关系网络掩盖其非法行为，单点处置难以识别隐藏在多层网络中的整个团伙。而通过基于图计算思想的子图模式匹配算法（Subgraph Pattern Matching），可以将多个节点（账户、交易、联系信息）和边（交易流、联系链）构成的关系图作为一个整体进行分析，识别出与典型洗钱网络类似的子图，从而挖掘出整个洗钱的模式与结构，帮助银行快速定位团伙。相较于其他方案，基于图思想的子图模式匹配能够更好的识别团伙之间的复杂关联，优势主要体现在以下几个方面：

- **捕捉复杂、多层次的网络结构**：传统方案多依赖单点异常识别，无法发现多个账户间的潜在关系。子图模式匹配通过分析多节点间的复杂关系，可以捕捉到洗钱团伙分布式作案的模式，提升识别效率；

- **高效模式识别**：一旦银行识别出一个洗钱团伙的操作模式，子图匹配技术可以将其作为模板，应用于后续的交易网络分析中。这种模式化检测提高了系统的复用性和识别速度，可以有效地发现相似的洗钱行为，并提前预警；
- **抵抗动态变化的风险**：子图模式匹配不仅能识别静态模式，还能通过模糊匹配识别洗钱团伙的变种手法。当团伙改变交易路径或增加伪装账户时，传统规则可能难以应对，但子图匹配算法可以容忍部分变化，从而提升算法鲁棒性和识别稳定性。

在面向大规模实时数据场景时，对图数据的计算和查询时效性提出了挑战，信雅达构建了一套基于图原生数据库的图计算平台，部署高性能图数据库配合图索引，通过优化面向图的计算度加速子图匹配效率。

2、基于知识图谱的风险传导分析与用户风险评分应用

在甄别涉赌涉诈风险账户的场景中，利用知识图谱技术可以深入挖掘复杂交易网络中的风险传导路径。首先，通过司法冻结记录、惩戒与风险警告信息等手段，识别出初始的“黑标客户种子”作为风险挖掘的起点。这些账户作为高风险节点，通过基于图的迭代算法（如 GAS 算法）逐步扩展其与其他账户的交易关联关系。通过此扩展路径的分析，可以有效识别与这些黑标客户有直接或间接联系的潜在疑似风险账户，揭示隐藏在复杂交易网络中的关联风险。

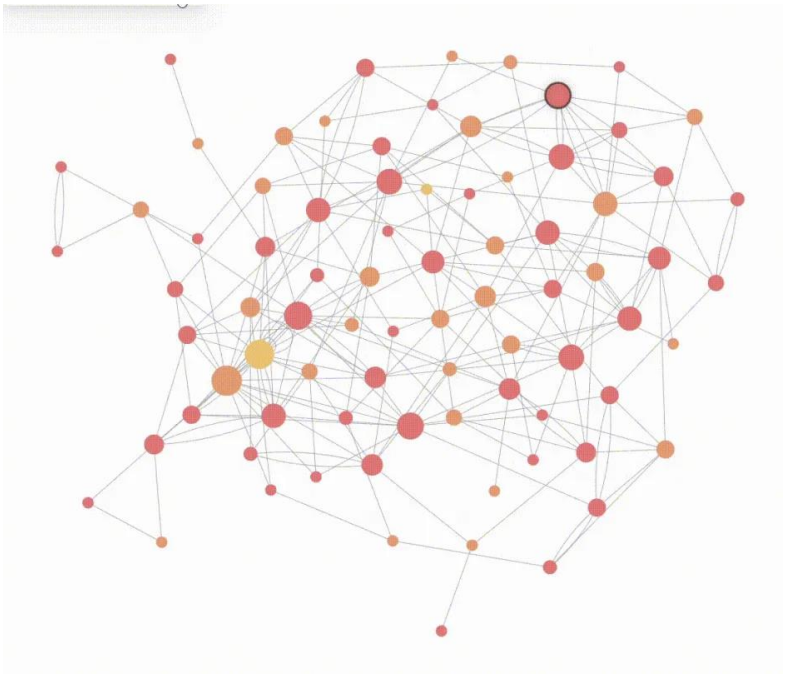

图 5.2 风险传导示意图

在识别出潜在风险账户后，通过构建风险评分模型来评估每个疑似账户的风险程度。以使用时序图神经网络（TGNN）模型为例，首先构建以账户为节点、资金流动为边的动态图网络，捕捉账户之间的交易行为及其时序特征。该模型同时将交易频率、交易时间等关键信息嵌入图结构中，

学习账户之间的资金流动模式和时间维度上的异常交易特征。通过这些分析，精确计算每个账户的风险分值，帮助金融机构优先处理高风险账户。此基于知识图谱的风险传导分析与评分体系能够有效应对复杂的涉赌涉诈交易网络，显著提升银行对潜在风险的识别和管控能力。

5.1.3 电商

5.1.3.1 租赁反欺诈

芝麻免押租赁是一种基于信用评价体系的租赁服务，主要依托于支付宝的芝麻信用分。用户在租赁物品时，如果芝麻信用分达到一定标准，可以选择免押金租赁。这种方式可以降低租赁门槛（用户不需要支付押金）、简化流程增加便捷性，并且鼓励用户维护良好的信用记录。然而，有大量的中介人员，通过教唆或帮助用户，利用各种不良手段，骗取租赁物品从而获得利润。这种类型的风险占到整体欺诈风险的30%以上。这类人员的典型特点是：以中介为核心，与中介相关的人员可能受到其教唆，形成一个“团伙”。

基于 TuGraph 图智能模型，构建基于半监督中介拓展的团伙识别框架，不再仅仅关注一层的风险，而是通过资金、媒介等关系，提取租赁用户的历史交互子图。通过这种深入的分析，可以推断出高可能性的中介，并通过这些中介定位到高风险的欺诈用户。然而，租赁场景中的中介标签严重不足，无法依赖传统的监督学习方法完成识别。因此系统采用半监督学习的方法，通过拓展疑似中介，从而减少对中介标签的依赖。

此外，基于图智能框架，设计了一个基于图神经网络的欺诈风险评估模块，引入用户与中介的交互时间、频率、金额等信息，以提升对高风险团伙欺诈的推断精度。

通过这一系列的创新，构建的系统能够有效识别和打击中介相关欺诈，不仅提高了芝麻免押租赁场景的风险识别能力，也大大降低了因中介欺诈行为带来的损失。模型上线以后，团伙风险发生率环比下降 17%。

5.1.3.2 跨境电商

在国际电商场景中，用户主要使用卡交易，其主要风险为盗卡风险。当黑产在电商平台使用盗来的卡交易时，如果电商不拦截该交易，那么卡的原持有者发现自己的卡被盗刷后，会向卡行发起拒付申请，电商平台需要承担赔付责任。由于国际电商准入门槛低，仅凭邮箱即可完成校验，因而吸引大量黑产。同时，管控方式相对单一、无线下打击，导致黑产猖獗持续作案。跨境电商场景的风控存在以下挑战：

- 案件报回慢：由于国际电商场景的盗卡风险报回链路慢长，平均一个月报回60%，三个月报回90%，导致该场景下风险感知慢，难以做到及时防控，当用户报回风险时，往往一批团伙作案已经完成了。

- 信息匮乏：在国际电商交易中，风控主体无法获取到个人、社交、关联和设备等相关信息，能够获取的主要是交易本身的信息。

针对国际案件报回慢的特点，如果等到案件报回后再去防控相应风险就已经错过了防控窗口期。因此，运营同学需要在事后对事件定性分析，以尽早捕捉风险，及时防控。在该场景下，利用图风控技术充分挖掘案件定性相关的交易信息，实现更快更准的风险定性。在国际电商场景中，盗卡风险分为无关联性的单点风险和呈现批量性的团伙风险。据统计，团伙一次批量性的盗卡行为 90%交易在首次交易后 3 天内完成。因此，风险定性窗口主要分为事中、事后 3 天内和事后 3 天后三个场景。

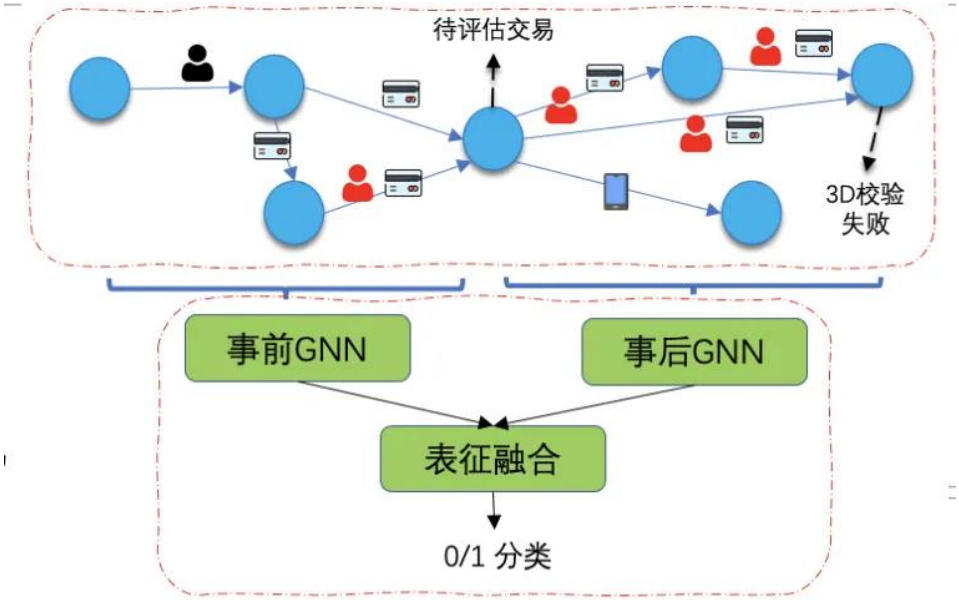

图 5.3 国际电商场景团伙风险定性流程

在事中定性场景中，风险定性模式主要为单点风险定性为主。利用高效的图数据库引擎，实现了海量数据下图上特征的毫秒级计算。将该类图统计特征加入到交易盗卡风险实时定性模型中，可以实时地有效防控在图上已经呈现关联性的交易风险。

在事后 3 天内定性场景中，风险定性分为单点风险定性和团伙风险定性两种模式。针对于单点风险定性，相比于事中风险定性，此时有了更多可利用的交易事后信息。基于图风控平台，使用时序图表征建模以充分挖掘交易事前事后信息，在交易发生后 T+1 时效感知有风险交易，及时更新防控策略，进行止损。

针对于团伙风险定性，利用图风控平台深度挖掘不同用户在交易强、弱介质上的关联性，并通过连通图、Louvian 等算法分析尽可能让黑白用户在图上呈现聚类性，实现 T+1 时效的团伙风险防控。同时，基于近线团伙防控平台，通过长短周期构图和社区搜索的方式大幅缩减团伙风险定性耗时，实现部分图上聚集性风险的秒级定性和防控。

在事后 3 天后定性场景中，同一批风险已基本定性。但是在该场景下，依然需要充分挖掘所有隐性案件，例如拒绝交易案件和未报回案件，为策略和模型的开发提供准确的评估标签，以防控之后再有类似手法的风险。在该场景下，基于图风控平台，使用标签传播算法充分挖掘与案件在设备、账号等交易介质上存在关联的隐性案件。

5.1.3.3 芝麻职业图谱

芝麻信用提供了职业认证服务。用户可以通过芝麻信用进行职业认证，这有助于提升芝麻信用分，同时也可能为用户提供更多的信用服务。此外，芝麻信用还会根据用户的职业信息，提供一些与职业相关的信用服务，如职业培训、职业指导等。

传统上，对于用户求职、参与培训、证书认证等环节，对于岗位、场景、技能、证书内容等大量信息的精确度不够。因此，基于 TuGraph 建设的职业图谱，统一各场景数据，还能完善用户画像，有助于带来场景增益、提升用户人岗匹配的效果。这里利用了用户填写的职业档案，并结合图学习技术，将用户行为子图和用户档案子图的信息表示为通用知识表征，提升职业推荐精度。

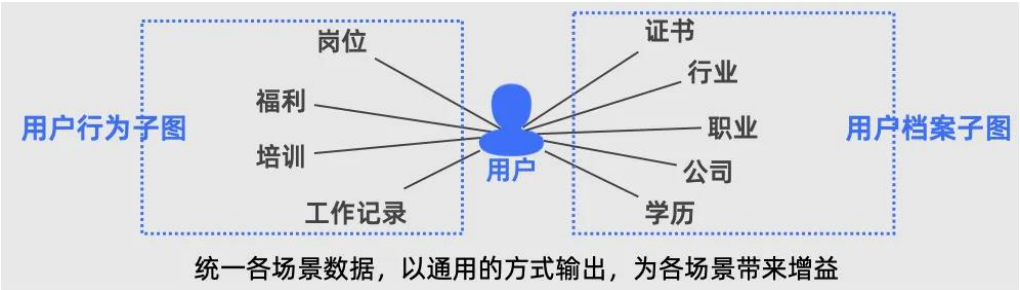

图 5.4 职业图谱统一场景数据

在图学习知识表征中，通过采用无监督多任务学习，获得独立于下游的通用图表征向量，轻量引入多个场景。主要步骤包括：

1、特征组合：将节点包含的点边类型编码特征、点边特征、文本编码特征等进行组合，丰富邻居节点特征表达；

2、子图采样和汇聚：从全图中采样得到子图，然后对邻居进行汇聚传播，得到中心节点的嵌入向量表征；

3、无监督多任务学习：构造链接预测任务，结合无监督学习任务，强化嵌入向量表达能力

从而在下游场景应用中，将图向量引入点击率模型，提升岗位推荐的有效性。有图嵌入特征的模型，点击率提升近 3%，带来业务价值。

5.1.4 游戏

5.1.4.1 背景

腾讯游戏作为全球领先的游戏厂商，拥有海量用户和众多游戏产品。为了更好地运营和推荐游戏，腾讯游戏构建了游戏知识图谱，并将其与 AI 技术相结合，形成了独特的 Graph+AI 体系。本文将介绍腾讯游戏在游戏知识图谱构建和应用方面的探索与实践。

5.1.4.2 游戏知识图谱-游谱

早期对游戏的刻画主要依赖游戏描述文本，通过分词技术提取关键字进行简单描述。然而，游戏作为一种融合了美术、音乐等多种艺术形式的“第九艺术”，其复杂性和多样性远非简单关键字所能涵盖。为了更全面地刻画游戏，我们构建了名为“游谱”的游戏垂直领域知识图谱，这是一个多模态的知识图谱，汇聚了全球约 200 万款游戏实体，涵盖主机、PC 和手游三大类游戏。游谱融合了文本、图像、音效等多种信息，并通过 NLP、CV、音频等技术进行处理和分析，从而更全面地刻画游戏的各个方面，例如：

- **文本信息：** 游戏名称、类型、描述、关键词等。
- **图像信息：** 游戏截图、图标、角色形象等。
- **音效信息：** 游戏音乐、音效等。
- **交互信息：** 游戏玩法、机制、操作方式等。

通过多模态知识图谱的构建，我们能够更深入地理解游戏，并为游戏推荐、评估、运营等场景提供更精准的数据支持。

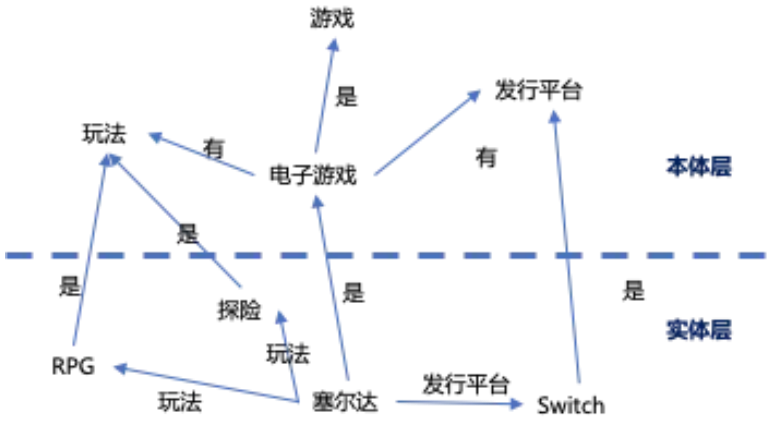

图 5.5 游戏知识图谱示意图

5.1.4.2.1 第一版

游谱的第一个版本，也就是游戏知识库，于 2015 年启动构建。当时，为了支持新游戏的拉新工作，需要提取游戏的特征信息。然而，游戏作为一种超媒体，包含了文本、图像、音效以及与玩家的交互等多种难以量化的特性，对其进行全面刻画面临着巨大挑战。

由于当时的技术条件和人力资源有限，我们选择从文本信息切入，构建了第一个版本的图谱。该版本主要包含公司、游戏名称、游戏类型、游戏关键词、游戏描述等信息，并利用分词、词性标注、LDA 等自然语言处理 （NLP） 技术对这些文本信息进行处理和分析。

通过 NLP 技术的深入挖掘，我们成功构建了约 80 维度的标签体系，覆盖了 10 万款游戏。这一标签体系有效地刻画了游戏的特征，并为游戏推荐、搜索、分类等场景提供了重要的数据支持。

游谱 1.0 版本的优势：

- **高效构建：** 基于文本信息，利用成熟的 NLP 技术，能够快速构建知识库。
- **覆盖面广：** 涵盖了 10 万款游戏，能够满足大多数游戏场景的需求。
- **易于扩展：** 可以根据需要添加新的标签维度，进一步提升刻画能力。

游谱 1.0 版本的局限性：

- **刻画能力有限：** 仅基于文本信息，难以全面刻画游戏的复杂性和多样性。
- **缺乏多模态信息：** 无法利用图像、音效等多模态信息进行更深入的刻画。

尽管游谱 1.0 版本存在一定的局限性，但它为后续版本的迭代和完善奠定了基础，并成功应用于多个游戏场景，取得了良好的效果。

5.1.4.2.2 第二版

随着游戏业务的不断发展，对游戏刻画的需求也日益增长。早期版本的游戏知识库，由于维度有限，难以满足日益复杂的需求，例如：

- **游戏市场排名预测：** 需要更全面地了解游戏特征，才能准确预测其在市场上的表现。
- **游戏用户规模预估：** 需要更深入地分析游戏特性，才能有效预测其潜在的用户规模。
- **游戏玩法和手感刻画：** 需要更细致地描述游戏的玩法和手感，才能更好地满足用户需求。

为了解决这些问题，我们参考了 DPE（Design Patterns and Elements）和 MDA（Mechanics-Dynamics-Aesthetics）等常见的游戏设计框架，设计了 40 维度的游戏画像 Demo 版本。该版本的游戏画像不仅包含了游戏类型、题材、画面风格等基本信息，还涵盖了游戏玩法、难度、目标、奖励机制、交互方式、故事背景等多个维度，能够更全面地刻画游戏的特性。

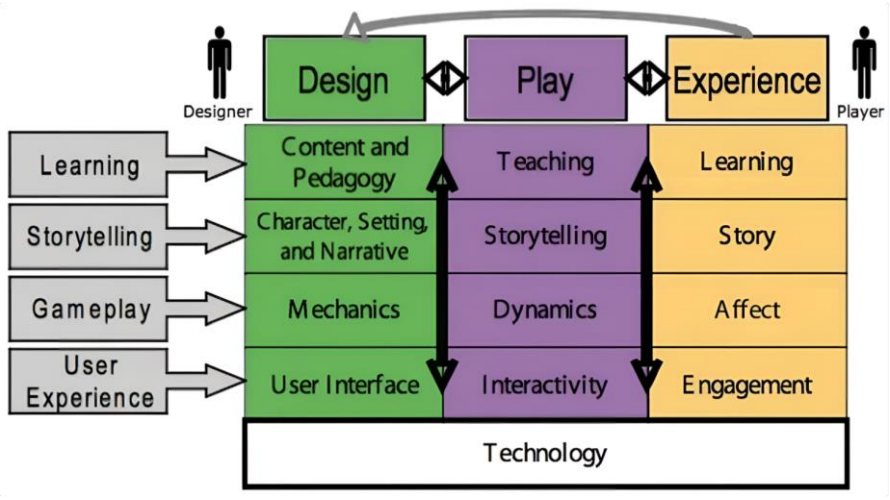

图 5.6 Design Patterns and Elements 框架

游谱 2.0 版本的游戏画像 Demo 版本的优势：

- **刻画维度有提升：**40 维覆盖游戏核心能力。
- **参考游戏设计框架：** 保证了游戏画像的科学性和合理性。
- **初步验证有效：** 通过对 10 款游戏进行打分，初步验证了游戏画像的有效性。

然而，游谱 2.0 版本的游戏画像 Demo 版本也存在一些局限性：

- **维度仍然有限：** 40 个维度可能无法涵盖所有游戏特性。
- **打分难度较高：** 需要大量人力进行人工打分，成本较高。
- **刻画粒度不够细：** 部分维度可能过于粗略，无法满足精细化运营的需求。

为了克服这些局限性，我们进行了更深入的研究，并最终抽象出近 2000 个游戏维度，并对上百款游戏进行了人工标注。基于这些数据，我们设计了专门的矫正算法对打分结果进行调整，并进行了包括新近用户预测、用户规模预估等在内的一系列实验，均取得了良好的效果。

游谱 2.0 版本的游戏画像的优势：

- **刻画维度更全面：** 近 2000 个维度能够更全面地刻画游戏的各个方面，包括玩法、手感等难以量化的特性。
- **打分结果更准确：** 通过人工标注和矫正算法，保证了打分结果的准确性和可靠性。
- **应用场景更广泛：**可以应用于游戏推荐、搜索、分类、评估等多个场景，并取得了良好的效果。

游谱 2.0 版本的游戏画像标志着我们在游戏知识图谱构建和应用方面取得了重要进展，为后续版本的迭代和完善奠定了坚实基础。

5.1.4.2.3 第三版

第一版本游戏知识库涉及的维度少，不过可以覆盖大规模的游戏。第二部版游戏画像可以深度刻画游戏，不过构建成本高。 通过总结了各自方法的优缺点，我们并取长补短，形成了多模态图谱的构建流程。主要包括持续迭代优化游戏画像的维度， 并扩大打分的范围。同时为了降低打分难度，我们会根据图谱在不同场景下的表现，对实体与属性进行调整， 使得更少的选项可以刻画更加丰富的内容。同时也在构建流程中引入更多的技术，让打分过程变得半自动话。这个过程会涉及知识图谱，NLP，CV，音频等诸多相关技术。 譬如非结构化数据处理中， 可以通过 LLM 提取实体间关系。或者通过游戏截图产生的隐空间向量表征来提取游戏画风等图片相关属性。 类似的在校验补全上也寻找相应的配套方案，譬如游戏名的相似并不仅仅基于文本来做，还可以根据游戏的图标来做相似度计算。

游谱 3.0 版本的优势：

- **多模态信息融合：** 不仅包含文本信息，还融合了图像、音效等多模态信息，能够更全面地刻画游戏的特性。
- **维度更丰富，成本更低：** 通过优化维度和引入半自动化打分技术，在保证刻画能力的同时降低了构建成本。
- **应用场景更广泛：** 可以应用于游戏推荐、搜索、分类、评估等多个场景，并取得了良好的效果。

通过多年积累， 目前游谱为游戏垂直领域规模最大的图谱。并且提供了一系列解决方案，包括新游戏发现、游戏及公司评估、发行运营等阶段的服务。这些服务可以帮助用户快速找到合适的游戏，预测游戏的市场表现和用户规模，以及提供精准的玩家画像和游戏推荐。譬如在新游戏拉新场景，推荐等场景效果提升在 10%+。最后我们总结相关经验参加了 OGB-LSC Wiki90mV2 的比赛，获得了第三名的好成绩，具体可以参考 Solution for OGB-LSC Wiki90mV2。游谱 3.0 版本的构建和应用，标志着我们在游戏知识图谱领域取得了重要进展。未来，我们将继续探索多模态信息融合、图神经网络等前沿技术，为游戏行业带来更多创新和价值。

5.1.4.3 游戏社交网络-游缘

社交网络通常指的是人与人之间的联系和互动，例如熟人网络、陌生人网络等。然而，除了这些常见的社交网络，许多垂直领域也存在着独特的社交网络，游戏领域便是其中之一。在游戏中，玩家之间会建立起各种社交关系，例如好友、队友、公会成员等。这些社交关系会对玩家的游戏行为产生重要影响，例如更紧密的社交关系往往意味着玩家在游戏中更活跃、粘性更强。

现实中的社交网络通常由许多小结构组成，例如社区、圈子等，其复杂度非常高。以大型游戏为例，其好友关系网络可能包含上亿个节点和几十亿条边，从中找出规律并进行有效分析，面

临着巨大的挑战。为了更好地刻画和分析游戏社交网络，我们构建了名为“游缘”的游戏社交网络知识图谱。游缘在基础的关系链数据上进行抽象，形成了包含网络标签和个人标签的社交关系画像体系。

5.1.4.3.1 高影响力玩家识别

在游戏社交网络中，一些玩家由于其活跃度、影响力等因素，对其他玩家产生着重要的影响。识别并利用这些高影响力玩家，可以帮助游戏运营者有效地提升用户活跃度和留存率。我们基于 Topical Affinity Propagation （TAP） 算法识别游戏内的高影响力玩家。TAP 算法主要基于概率图网络，通过计算影响力在网络中的传播过程，识别出对其他玩家影响最大的玩家。

- **考虑影响力传播：** TAP 算法不仅考虑玩家自身的活跃度，还考虑其在网络中的影响力，能够更准确地识别高影响力玩家。
- **无监督学习：** TAP 算法无需人工标注数据，可以自动学习网络结构，识别高影响力玩家。

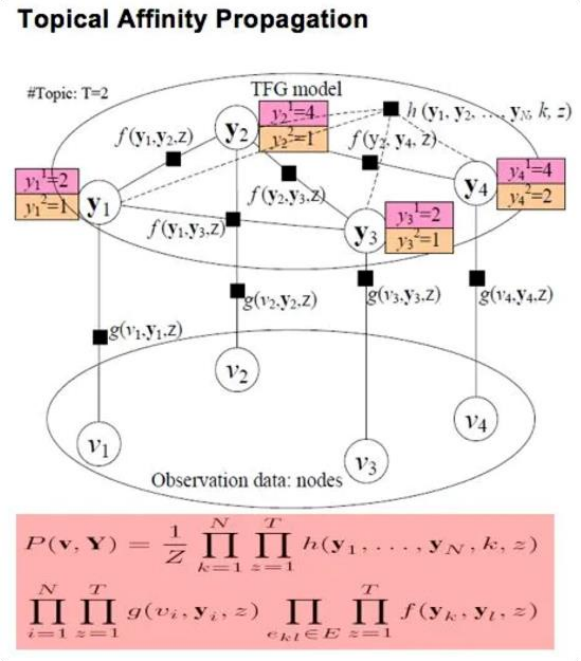

图 5.7 Topical Affinity Propagation 原理图[151]

基于识别出来的高影响力玩家，我们在多款业务做了好友召回活动。具体是为高影响力的玩家提供潜在召回玩家的列表， 由玩家自己选择应该召回谁。最终在多个业务上实验，并取得了不错的效果。

5.1.4.3.2 千亿图神经网络

随着游戏社交网络规模的增加，形成千亿的关系链和社区网络，涵盖了不同的社交关系。分析研究这种大规模游戏复杂网络，并维持它的稳定，成为了游戏生态中的重要一环。首先，需要

高效的图计算算法支持。其次，在真实场景，我们能收集到的标签往往很少。如何将这一大部分无标签数据为我们所用，提升在推荐任务上的成功率，是我们重点探索的方向之一。

业界主流的针对大图计算的优化方法主要是采样。Fastgcn 在每一层以度数为权重采样固定数量节点，采出来的子图可能过于稀疏。而 clustergcn 这样分而治之的方法，可以比较高效的实现高度并行。然而，大图分割难度大，metis 图分割算法不能处理千亿规模的大图。切图必然会带来信息损失，影响图计算效果。GraphSAGE 学习一个对邻居节点进行聚合表示的函数，以此来生成目标节点的嵌入向量，而 FastGCN 则采取不同的方法。FastGCN 直接对图中的节点进行采样，而不是邻居节点。它通过定义样本的损失，并使用蒙特卡洛近似方法计算样本梯度，从而进行积分计算。此外，还可以通过调整采样分布来减少近似方差。

为此，我们提出了 lps-gnn 框架，分别对图分割和子图数据增强进行了优化。此外，我们的框架可以灵活的选择任意 gnn 算法。首先对于大规模 gcn 框架的第一部分图分割算法。目前主流的图分割算法有两大缺点，一是能处理的图大小仍有限制，二是很容易行程超级社区，而超级社区对于并行计算的负载均衡性能有非常大的影响。在现有图分割算法里，metis 是表现较为稳定的经典算法。然而，metis 可以处理的图大小有限。为此我们设计了基于 label propagation 的 lpmetis 图分割算法。它结合了社区发现和 metis 的优点。我们首先用标签传播算法多伦迭代进行多层次图合并，在对最终的缩略图做 metis 分割图后，递归得到原图分区。标签传播算法保留了原图的重要结构信息。但是针对其容易形成超级社区的缺陷，我们设计了 stable 机制，在标签传播时同时考虑邻居的情况和子图的大小，以此来保障并行算法的负载均衡。最终我们的 lpmetis 图分割算法，可以在 9 小时跑完千亿大的图，并且与其他图分割算法对比，它切图更均匀，保留的边更多，速度更快，且切出来的图应用于下游 gcn 任务准确度更高。

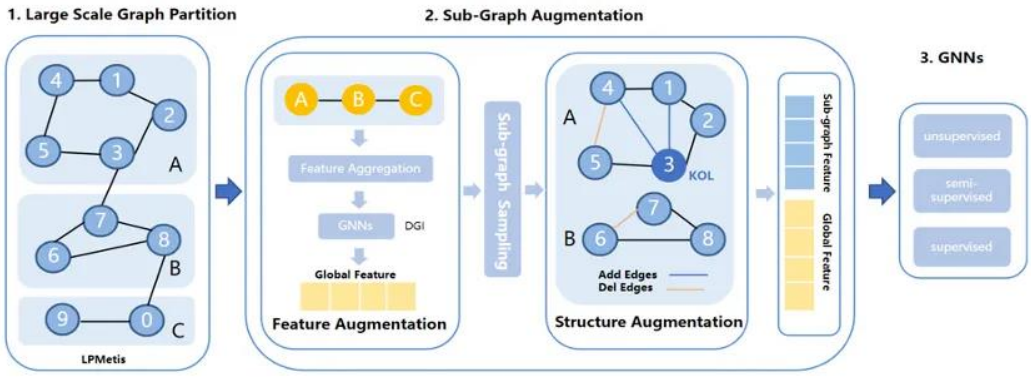

图 5.8 LPS-GNN 算法流程示意图

切图会带来信息的损失。为此我们设计了一系列子图数据增强的方法，来提升子图质量，已减小图切割带来的信息损失。我们首先借助大盘关系数据的力量，使用无监督 gnn 算法学习基于玩家大盘关系网络来得到玩家整体的游戏社交偏好作为单游戏单场景的信息补充。在识别作弊玩家这样的标签较少的场景下，可以有效的提升预测准确度。其次，考虑到每个子图只保留了局部信

息，我们将子图看成节点，子图之间形成有权重的边，构成了包含了全局信息的缩略图。对该缩略图进行表征学习可以得到整个大图的全局信息。除了特征增强，我们还尝试了对子图结构进行优化。我们尝试了不同的策略，比如随机删掉一些边，基于表征相似度来对图的边进行调整。但从结果显示，最优的方法是基于意见领袖对图的影响更大，我们去除了子图中 pagerank 最低的 5%的节点对子图结构进行去噪。这一步是带来了 4.6%的准确度的提升。此外，我们还发现对于超级大图，只采样一部分子图进行多伦迭代训练由于对全图进行训练。不仅时间显著下降，准确度也显著提升。

最终，应用到实际场景时，基于不同的目标和数据，我们可以灵活的选择合适的 gnn 算法。该框架在腾讯游戏多个场景落地，在线上 AB 实验中均获得显著效果提升。

5.1.4.3.3 好友推荐

在游戏内有很多排序的场景，需要给玩家按照他们的喜好来推荐他们可能感兴趣的好友、道具、游戏、玩法模式等。其中一个场景是好友召回活动，当玩家流失不再登录游戏时，我们会利用已经流失玩家的好友来邀请他们回归游戏，并给予双方奖励。为了实现这一目标，我们需要对活跃玩家的所有已流失好友进行排序，将更有可能接受邀请回归游戏的好友排在前面。

这个问题面临两个挑战。首先，玩家在游戏内的行为多样，我们需要利用他们的历史行为和特征来提高转化率。其次，现实场景中存在大量无标签数据。以好友召回活动为例，有标签的是指在往期活动中曝光过的好友，被邀请且回流是正样本，未被邀请和被邀请未回流是负样本。但只有 0.5%的数据有标签，我们需要利用 99%以上的无标签数据来提升推荐任务的成功率。

以前的推荐方法通常采用规则或节点分类方法，比如优先推荐与玩家交互更多的好友。但这些方法没有同时利用所有参与用户的特征和历史交互。为了解决这个问题，我们将好友排序问题重新定义为链路预测问题，判断两个玩家之间是否存在成功的邀请边。这样，我们可以同时考虑两个玩家的特征、历史交互特征和历史活动信息。我们尝试了多种链路预测方法，包括传统的启发式算法、基于 embedding 的方法、直接对边的特征训练分类模型以及使用模型自动学习权重的 bilinear 方法[146]。在离线实验中，bilinear 方法的效果明显优于其他方法。

然而，bilinear 方法没有充分利用交互特征，而我们观察到交互特征在预测邀请边时是重要的判断依据。受到知识图谱表征学习论文 ConvKB[147]的启发，我们设计了 Edge CNN 算法，将边两端玩家的特征和交互特征融合起来建模。该算法克服了 bilinear 方法没有考虑交互特征的缺点，在线上好友排序场景中提升了 4.23%的转化率。然而，EdgeCNN 只能学习同一维度特征的相关性，不能自由学习任意两维特征之间的关系。因此，我们设计了 EdgeTransformer[148]，利用多头注意力机制充分学习任意两个特征之间的相关性。在在线上实验中，Edge Transformer 进一步提升了 2.2%的转化率。

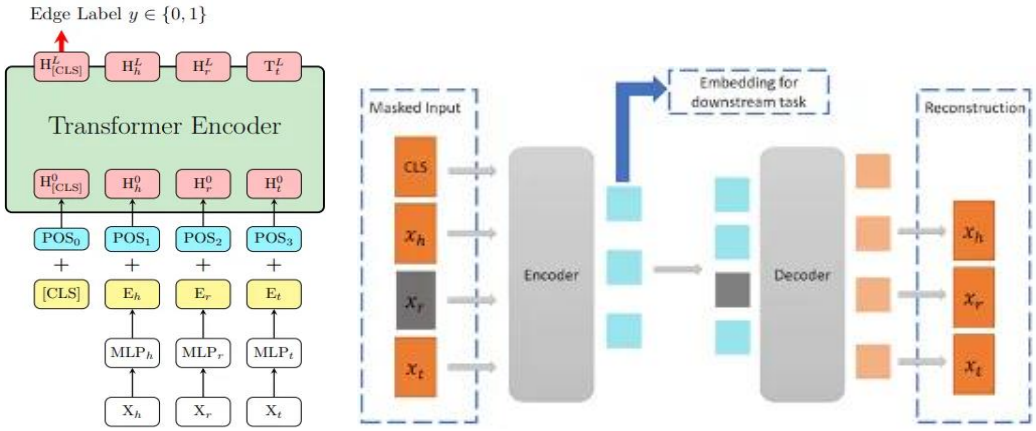

图 5.9 左：Edge Transformer 模型; 右：Edge MAE 模型

由于链路预测没有考虑图结构的全局信息，我们使用特征传播增强策略，在判断一条边是否存在时，考虑了该玩家的其他好友的状态和特征。这个操作可以用 spark 实现分布式并行计算，并且在 OGB 竞赛中获得了第一名[149]。此外，为了充分利用无标签数据，我们设计了 EdgeMAE 预训练模型[148]，通过 encoder-decoder 结构学习无标签样本中的网络结构信息。在有监督任务上，EdgeMAE 表现出了稳定的效果提升，并超过了学术界最前沿的算法。

我们的链路预测算法不仅在好友推荐场景中落地，还应用到了其他场景，比如道具推荐。在线上实验中，我们的算法带来了不错的购买率提升。

5.1.4.2 总结

腾讯游戏在游戏知识图谱构建和应用方面取得了丰硕成果，并将其成功应用于多个场景，有效提升了游戏运营和推荐的效率和效果。未来，腾讯游戏将继续探索 Graph+AI 技术，为游戏行业带来更多创新和价值。

5.1.5 犯罪网络检测

5.1.5.1 研究背景与挑战

在当今数字时代，加密通信技术已经成为网络通信的主流。随着超过80%的网络通信采用加密形式，犯罪网络检测工作面临着前所未有的挑战。为了应对这一挑战，本研究提出了一种创新方法，通过结合图技术与大语言模型(LLMs)来增强加密平台上的犯罪活动检测能力。

5.1.5.2 图技术应用基础

在犯罪网络分析领域，图技术展现出了独特的优势。它通过节点、边和属性的结构，能够直观而有效地展示各实体之间错综复杂的关系和交互模式。值得注意的是，作为一款专为 GenAI 时代设计的多模态数据库，ArcNeural 不仅具备强大的图数据处理能力，还可以同时处理文本、图像、音频等多种数据类型，为犯罪网络检测提供了全方位的技术支持。

5.1.5.3 创新技术方法

本研究开发的检测系统主要采用两种互补的技术方法。第一种是 Text2Cypher 方法，它能够将自然语言直接转换为 Cypher 图查询语言。在具体实现过程中，系统首先运用自然语言处理技术对输入内容进行语义分析，识别出关键实体、关系和查询意图。随后，系统将分析结果与预定义的查询模板进行匹配，并基于匹配结果生成相应的 Cypher 查询语句。最后，系统将生成的查询发送到图数据库执行，完成整个查询过程。

第二种是基于大语言模型的 GraphAgent 方法，这种方法专门用于处理较为复杂的查询场景。在实际操作中，系统首先利用 LLM 深入分析用户输入，全面理解查询的复杂性和所需的图操作。基于这种理解，系统会生成一个详细的查询执行计划，将复杂的查询任务分解为一系列可执行的步骤。接着，系统会选择并调用适当的 API 来执行这些步骤，最后将各个步骤的结果进行综合，形成最终的答案。

为了支持 GraphAgent 的高效运行，系统设计了两套完整的 API 体系。其中，基本图操作 API 主要负责节点度数获取、节点特征提取、节点连接检查等基础性操作。而高级业务 API 则封装了更为复杂的业务逻辑，能够执行通信频率分析、异常行为检测等高级功能。

5.1.5.4 实际应用策略

在实际应用中，系统采用了灵活的分层查询处理策略。对于简单的数据检索需求，如查找特定用户的 IP 地址，系统会优先使用 Text2Cypher 方法直接处理。而面对需要多步推理的复杂查询，例如分析过去一周内与多个犯罪分子保持频繁联系的用户，系统则会启用 GraphAgent 方法进行深入分析。在某些情况下，系统还会根据具体场景需要，灵活结合这两种方法的优势，实现更高效的查询处理。下面是一些实际使用中的效果截图：

图 5.10 通过图算法发现的隐藏团伙

图 5.11 通过 LLMs 辅助犯罪发现

5.1.5.5 实践经验与未来展望

通过实际应用，我们积累了丰富的实践经验。准确的问题复杂度分类和有效的工作流程设计是系统成功运行的关键因素。同时，采用迭代方法不断优化系统性能，并注重验证结果的合理性也十分重要。特别值得一提的是，思维链(CoT)和少样本学习的引入显著提升了系统的整体表现。

然而，系统当前仍面临着一些技术挑战。数据质量与完整性的保障、大规模图数据的实时处理、模型可解释性的提升、动态图分析的优化，以及跨语言和跨文化分析等问题都需要进一步研究和解决。

展望未来，我们的研究将重点关注图神经网络与 LLMs 的深度集成、联邦学习在隐私保护中的应用、自适应学习系统的研发，以及多模态数据融合技术的探索。这些研究方向将为加密环境下的犯罪网络检测带来新的突破和发展机遇。

5.2 科学研究

在这一节，我们将具体阐述 Graph+AI 在科学中的应用。近几年，随着人工智能的迅猛发展，人工智能与科学开始深度交叉融合，产生了一批令人瞩目的成果，AI for Science（科学智能）已经成为了一个大趋势。由于图（graph）结构在科学中的普适性，基于图的人工智能也在科学领域中成为最重要的范式和方法之一。因此，在本节中，我们将介绍 Graph+AI 如何具体用于各个科学领域，着重在地球科学、连续体仿真、材料科学、粒子物理、生命科学、运筹学等领域阐述。

5.2.1 地球科学

学科领域知识是隐藏在海量科学文献中的理论积淀和人类共识。构建领域知识图谱，有助于建立系统化、标准化、结构化的学科概念体系，将人类可读的非结构化数据语料，转变为机器可读、可计算、可推理的结构化数智资源。基于知识图谱检索、知识推理、统计计算，能够实现知识的定量关联、长程连接、与隐式依赖发现。

地学知识图谱能用于发现地学子学科的知识演化规律，发现物质和生命演化的新机理，推断不同系统间的指示作用，为地球资源的开发和利用提供决策支撑。因此，精准全面的领域知识图谱是地学科学家亟需的科研工具[1]。然而，现有的依赖专家经验的自顶向下构建方法效率极低。以地球科学为例，大量的地学实体散落在前沿的科学文献中，由于概念定义模糊、学术界未达成共识、研究课题小众等原因难以被有效穷举和人工定义。因此，本案例利用大模型的数据汇聚、理解与推理能力，从海量科学文献中自动挖掘有价值的实体、关系及属性，辅助科学家建立地学全领域最完整的知识图谱和有学科纵深的地学知识体系[90]。同时，利用知识图谱对地学学科体系化的建模与表示，为大模型提供可靠的信息检索服务[88]，能有效提升模型的分析推理能力，在关系推理、逻辑推理和推断任务中提高准确性，减少大模型的幻觉，提高可解释性[92]。

5.2.1.1 自底向上的知识图谱构建

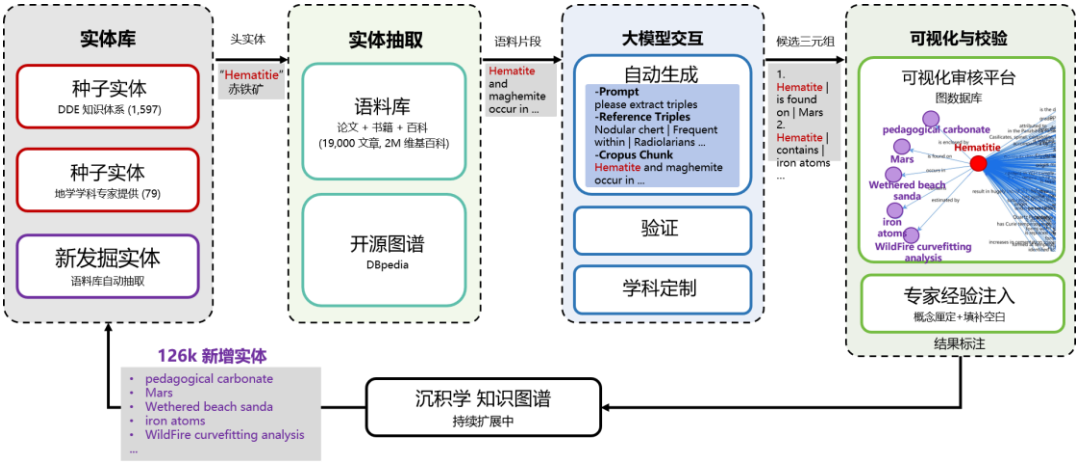

图 5.12 全域语料感知的大模型实体语义理解与隐含关联推断

自底向上（bottom-up）构建是借助大模型的技术手段，收集和整理公开采集的数据，再根据数据内容总结、归纳其特点，选择其中置信度较高的信息，添加到知识库中，然后对这些知识要死进行归纳和组织，逐步抽象为更高层次的概念，最终形成模式层。一般适用于涉及海量数据、内容繁杂且架构不清晰的公共领域通用知识图谱。

自底向上的构建方式基于专业领域的学术论文作为基础语料，采用人在回路的迭代扩展方式，从初始的种子节点出发，逐步扩展图谱的结构。这是一种三阶段的抽取算法，每一轮迭代中依次执行三个阶段流程，分别抽取知识图谱中的实体、关系和属性。首先建立包含人工种子实体和新

增实体的实体库，从种子实体出发执行第一阶段算法，依次将每个种子实体作为头实体，通过语料库和开源图谱检索，获得与当前头实体匹配的语料片段和三元组样例，通过大模型自动抽取、关联验证和学科规则过滤等操作，获得新发现的尾实体。随后，采用大模型优选策略，让大模型扮演专家角色来评判和筛选抽取的尾实体。然后执行第二和第三阶段算法，抽取关系和属性。抽取出的图谱导入可视化校验平台，通过专家经验注入进行校验，验证合格的实体和关系输出成知识图谱，同时动态扩充实体库。

5.2.1.2 自顶向下的知识图谱构建

通过自顶向下的方法，从百科类网站等结构化数据源中提取高质量数据，获取本体和模式信息，并将其整合到知识库中，即先为知识图谱设计数据模式（data schema），再依据设计好的数据模式进行有针对性的数据抽取。一般适用于数据相对集中、知识结构相对确定的垂直领域行业知识图谱。自顶向下的方式首先进行本体构建，构建知识图谱的模式层。从地学最顶层的概念开始构建顶层本体，然后细化概念和关系，形成结构良好的概念层次树。需要利用一些高质量数据源提取本体，即本体学习。然后进行实体学习，将知识抽取得到的实体匹配填充到所构建的模式层本体中。

5.2.1.3 基于地学知识图谱的应用

1、检索增强

地学知识图谱用于地学领域大模型的检索增强生成主要有两种技术路线：基于自然语言的图谱查询与子图匹配。

基于自然语言的图谱查询是将用户的自然语言输入翻译成精确的图数据库查询，对节点、路径、子图进行精确搜索，将查询的到的结果返回给大模型进行内容增强生成。

子图匹配广泛用于各个 GraphRAG 系统，首先基于聚类、社区挖掘等算法将知识图谱分割成语义关联紧密的（多级）子图，然后从用户输入中抽取出关键词，进行模糊匹配子图查询，查找出语义相关度最高的社区结构，返回给大模型进行内容增强生成。

2、推理预测

基于领域知识图谱的推理预测的任务可以用于矿产预测、地热资源利用、沉积学知识体系演化等学科应用场景。例如在矿产领域，矿产资源预测是国家重大战略需求，通过精准人工智能技术研发推动破解成矿理论和精准矿产预测的技术难题。新能源革命导致铜供应的巨大缺口，斑岩铜矿供应了全球超过 70%的铜，易于寻找的矿床已寻找殆尽，找矿方向转向深部与覆盖区等高难度地区，现有小模型+专家经验的方法预测精度低、可解释性差。如何利用知识图谱和大模型解释推断矿物形成、分布、预测的机理，完成斑岩铜矿的高精度、端到端、可解释性好的定位与定量预测，是一个关键的科学问题。基于知识图谱进行图计算、数据统计和关联分析的知识检索、知识

分析和基于机器学习的知识推理等，实现间接、模糊、复合等多种关系的发现，发现弱关联和间接关联的实体，识别实体间的长程关联和隐式依赖，进而发现化学元素与成矿之间的指示作用和不同矿产形成的新机理。

5.2.2 材料学

材料体系的仿真计算常规手段主要依赖于密度泛函理论（density functional theory，DFT）计算，具有计算成本高昂、适用尺度小等局限性。因此，如何利用机器学习技术有效加速或替代 DFT 计算，一直以来都是学术与工业界亟待解决的难题。

2018 年由 Xie 等人提出的 CGCNN[68]通过将晶胞表示为图（节点为原子元素属性特征向量，边为原子间距离的高斯基展开特征向量），利用图卷积神经网络（graph convolutional neural network，GCNN[69]）提取晶体特征。在每层网络中，每个节点的邻居节点特征和连边特征级联后，通过一个线性变换后得到信息，并与各节点经过一个不同线性变换后的特征表示相加并偏置后，经过激活层（activation）变换，得到该层输出的节点特征。经过多层的信息汇聚后，池化所有节点特征向量，得到代表全局特征的向量用于晶体属性预测任务。CGCNN 在 DFT 计算得到的训练数据上学习，实现了接近 DFT 计算精度的属性预测。

在此基础上，ALIGNN[70]则进一步引入键角信息以优化材料属性对键角信息敏感的问题，该模型由一个原子图（节点表示原子，边表示键）和对应的原子线图（节点表示键，边表示键角）构成。通过在两个图上交替进行图卷积操作，将键角信息通过原子间键的表示传播到原子节点的表示，反之亦然，该模型实现了更优精度的晶体属性预测。

类似的，M3GNET[71]利用包含键长、键角在内的原子三体信息更新原子间连边的特征向量，并通过图卷积操作更新原子节点表示，在得到的原子特征上汇聚了更为丰富的原子间相对几何信息。该模型基于大量第一性原理分子动力学（AIMD，ab initio molecular dynamic）数据上对原子间势能（interatomic potential，IAP）进行训练学习，可作为通用力场模型广泛用于不同化学体系的晶体结构松弛、动态模拟和性能预测，以高效率和低计算成本为昂贵的 AIMD 模拟提供了替代选择。

相比于直接对 DFT 计算得到的属性进行学习，DeepH[72]则利用电子相互作用和 DFT 原子轨道基组的局域性质，通过局域坐标系的构建，利用晶体图（节点表示原子属性，边表示原子间距）对 DFT 计算中间量哈密顿矩阵（Hamiltonian）进行学习，并通过哈密顿矩阵确定性计算得到晶体属性，取代了原本 DFT 计算中最为耗时的自洽场（self-consistent field，SCF）迭代过程。得益于哈密顿矩阵元仅取决于材料局域结构的性质，该模型能够在 DFT 成本较低的小型晶体结构-哈密顿矩阵数据上进行训练，学习不同局域结构对应的哈密顿矩阵元，并在常规情况下 DFT 计算成本难以接受的大尺度材料结构（例如转角双层二维材料和纳米管）上进行推理得到其哈密顿矩阵，并用于计算各项材料属性。

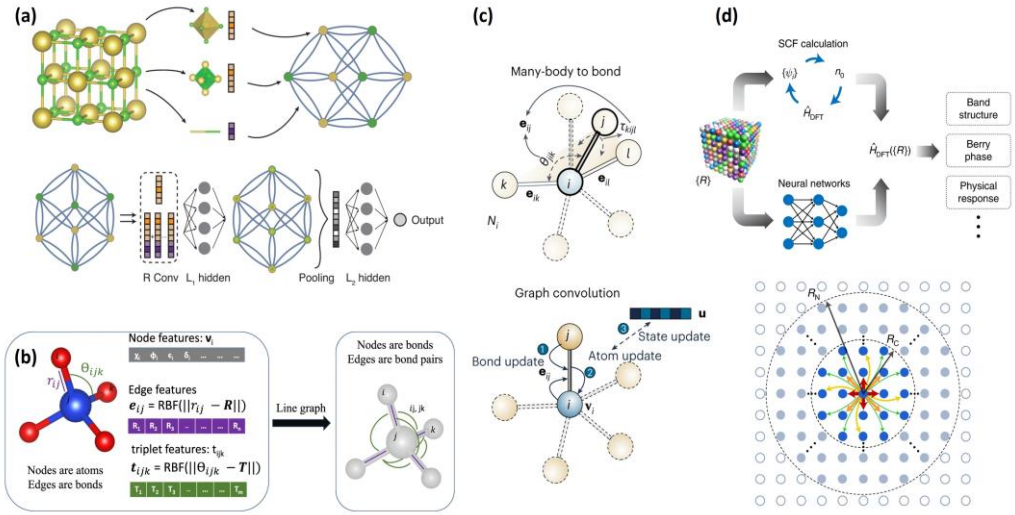

图 5.13 (a) CGCNN 晶体图表示的构建与模型架构；（b）ALIGNN 原子图与原子线图的构建；（c） M3GNET 晶体图表示及信息聚合示意图；（d）DeepH 工作流示意图；电子相互作用的局域性质，只有范围内原子基组间的哈密顿矩阵元不为 0，只有范围内的原子参与信息传递

除了在晶体材料中的应用，图神经网络也被用于多晶材料以及无规非晶体材料的特征提取与属性预测。Hu 等人[73]通过将多晶材料中不同晶粒表示为包含其取向角度、体积、相邻晶粒个数的特征向量，基于晶粒间的连接关系构建图结构，利用可学习的信息传递更新节点特征，并将所有节点特征级联后经过 MLP 输出预测材料整体属性。该模型可将外加磁场强度作为 MLP 共同输入实现对不同磁场下材料磁性响应的预测，并分析材料中不同晶粒对宏观相应的贡献。

Kondor 等人[74]基于无规体系中原子的位置构建图，利用信息传递神经网络（message passing neural network，MPNN）更新节点和连边特征，并将节点特征加和后得到全局特征，用于分类该体系为玻璃态或液态。在此过程中，将各个节点的连边特征经过经过单层无偏置的 MLP 后与可学习的注意力向量点乘，得到注意力分数，并归一化后作为权重更新各连边特征。通过这样的自注意力机制，可以显式区分图结构中对分类结构影响大的连边，为模型推理的可解释性提供依据。

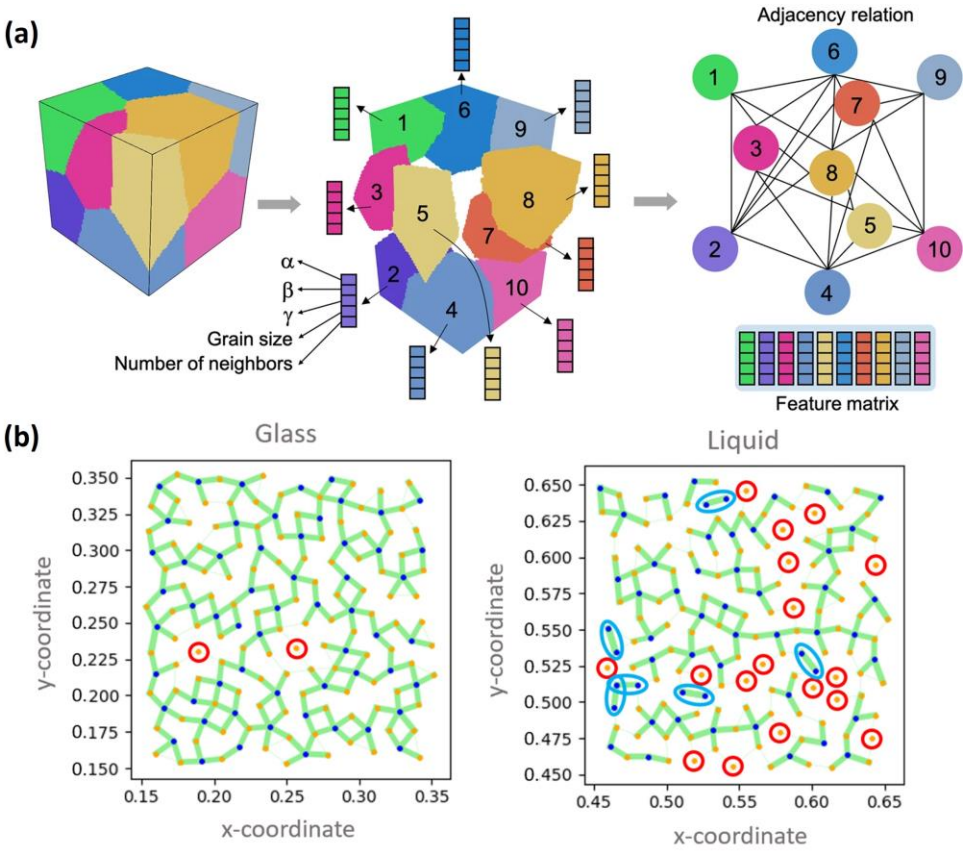

图 5.14 （a）多晶材料图结构的构建，其中晶粒由取向角度、体积、相邻晶粒数表示，这些特征均可由实验表征获得；（b）仅根据原子位置对材料处于玻璃态或液态进行分类的 GNN 模型，其中红色和蓝色圈标注了由高注意力权重边连接的原子及原子团

5.2.3 物理学

5.2.3.1 连续体仿真

许多的物理系统的动力学都可以通过偏微分方程（PDE）来描述，我们称之为连续体（continuum）。其可以建模流体、固体、等离子体、大气等系统。对于这些系统，一个重要的问题就是如何更快、更好地对其进行仿真，预测其状态随时间的演化。同时基于此，人们可以进行下游的设计、控制等任务。

对于连续体系统，进行仿真时，往往将其在空间离散化为网格（mesh）或者粒子（particle）表示，在时间离散化为离散时间的状态（图 5.15）：

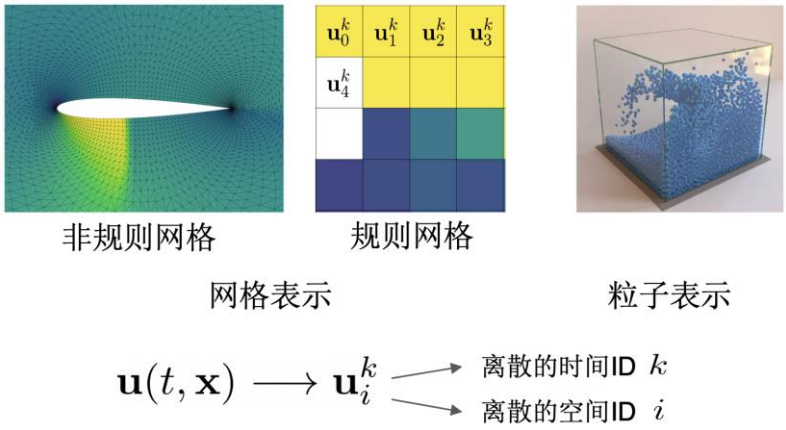

图 5.15 连续体系统的网格表示和粒子表示

传统方法的仿真，往往在离散的空间和时间上，基于有限元、有限差分或者有限体积等数值方法求解，尽管比较精确，但速度非常慢，往往几小时或者更长时间才能模拟一个千万或者上亿网格的系统。近几年，基于代理模型的深度学习方法迅猛发展，其主要优势是速度快，往往比传统数值方法快一到四个数量级。具体来说，该方法通过一个参数化的代理模型f_θ，其输入为系统在 t 时间的状态u_t以及参数 a、外界控制w_t等，输出为对系统 t+1 在时间状态u_{t+1}的预测$\hat{u}_{t+1} = f_\theta(u_t; a, w_t)$，在训练时通过最小化预测误差$\ell(\hat{u}_{t+1}, u_{t+1})$学习系统的演化。这里$\ell$为损失函数，通常为均方误差（MSE）。推理时，该代理模型便可基于初始条件和边界条件，通过自回归方式模拟系统的演化，从而预测系统未来长时间的状态。由于根据系统的数据结构，可选用不同架构的代理模型。常用的架构为图神经网络，U-Net，变换器（Transformer），神经算子等。以下将选取一些代表性工作具体介绍。

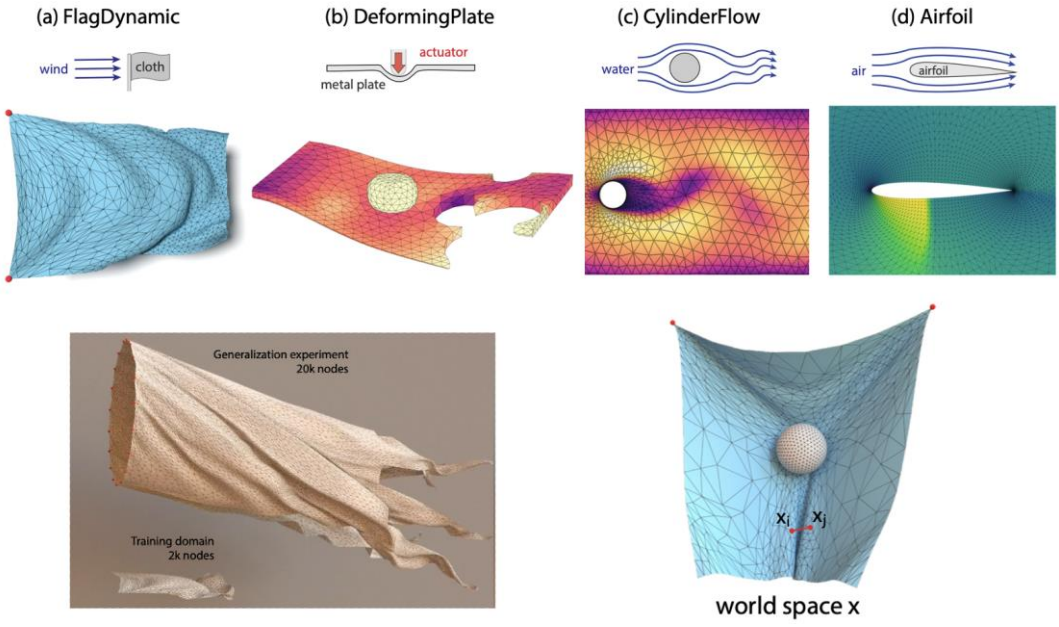

图 5.16 MeshGraphNet[63]将物理系统建模成非规则网格，可模拟各类物理系统的演化

2021年，DeepMind提出MeshGraphNets（网格图网络）[63]，通过非规则网格建模多个刚性或柔性物体，并模拟他们之间的相互作用，例如，该方法可以模拟布料、板材、圆柱扰流、机翼等的运动（图5.16）。

在具体方法上，该工作将每个物体建模为一个单独的非规则网格（Irregular mesh）。接下来，MeshGraphNets作为代理模型，由t时刻系统的状态（网格上每个顶点的位置和速度）预测t+1时刻的状态（图5.17）。具体来说，模型的输入首先进入编码器，将网格每个顶点的位置和速度映射到一个高维隐空间，作为图的节点特征。同时，构建两类边：网格上的连边，以及物理世界的连边（如果两个节点在物理世界的距离小于某阈值）。

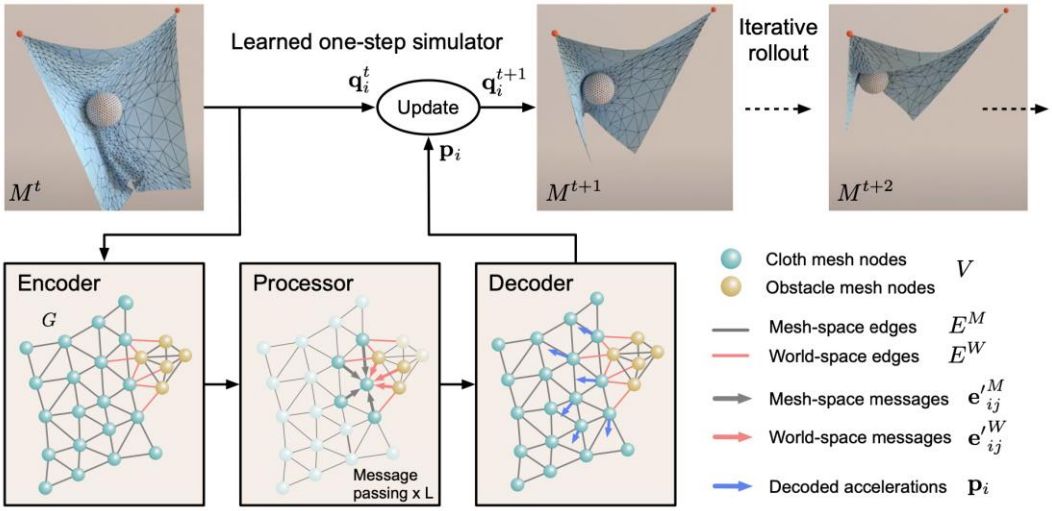

图5.17 MeshGraphNets[63]网络架构。系统在t时刻的状态M^t（包括网格结构，每个格点的位置、速度）作为输入，同时，连接物理空间的连边。接下来，经过编码器、消息传递和解码器，预测t+1时刻的状态改变。

接下来，通过层的消息传递（message passing）实现图的状态更新。每一层的信息传递包括三步：**（1）相互作用的学习：**在每一条边上，根据边的特征和相邻的两个节点的特征，通过一个多层感知机（MLP）预测该边上的消息（message）；**（2）相互作用聚合：**在每个节点上通过相加方式聚合所有相邻边的消息，包括网格连边和物理世界连边；**（3）节点特征更新：**在每个节点上，将聚合的消息与节点当前层特征进行级联，通过另一个多层感知机预测该节点下一层的特征。最后，输入解码器，预测下一个时间每个节点的状态改变。训练时，通过最小化预测误差，通过反向传播学习以上的编码器、消息传递和解码器的参数。以此，MeshGraphNets可以学习不同网格表示的物理内部以及之间的相互作用，以及这些相互作用如何影响每一个节点的状态。该工作在一系列物理系统的仿真任务中进行了评估，取得了优异的效果。

基于图神经网络的代理模型已应用部署在多个具体工业场景中。例如，[64]提出混合图网络模拟器（图5.18），首次将图神经网络用于大规模地下流体仿真并在工业管线中部署。该方法首先将地下三维区域划分成一个六面体网格（规模可达百万或千万量级）。图的每个节点为每个单元格

的特征，包括地下水、石油、天然气的饱和度、压强等，通过混合图网络模拟器模拟每个单元格中状态的演化。为提高长期预测的准确度，该方法提出最小化多步自回归预测误差的训练目标。针对网格的大规模的特性，完整的图无法放在一个 GPU 里面，该方法提出基于子图的训练方式，使得在训练中的每个样本为一个子图（五万节点量级），但是在推理中模型能泛化到百万甚至千万量级的网格。实验表明，该方法比传统求解器加速 18 倍，能够准确预测系统 20 个月的状态演化。该方法已经部署在沙特阿美（世界最大石油公司）的仿真管线中，配合原有求解器加速下游的设计和推理任务。

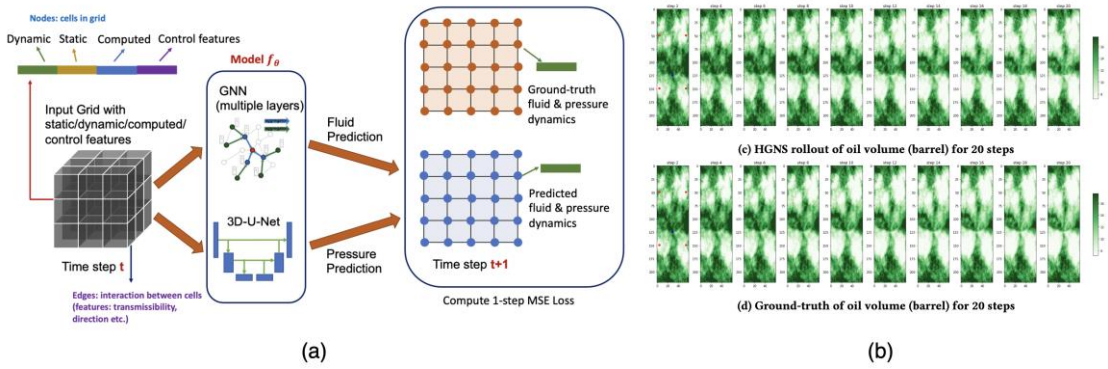

图 5.18 混合图网络模拟器架构（a）和对石油饱和度 20 个月的预测结果以及与真实值比较

除了图神经网络这一最常用的神经网络架构之外，其他架构（包括 Transformer、神经算子）也是AI用于连续体仿真的常用架构。例如，几何信息神经算子（Geometry-Informed Neural Operator）[65]首先通过一个图神经算子（GNO）[66]作为编码器，把复杂的几何形状映射到规则网格空间，接下来，通过多层的傅里叶神经算子层在傅里叶空间和原空间将规则网格上编码的状态进行处理，最后，通过另一个图神经算子作为解码器，映射回原空间的预测。通过这一方式，文章表明该方法在三维汽车风场预测等任务中达到了最先进的性能（state-of-the-art）。此外，工作[67]基于 Transformer 架构，将非规则的网格通过注意力机制映射到固定长度的词汇单元（token），接下来，在隐空间进行演化，最后通过注意力机制预测给定空间位置的系统状态。

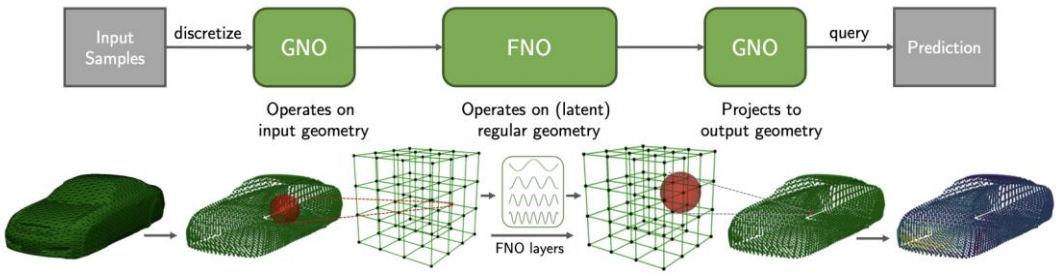

图 5.19 几何信息神经算子架构

5.2.3.2 将对称性嵌入神经网络

对称性，是各类物理系统中非常普适的性质。例如，分子的演化过程对于坐标的平移和旋转具有等变性，粒子物理射流标记任务对于粒子交换和洛伦兹变换具有不变性等等。将对称性植入

神经网络，能够使得这些对称性被严格满足，提高预测的准确度，同时大大减少需要的训练样本的数量。以下，选取一些代表性工作进行介绍。

在工作[78]中，作者根据以下两个定理构建对于各类对称群不变或者等变的神经网络：（1）向量输入的函数返回不变标量的充要条件为它可以写为只有输入向量的不变标量乘积的函数；（2）向量输入的函数$h(\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n)$返回等变向量的充要条件为它可以写为不变标量函数乘以输入向量的线性组合，即

$$h(\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n) = \sum_{t=1}^{n} f_t\left(\langle \mathbf{v}_i, \mathbf{v}_j \rangle_{i,j=1}^{n}\right)\mathbf{v}_t$$

这里$\langle \mathbf{v}_i, \mathbf{v}_j \rangle$为向量$\mathbf{v}_i$和$\mathbf{v}_j$的内积，具有对相应对称操作的不变性，$f_t(\cdot)$是一个标量函数。基于以上这两个定理，这一工作构建了对于平移、旋转、洛伦兹变换、庞家来群、交换不变或者等变的神经网络，是一个非常普适的基于对称性构建神经网络架构的方法。

由于图神经网络在各类物理系统中的广泛应用，将常见的旋转和平移的等变性植入成为一个重要问题。等变图神经网络（EGNN[82]）将图节点特征分为坐标特征和其他特征，通过构建对于坐标特征等变的消息传递和节点特征更新，实现其对坐标旋转、平移和镜面反射的等变性。其在多体问题、图重建、分子性质预测等任务中，实现了很大的准确度提升，展示了对称性的植入对于神经网络的重要性。

此外，洛伦兹群神经网络（Lorentz Group Network[83]）将洛伦兹群分解为不可约表示，通过张量积实现对于洛伦兹变换等变的神经网络。工作[84]通过将神经网络参数通过特定方式捆绑（weight tying）实现其等变性。除了以上对于全局对称性的植入，工作[85] [86]提出方法将局域的规范对称性嵌入神经网络，可以建模更广泛的、非欧式的几何空间。

5.2.3.3 粒子物理

粒子物理研究物质和力的基本定律。通常，研究人员通过构建巨大的粒子加速器，例如大型强子对撞机（LHC），将粒子加速到几乎光速，通过相反方向粒子的对撞产生出新的粒子（每秒有超过四千万次的粒子对撞）。这些新粒子往往具有极短的衰变时间，其迅速进行多次衰变成为多个粒子，形成窄锥体形态的粒子射流（jet），并在探测器上形成轨迹（track）。这里最重要的任务，就是根据探测器上的轨迹，重建（reconstruct）以上过程，并推测最初形成的粒子。由于图神经网络能充分利用节点和边的特征并进行整合，非常适合于这一重建任务。具体来说，重建任务包括以下子任务：（1）寻迹（track finding）：通过图构建、边分类、图分割，将探测器上的信号重建成为不同粒子的入射轨迹（图 5.20）；（2）射流标记（jet tagging）：基于以上重建信息，对射流的起始粒子进行分类；（3）事件（event）分类：基于以上信息，对最初的产生新粒子的物

理过程进行分类。通常，这些加速器产生的数据量极大（可达每秒 TB 量级），因此，对人工智能方法的精度、速度和可扩展性提出了很高的要求。以下，将介绍在粒子物理中 Graph+AI 的几个代表性工作。这些工作将对称性通过不同方式嵌入神经网络，实现更高的精度和样本效率。

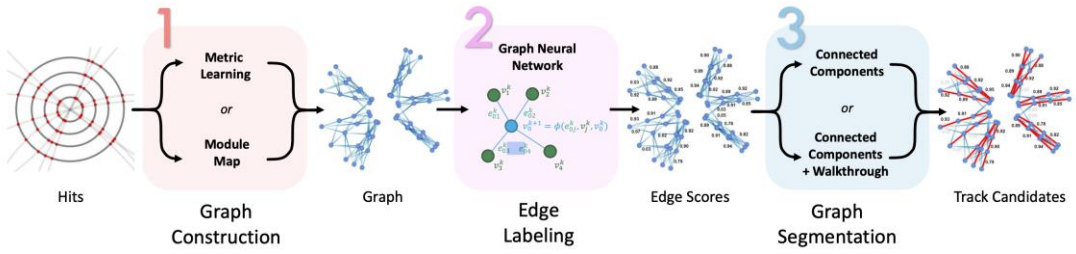

图 5.20 图神经网络用于寻迹（track finding）：包括图构建、边分类、图分割，将探测器上的信号重建成为不同粒子的入射轨迹

对于射流标记（jet tagging）任务，ParticleNet[75]（图 5.21（a））将射流看成不同粒子（particle）的集合，构建一个对于粒子的 ID 交换不变的神经网络。具体来说，其使用 EdgeConv（边卷积层[76]）对每个粒子周围的 k 个最近邻粒子进行消息传递和平均聚合，通过多层 EdgeConv 的叠加聚合更远邻居的信息，最后，通过对粒子的全局池化（global pooling）预测该射流的标签。可以证明，这一方法对于输入的任意 ID 的交换具有不变性。与之前没有考虑对称性的方法相比，该方法实现了更高的准确度。

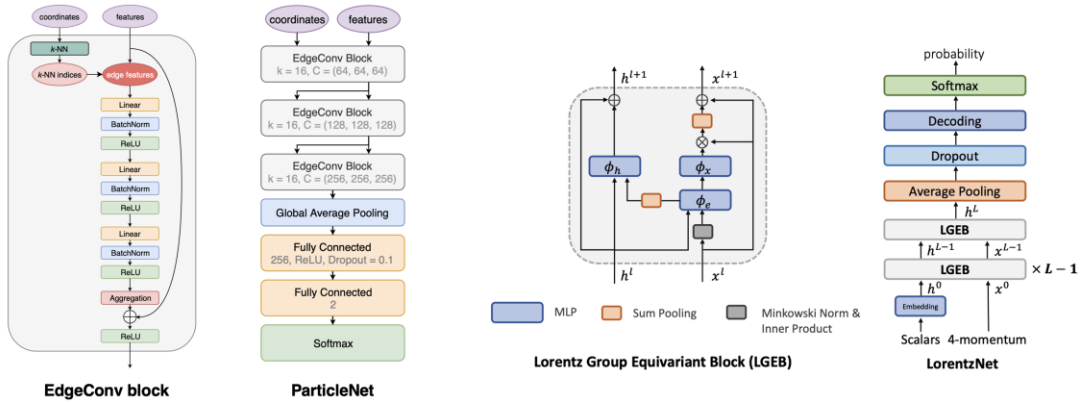

图 5.21 (a)ParticleNet 和其 EdgeConv 层；(b)LorentzNet 和其 LGEB 层

与 ParticleNet 不同，LorentzNet[77]（图 5.21（b））考虑粒子满足的另一个基本对称性：洛伦兹对称性。其根据洛伦兹对称性的万能逼近定理[78]，提出对洛伦兹变换等变的图神经网络层。具体来说，其在第层的消息由以下公式给出：

$$\mathbf{m}_{ij}^{l} = \phi_e\left(h_i^l, h_j^l, \psi\left(\left\|\mathbf{x}_i^l - \mathbf{x}_j^l\right\|^2\right), \psi\left(\langle \mathbf{x}_i^l, \mathbf{x}_j^l \rangle\right)\right)$$

其中为第个节点第层的隐空间标量，为第个节点第层的坐标嵌入，为闵可夫斯基点积，其对洛伦兹变换不变。 是一个可学习的神经网络。经过求和聚合，第个节点的第层的坐标嵌入为：

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + c\sum_{j\in[N]} \phi_x\left(\mathbf{m}_{ij}^l\right)\cdot \mathbf{x}_j^l$$

这里是一个标量函数。可以证明，以上构建的图神经网络满足满足洛伦兹对称性的万能逼近定理，其对于洛伦兹变换具有等变性。该方法在顶夸克标记和夸克-胶子标记任务中，其准确度相比起其他模型有显著提升。

除了结合对称性的图神经网络之外，基于 Transformer 的架构也可被用于粒子物理的任务，例如 NodeFormer[79]，SGFormer[80]等。尤为突出的是，最近的一个架构 HEPT[81]提出基于局部敏感哈希（LSH）的高效的点变换器（point transformer），实现了近线性的算法复杂度，并且对于具有局部归纳偏差的任务，可证明较低的近似误差。相对于其他基于图神经网络和 Transformer 的基线模型，其在 GPU 上实现了 203 倍的加速，并达到了目前最高的准确度。

5.2.4 生命科学

生命科学是研究生物体及生命过程的学科，涉及到化学、生物学、医学、药学等相关领域。生命科学中的研究工作产生并使用大量的关系型和图结构数据[93]。因此，图作为一种强大的表示工具，正展现出广阔的应用潜力。同时，专家知识常常通过图（谱）进行结构化沉淀，使复杂的信息和经验能够以系统化的形式保存和传播。知识图谱（KGs）是用于建模专家知识的一种流行方法，广泛应用于各种系统和应用中。目前，它们已成为许多语义网页搜索引擎和问答系统在学术界和工业界的核心支柱[94]。生命科学知识图谱旨在构建一个结构化的知识网络，该网络涵盖了生命科学领域的广泛知识，包括但不限于化学、生物学、医学、药学以及跨学科研究等。通过将科学事实、理论、实验数据、研究成果及科学家之间的关联以图谱的形式组织起来，生命科学知识图谱能够增强科学理解的深度、促进新发现以及加速科学传播，为解决复杂科学问题提供强有力的支持。

5.2.4.1 生命科学知识图谱的构建

生命科学，作为探索生命本质及其运作机制的庞大领域。构建生命科学知识图谱是一项旨在系统化整理并关联这些海量知识的开创性工作。这项工作的起始，首先是认识到传统信息存储和检索方式已难以满足当今生命科学研究的深度与广度需求。随着基因测序技术的飞速进步、生物大数据的爆发式增长，以及人工智能技术的应用拓展，构建一个全面、精准且可计算的生命科学知识图谱显得尤为重要且迫切。这一知识图谱的构建，从定义核心实体（如基因、蛋白质、疾病、代谢途径等）开始，通过标准化的词汇和本体（如 Gene Ontology、Human Phenotype Ontology 等）

对这些实体进行精确描述，并利用先进的计算技术手段捕获它们之间的复杂关系，如基因调控关系、蛋白质互作网络、疾病与基因的关联等。这样的知识图谱将成为生命科学研究人员的强大工具，不仅能够加速新知识的发现，还能促进药物研发、精准医疗、生物技术等领域的革新，为解开生命奥秘、改善人类健康提供坚实的基础。因此，启动构建生命科学知识图谱的旅程，是迈向生命科学新时代的关键一步。

生命科学知识图谱的构建大致可以分为六个阶段：

1、数据源的选择

这个阶段要确定生命科学知识图谱需要使用的数据源，根据结构化、半结构化和非结构化数据源的不同，进而影响到知识抽取阶段技术的选择。结构化数据指的是以固定的格式存储，通常是表格或者公共数据库（例如 UniProt[95]或 ChEMBL[96]）中的数据。半结构化数据指的虽然没有严格的结构，但包含一些可以被识别的模式或标记，使得数据在一定程度上被解析，通常是 XML 文档。非结构化数据指的是缺乏预定义的格式或结构，即自由文本源（例如 PubMed）。

2、本体的构建

这个阶段是为特定领域创建形式化知识模型的关键步骤。本体不仅定义了领域中的概念和类别，还描述了它们之间的关系，以促进知识的共享和重用。构建本体的过程通常包括需求分析、概念建模以及最终的验证，确保其在实际应用中的有效性。

3、生命科学知识抽取

这一阶段的重点是从生物医学文献和数据库中提取原始信息。通过应用自然语言处理（NLP）、信息检索和文本挖掘等技术，研究人员能够自动化识别重要的实体（如基因、蛋白质、疾病）及其相互关系。这一过程的核心任务是从大量的文本和数据中找到相关的知识，并初步分类和组织，以便为后续阶段提供数据基础。此阶段输出的结果往往是未经过深度验证的、初步提取的知识，属于知识体系构建的原始输入。

4、知识摄取和验证

在此阶段，目标是从抽取的原始知识中进行深入的验证和整合。不同来源的知识数据可能存在模糊性、不完整性或冗余现象，因此需要通过知识建模和对齐技术对这些知识进行清理和合并。同时，通过与本体的对照和注释，确保知识的语义一致性和结构完整性。该阶段不仅是对知识进行摄取和整理，更重要的是对其进行验证和完善，确保其准确性和可靠性，最终为科研和临床决策提供高质量的知识基础。

5、生命科学 KG 存储及检验

知识图谱的存储需要高效的数据管理系统，通常使用图数据库，如 Neo4j 或 GraphDB，以便支持复杂的关系查询和大规模数据处理。检验知识图谱的质量和准确性至关重要，通常通过一致性

检查、完整性验证和语义推理等技术进行。这一过程确保图谱中的信息可靠且有助于推理新的知识。

6、生命科学 KG 维护和优化

这个阶段强调生命科学知识图谱的维护和优化。这一过程涉及定期更新和扩展知识图谱，以确保其准确性和时效性。维护工作包括监测新数据源、整合来自最新研究的发现，并纠正潜在的错误。此外，优化技术如版本控制和数据清洗被广泛应用，以提高知识图谱的质量和可用性。通过持续的维护，生命科学知识图谱能够适应快速变化的科学环境，支持更有效的研究和临床应用，促进跨学科的合作和知识共享。

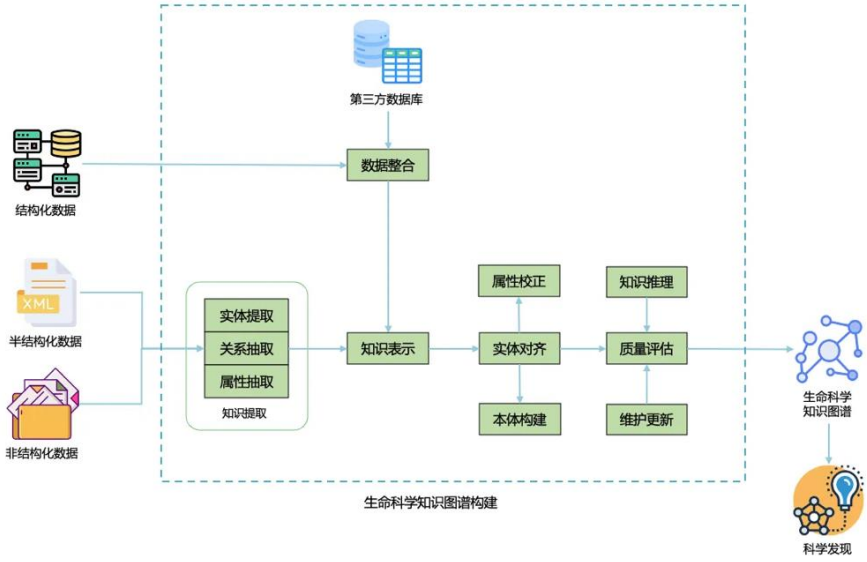

图 5.22 生命科学知识图谱构建

5.2.4.2 利用生命科学知识图谱进行科学发现

对人工智能技术的研究（包括机器学习和基于知识图谱的推理）旨在加速科学发现的步伐，这一领域正处于新兴且快速发展的阶段。其挑战在于帮助科学家揭示新的知识和解决方案，例如发现新的治疗机会，识别用于治疗复杂疾病的候选分子药物或现有药物的新用途，以及支持更加个性化的预测。

1、药物发现和治疗

药物发现涉及探索及其庞大的潜在药物候选空间，人工智能能够在进行昂贵的实验之前筛选出最有前景的候选者从而加速这一过程。最新研究表明，融入现有的表达性事实知识能够有效提升下游机器学习任务的效果。比如为了加强蛋白质语言模型（PLM），诸如 OntoProtein[97]和[98]等模型均采用了蛋白质序列的知识图谱，并通过基因本体（GO）的文本注释进行增强。Otter-Knowledge[99]融合了来自多种来源和不同模态的异构知识图谱（基于模式，包含概念及其属性），即每个节点都有特定的模式来限定其类型（文本、图像、蛋白质序列、分子等），并根据其模态

计算初始嵌入。然后使用图神经网络（GNN）丰富蛋白质和分子的表示，并训练模型以生成最终节点嵌入。该模型能够为训练期间未见的实体生成表示，并在药物-靶点结合亲和力预测的治疗数据公用库（TDC）基准测试中实现最新的结果[100]。TxGNN[101]在从各种知识库构建的大型异构多关系疾病和治疗候选者知识图谱上进行预训练。TxGNN 根据疾病的邻近蛋白、暴露和其他生物医学实体，获得每种疾病的特征向量，以计算疾病相似性，并预测药物的适应症或禁忌症。

2、预测蛋白质功能

使用物理实验来识别蛋白质功能是非常耗费时间和资源的，随着生命科学知识图谱的发展，利用其中的蛋白质知识图谱进行蛋白质功能的预测近年来得到了广泛的研究。例如InstructProtein[102]通过将蛋白质的功能注释、分子功能和生物过程等信息构建为知识图谱，模型能够系统地学习蛋白质与其功能之间的复杂关系。这种结构化的信息帮助模型理解蛋白质的功能注释，并通过因果关系推断出蛋白质的具体功能。DeepGraphGO[103]通过结合蛋白质序列和蛋白质网络数据，使用图神经网络（GNN）来增强蛋白质功能预测。同时也有一些研究尝试进一步利用 GO 中定义的函数间关系从而获得更好的性能。例如，DeepGOZero[104]和 HMI[105]在训练多标签分类器以预测蛋白质功能时，采用了包括类层次结构、类互斥公理及 OWL 中的复杂类限制在内的形式语义作为附加约束条件。蛋白质功能预测是一个典型的多标签分类问题，其中标签之间的复杂关系在知识图谱（KG）中被定义，能够被很好地用于增强蛋白质功能预测的性能。

3、使用本体和临床数据对医疗保健进行预测

数字医疗涉及利用临床数据和本体进行预测，包括诊断（例如罕见疾病）和程序预测（例如重症监护病房再入院）。一个相关概念是个性化医疗，它通过匹配和融合来自不同源头的知识来实现，并在预测任务中发挥重要作用，从而增强对不良反应或因缺乏足够标注数据集的罕见疾病的预测。例如，SHEPHERD[106]融入了一个多关系知识图谱（从 PrimeKG[107]提取），包含疾病、表型和基因信息，并利用患者模拟数据发现患者临床、表型和基因信息之间的新关联，以加速罕见病的诊断。GraphCare[115]通过利用大型语言模型和外部生物医学知识图谱来构建患者特定的知识图谱，并结合双注意力增强的图神经网络，以提高基于电子健康记录的医疗保健预测任务的准确性。DALK[116]通过动态地将大型语言模型和知识图谱相结合，从科学文献中提取关于阿尔茨海默病（AD）的结构化知识，并通过粗到细的采样方法与自感知知识检索技术，选择适当的知识来增强 LLMs 的推理能力，从而提高对 AD 相关问题的回答准确性。[108]构建了一个 KG（使用表达性 OWL 本体）来通过 BioPortal 中各种生物医学本体的语义注释丰富 EHR 数据来预测 ICU（重症监护病房）再入院风险。这些预测基于 KG 嵌入（例如 RDF2vec、OPA2vec 和 TransE）以及经典机器学习方法（例如逻辑回归、随机森林、朴素贝叶斯和支持向量机）。

4、分子属性预测与分子图对比学习

传统的分子属性实验研究耗时且资源密集，随着生命科学知识图谱的发展，通过分子知识图谱进行分子属性预测和分子图对比学习的应用日益广泛，为加速科学发现带来了新可能。例如KCL[117]通过构建化学元素知识图谱（KG）并利用其指导分子图的对比学习，从而在不需要人工标注的情况下，利用自监督信号来学习分子的表示，为药物设计和分子属性预测等下游任务提供支持。KANO[114]通过整合化学元素知识图谱（ElementKG）来增强分子属性预测任务，利用生命科学领域内的基本化学知识，通过对比学习预训练和功能提示微调，提高了模型在多个分子属性预测数据集上的性能，并为科学发现提供了新的视角。

5.2.4.3 利用生命科学知识图谱增强可解释性人工智能

人工智能（AI）和机器学习（ML）方法被广泛应用于解决许多领域的复杂问题，包括化学或生物医学等生命科学领域。然而，其中许多方法都是作为“黑匣子”运行的，无法使领域专家理解其预测背后的推理。这是一个主要问题，特别是对于对人类生命有潜在影响的领域尤其是生命科学领域。此外，理解人工智能方法的工作原理在生命科学应用中也至关重要，解释预测过程可以帮助阐明自然现象。

在生命科学领域有两个不同的受众：科学家（研究人员）和医疗保健从业者。对于科学家而言，解释被用作理解生命科学研究中科学发现的指南。因此，解释可能存在于假设或研究项目的明确背景中。另一方面，从业者直接参与影响医疗保健的决策。他们需要在开放的环境中考虑模型的输出，有时还需要向非领域专家的患者解释输出。使用知识图谱可以显著提升可解释性人工智能的质量，因为知识图谱非常适合提高模型的可解释性、可说明性和可理解性。

1、用于医疗保健实践的可解释人工智能

在医疗保健实践中，人工智能的广泛应用引发了将生命攸关的决策交给“黑盒”模型的担忧。由于医疗决策对患者的健康至关重要，可解释性成为确保人工智能能够被临床医生信任和采纳的关键因素。特别是在精准医疗领域，医生不仅需要依赖模型提供的治疗建议，还必须清楚地理解模型是如何得出这些结论的。模型仅给出一个决策结果是远远不够的，医生还需要通过可解释的推理过程来验证这些决策是否符合医学知识和患者的具体情况。因此，可解释人工智能在医疗领域的核心任务是提供透明、可审查的推理路径，确保每个决策都能经得起临床验证。这将通过补充而非替代临床医生的解释，增强医患双方对模型的信任。举例来说，这一方向已在多个医疗场景中被预见。可解释的人工智能模型能够辅助专家根据患者的病史找到合适的临床试验。对于可能导致严重后果的反直觉或不可靠预测，通过解释可以加以阐明并预防。还有人设想利用此类模型来解释并驳斥与医疗健康相关的错误信息。值得注意的是，应根据目标受众采用不同类型的解释，例如针对证据的科学解释，或针对治疗的轨迹式解释。

2、可解释的人工智能知识发现

生命科学知识图谱不仅在知识发现过程中发挥着基础性作用，也为推理和解释提供了结构化支持。生命科学知识图谱通过明确地表示实体及其相互关系，使复杂的生物学现象具备了更强的可解释性。它不仅帮助模型在推理过程中发现潜在的关联，还能够提供系统化的解释框架，揭示这些关联背后的因果机制。例如，Bresso 等人[109]利用从知识图谱中抽取的特征（如路径、邻近节点、路径模式等可解释特征）与白盒模型（如决策树）来复现专家对药物是否引发特定不良反应的分类。从决策树中提取的规则包含了能够依据专家意见解释这些不良反应背后的分子机制的特征。Sousa 等人[110]则利用知识图谱，基于共享的语义层面，同时解释蛋白质-蛋白质相互作用预测及基因-疾病关联预测。

3、用于 KG 构建的可解释人工智能

近年来，知识图谱的构建逐渐依赖于数据驱动的深度学习方法，即自动从数据中提取知识进行构建。这些深度学习模型具有不透明性，因此需要解释性，由此生成的知识图谱可能在下游应用中缺乏可靠性。文献[111]提出了可信的知识图谱工程，强调在知识图谱生命周期中嵌入可解释的人工智能和人类干预的重要性。在从文本构建知识图谱过程中，可解释人工智能方法已被应用于许多与自然语言处理相关的任务（如实体和关系抽取、实体解析、链接预测等）。这些方法依赖于基于特征的解释或基于知识的解释。基于特征的解释试图从数据或模型对数据的理解中推导出解释，而基于知识的解释则旨在通过规则、推理路径和结构化的上下文信息来解释过程。

5.2.5 运筹学

5.2.5.1 组合优化与图

运筹学中一个重要的基本问题是求解组合优化问题，即在一个有限的可行解的集合中找出一个最优的解。从优化的角度看，组合优化问题的决策变量是离散的，其可行解集也是离散的。组合优化问题在许多科学与工程领域都起着举足轻重的作用，从基本的数学问题如图着色问题，到生物、化学及工程科学中的蛋白质设计、分子结构设计、药物设计、材料设计、集成电路，到流程工业、复杂制造过程以及物流等领域。

许多组合问题具有图数据结构，比如旅行商问题、选址问题及图着色问题，如图所示。

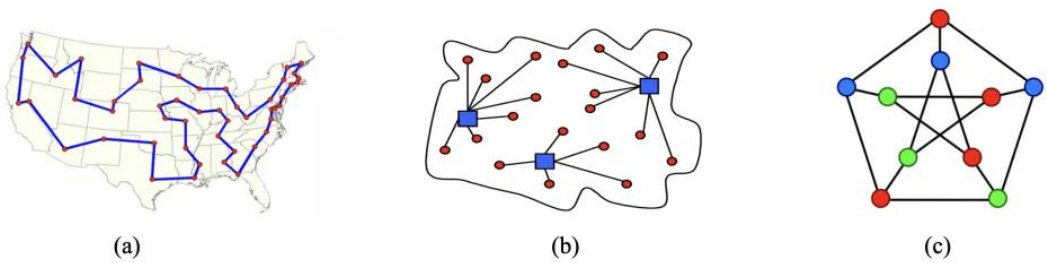

图 5.23 (a) 旅行商问题；(b) 选址问题； (c) 图着色问题

事实上，几乎所有的组合优化问题大概可以分成两类，一类可以自然地表示成图，而另一类则需要通过转换才能表达成图结构，如图所示[128]。

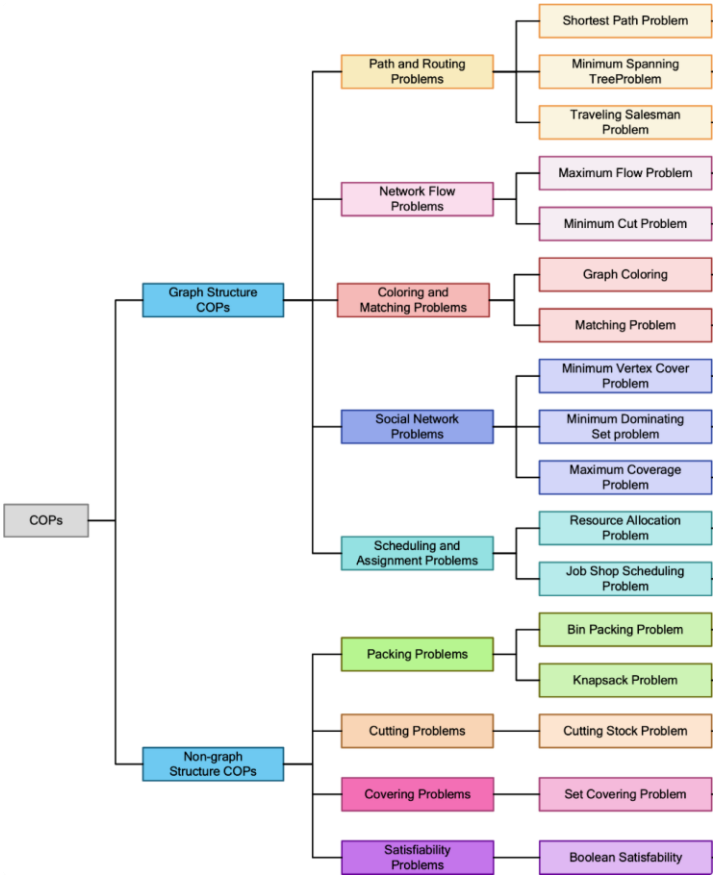

图 5.24 组合问题分类

5.2.5.2 基于图神经网络的组合优化问题求解

目前，图神经网络在组合优化求解中得到了广泛应用。这些应用大致可以分成两大类。第一类是借助图神经网络提取优化问题的特征(embeddings),以更有效地用求解器求解这些问题。比如，文献[129]中采用图神经网络提取有效特征,辅助强化学习算法求解高铁重调度问题，不但能提升最优解的性能，也大大缩短了求解所需要的时间。图 (a)给出了一个 5 辆列车及 3 个车站的调度方案与有向图的对应关系，图 (b)给出了强化学习的基本框架，而图 (c)则说明图神经网络在强化学习算法求解高铁重调度问题中的辅助作用。

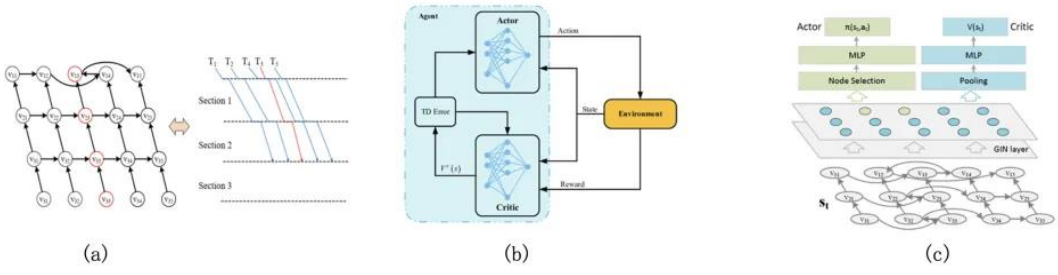

图 5.25 (a) 高铁调度方案的图表示; (b) 强化学习; (c) 图神经网络的特征提取

第二类方法则是一种数据驱动的组合优化问题求解方法。其基本思路是用以往求解过的案例训练图神经网络，然后在求解新问题时直接通过图神经网络进行推理。以两目标选址问题为例[130]我们可以训练两个图神经网络，分别学习每个候选地址被选中的概率，以及每个选中的地址给客户的供应关系（用概率矩阵来描述）。图 (a) 为选址问题示意图，图 (b) 则为基于图神经网络的求解框图。图 (a)中，方块表示所有的候选地址，实心方块则表示被选中的地址，而所有的圆点则为客户的位置，被选中的地址与客户之间的的实线则为供应关系。在这个两目标优化问题中，第一个目标为建造成本及运输成本的最小化，第二个目标则为可靠性最大化。当两个图神经网络经过已有的最优选址案例训练后，则可以通过采用推理采样获取一个新问题的 Pareto 最优解，如图图 (b) 所示。

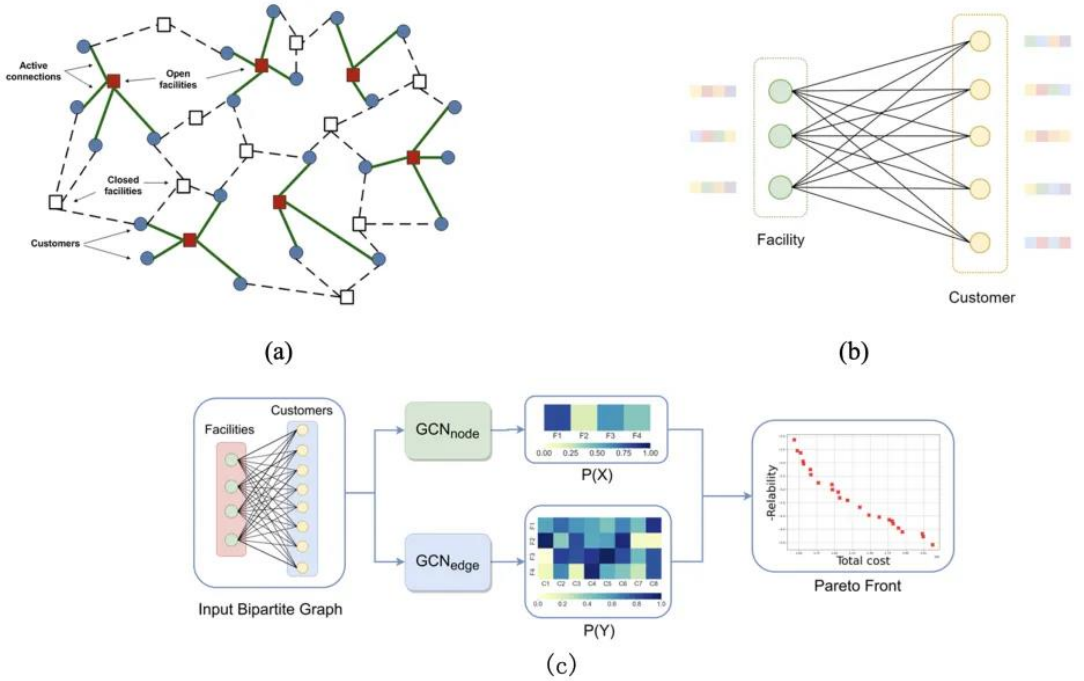

图 5.26 (a) 选址问题示意图；(b) 选址问题的二部图表示；(c) 基于图神经网络的选址问题求解

不管是第一类方法还是第二类方法，均需要预训练图神经网络。图神经网络的训练一般分成如下几个步骤[130]。首先，将低维的图特征映射到高维的潜空间。在选址问题中，对节点而言，其特征为成本（候选地址）或需求（客户），而边的特征则为运输成本及可靠性。之后，在潜空间再对相邻节点或边的高维特征进行消息传播（message passing），如图 5.27 (a)。这种高维特征进行 2～3 轮消息传播后则作为图神经网络的输出用于完成下游任务，比如作为一个多层感知器（MLP）的输入，如图 5.27 (b)所示。该 MLP 的输入为某个节点的 embedding，而输出则为相应节点被选中的概率。在训练时，MLP 输出的真实值为 0（选中）或 1（未选中）。在推理时，根据新的问题图神经网络输出的 embedding 作为 MLP 的输入计算这个概率。最后根据获得的概率（候选地址）或概率矩阵（供应关系）获得一些列 Pareto 最优解。

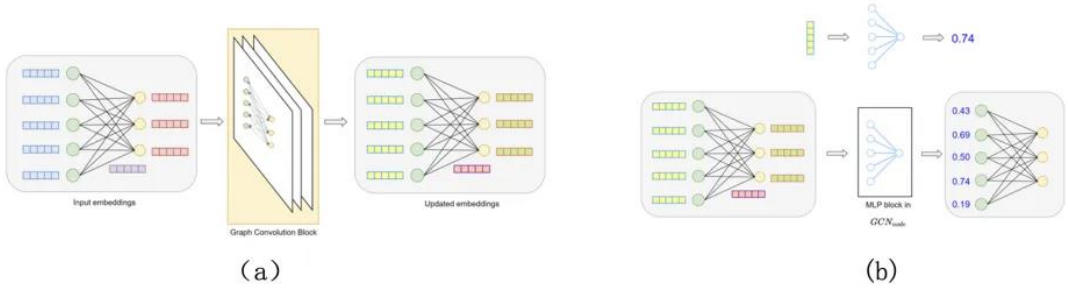

图 5.27 (a)选址问题的图神经网络训练；(b) 图神经网络的输出（embedding）作为下游任务（MLP）的输入

在选址问题的求解中，我们采用有监督学习训练下游的多层感知器。在求解其他组合问题如图着色问题时，则可以用无监督学习算法求解[128]。值得注意的是，用于求解组合优化问题的图神经网络类型，消息传播方法以及损失函数等，都需要根据具体的问题确定。另外，对于复杂的组合优化问题及动态组合优化问题，需要进行自动分解或搜索最优的图神经网络架构进行信息传播。

第 6 章 总结与展望

在本白皮书中，我们系统地探讨了图计算技术与人工智能的结合应用，涵盖了从基础技术到实际应用的各个方面。具体而言，我们深入分析了以下几个关键技术领域：

图数据：我们讨论了图数据的构建，强调了其在表示复杂关系和结构化数据方面的优势。我们总结了图数据增强技术和图采样技术，展示了如何通过这些技术提升图数据的质量和处理效率。

图神经网络：我们详细介绍了图神经网络的几种经典卷积和池化算子，深入探讨了图神经网络的训练和推理技术。我们还从多方面介绍了如何提升图神经网络的可信性。

图基础模型：我们探讨了图基础模型的概念、实现方法及其未来的发展前景。图基础模型为构建更复杂的图应用提供了坚实的基础，具有广泛的应用潜力。

知识图谱：知识图谱作为一种特殊的图结构，广泛应用于知识表示和推理。我们介绍了知识图谱的构建、维护和服务，展示了其在增强人工智能系统理解和推理能力方面的巨大潜力。

图应用：我们列举了图计算技术在多个领域的实际应用案例，包括能源、金融、科学研究等，尤其是与大模型结合的能力，展示了图计算技术在解决实际问题中的广泛应用前景。

此外，我们还提供了若干解决方案，展示了如何将图计算技术人工智能进行结合并应用于具体的业务场景中，以供企业和研究机构参考。

展望未来，图计算技术与大模型的结合将进一步推动图计算及人工智能的发展，带来更多创新和突破。以下是我们对未来发展的几点展望：

图技术与大模型的融合：随着预训练语言模型的广泛应用，图技术与大模型的结合将成为趋势。通过将图结构信息融入大模型中，可以显著提升模型在复杂关系和结构化数据处理方面的能力。同时，仿照大语言模型的模式，训练基于图的基础模型，可以解决多种图任务，进一步提升图技术的应用广度和深度。

跨领域应用的扩展：图技术在各个领域的应用潜力巨大，未来将看到更多跨领域的应用。例如，在医疗健康领域，结合图神经网络和大模型，可以实现更精准的疾病预测和个性化治疗。

增强解释性和可解释 AI：图技术在提升模型解释性方面具有独特优势。通过图结构，可以更直观地展示模型的推理过程和决策依据，增强模型的透明性和可解释性。这对于构建可信赖的人工智能系统至关重要，尤其是在涉及敏感数据和决策的应用场景中。

大规模图数据处理：随着数据规模的不断扩大，处理大规模图数据的能力将成为关键。未来的研究将致力于提升图算法的效率和可扩展性，开发能够处理超大规模图数据的分布式计算框架和优化算法，从而支持更大规模和更复杂的图应用。

开放图数据和标准化：为了促进图技术的发展，开放图数据和标准化工作将变得越来越重要。通过建立统一的数据标准和共享平台，可以加速图技术的研究和应用，推动整个生态系统的发展，促进跨领域和跨机构的合作。

总之，图计算技术与以大模型为代表的人工智能技术的结合，不仅将为人工智能领域的发展带来新的机遇，也将进一步促进图计算技术自身的发展。随着不断的研究突破和应用创新，我们有理由相信，这种结合将共同推动图计算技术和人工智能技术的协同发展，最终引领我们进入一个全新的图智能时代。

参考文献

[1] 叶育鑫，刘家文，曾婉馨，等.基于本体指导的矿产预测知识图谱构建研究[J].地学前缘，2024，31（04）:16-25.

[2] 张继元，钱育蓉，冷洪勇等.基于深度学习的命名实体识别研究综述[J].现代电子技术，2024，47（6）：32－42.

[3] Perozzi B, Al-Rfou R, and Skiena S. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.

[4] Grover A, Leskovec J. node2vec: Scalable feature learning for networks. Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 2016.

[5] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding. Proceedings of the 24th international conference on world wide web. 2015.

[6] Zhou J, Cui G, Hu S, et al. Graph neural networks: A review of methods and applications. AI open, 2020.

[7] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations, 2017.

[8] W. L. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. Neural Information Processing Systems, 2017.

[9] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. International Conference on Learning Representations, 2018.

[10] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. Modeling relational data with graph convolutional networks. European Semantic Web Conference, 2018.

[11] Ying Z, You J, Morris C, et al. Hierarchical graph representation learning with differentiable pooling. Neural Information Processing Systems, 2018.

[12] Gao H, Ji S. Graph u-nets. Proceedings of Machine Learning Research, 2019.

[13] Lee J, Lee I, Kang J. Self-attention graph pooling. Proceedings of Machine Learning Research, 2019.

[14] Diehl F. Edge contraction pooling for graph neural networks. arXiv preprint arXiv:1905.10990, 2019.

[15] Shi, C., Wang, X., & Yang, C. （2023）. Advances in Graph Neural Networks. Springer.

[16] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. （2020）. A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32（1）, 4-24.

[17] Zhao, T., Jin, W., Liu, Y., Wang, Y., Liu, G., Günnemann, S., ... & Jiang, M. （2022）. Graph data augmentation for graph machine learning: A survey. arXiv preprint arXiv:2202.08871.

[18] Hamilton, W., Ying, Z., & Leskovec, J. （2017）. Inductive representation learning on large graphs. Advances in neural information processing systems, 30.

[19] Rong, Y., Huang, W., Xu, T., & Huang, J. （2019）. Dropedge: Towards deep graph convolutional networks on node classification. arXiv preprint arXiv:1907.10903.

[20] Luo, Y., McThrow, M., Au, W. Y., Komikado, T., Uchino, K., Maruhashi, K., & Ji, S. （2022）. Automated data augmentations for graph classification. arXiv preprint arXiv:2202.13248.

[21] Li, R. H., Yu, J. X., Huang, X., & Cheng, H. （2014, March）. Random-walk domination in large graphs. In 2014 IEEE 30th International Conference on Data Engineering （pp. 736-747）. IEEE.

[22] Wang, Qiange, et al. "Neutronstar: distributed GNN training with hybrid dependency management." Proceedings of the 2022 International Conference on Management of Data. 2022.

[23] Thorpe, John, et al. "Dorylus: Affordable, scalable, and accurate {GNN} training with distributed {CPU} servers and serverless threads." 15th USENIX Symposium on Operating Systems Design and Implementation （OSDI 21）. 2021.

[24] Wang, Minjie, et al. "Deep graph library: A graph-centric, highly-performant package for graph neural networks." arXiv preprint arXiv:1909.01315 （2019）.

[25] Wang Q, et al. Neutronstar: distributed GNN training with hybrid dependency management[J]. Proceedings of the SIGMOD 2022.

[26] Jia Z, Lin S, Gao M, et al. Improving the accuracy, scalability, and performance of graph neural networks with roc[J]. Proceedings of Machine Learning and Systems, 2020, 2: 187-198.

[27] Wang Q, Chen Y, Wong W F, et al. HongTu: Scalable Full-Graph GNN Training on Multiple GPUs[J]. Proceedings of the ACM on Management of Data, 2023, 1（4）: 1-27.

[28] Ai X, Wang Q, Cao C, et al. NeutronOrch: Rethinking Sample-Based GNN Training under CPU-GPU Heterogeneous Environments[J]. Proceedings of the VLDB Endowment, 2024, 17（8）: 1995-2008.

[29] Zhang X, Shen Y, Shao Y, et al. DUCATI: A dual-cache training system for graph neural networks on giant graphs with the GPU[J]. Proceedings of the ACM on Management of Data, 2023, 1（2）: 1-24.

[30] Huang Y, Cheng Y, Bapna A, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism[J]. Advances in NIPS, 2019, 32.

[31] Rajbhandari J, Rasley J, Ruwase O, He Y, et al. ZeRO: Memory optimizations Toward Training Trillion Parameter Models[C]//Proceedings of the SC20,2020: 1-16.

[32] Gandhi S, Iyer A P. P3: Distributed deep graph learning at scale[C]//. Proceedings of the OSDI21,2021: 551-568.

[33] Karypis G, Kumar V, et al. METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices[EB/OL]. 1997.

[34] Lin Z, Li C, Miao Y, Liu Y, Xu Y. Pagraph.Scaling GNN training on large graphs via computation-aware caching[J].Proceedings of the 11th SOCC. 2020: 401-415.

[35] Wang Q, et al. Neutronstar: distributed GNN training with hybrid dependency management[J].Proceedings of the SIGMOD 2022.

[36] Wan C, Li Y, Li A, et al. BNS-GCN: Efficient Full-Graph Training of Graph Convolutional Networks with Partition-Parallelism and Random Boundary Node Sampling[C]//Proceedings of the 2022 MLSys Conference. 2022: 1-17.

[37] Hönig R, Zhao Y, Mullins R, et al. DAdaQuant: Doubly-adaptive quantization for communication-efficient federated learning[C]//Proceedings of the International Conference on Machine Learning. PMLR, 2022: 8852-8866.

[38] Ho Q, Cipar J, Cui H, et al. More effective distributed ml via a stale synchronous parallel parameter server[J]. Advances in NIPS, 2013, 1223-1231.

[39] Liu, J., Yang, C., Lu, Z., Chen, J., Li, Y., Zhang, M., ... & Shi, C. （2023）. Towards graph foundation models: A survey and beyond. arXiv preprint arXiv:2310.11829.

[40] Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., & Wang, L. （2020）. Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131.

[41] Hou, Z., He, Y., Cen, Y., Liu, X., Dong, Y., Kharlamov, E., & Tang, J. （2023, April）. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In Proceedings of the ACM web conference 2023 （pp. 737-746）.

[42] Gui, A., Ye, J., & Xiao, H. （2024, March）. G-adapter: Towards structure-aware parameter-efficient transfer learning for graph transformer networks. In Proceedings of the AAAI Conference on Artificial Intelligence （Vol. 38, No. 11, pp. 12226-12234）.

[43] Sun, M., Zhou, K., He, X., Wang, Y., & Wang, X. （2022, August）. Gppt: Graph pre-training and prompt tuning to generalize graph neural networks. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining （pp. 1717-1727）.

[44] Wang, X., Wang, D., Chen, L., Wang, F. Y., & Lin, Y. （2023, September）. Building transportation foundation model via generative graph transformer. In 2023 IEEE 26th International Conference on Intelligent Transportation Systems （ITSC） （pp. 6042-6047）. IEEE.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, GabrielGoh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[46] Zhang, X. and Zitnik, M. GNNGuard: Defending Graph Neural Networks against Adversarial Attacks. In NeurIPS, 2020.

[47] Jin, W., Ma, Y., Liu, X., Tang, X., Wang, S., & Tang, J. （2020, August）. Graph structure learning for robust graph neural networks. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining （pp. 66-74）.

[48] Bifet, A., & Gavalda, R. （2007, April）. Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM international conference on data mining （pp. 443-448）. Society for Industrial and Applied Mathematics.

[49] Rossi, R., & Ahmed, N. （2015, March）. The network data repository with interactive graph analytics and visualization. In Proceedings of the AAAI conference on artificial intelligence （Vol. 29, No. 1）.

[50] Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y. （2016）. A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering, 29（1）, 17-37.

[51] Shi, Y., Jiang, G., Qiu, T., & Yang, D. (2024). AgentRE: An Agent-Based Framework for Navigating Complex Information Landscapes in Relation Extraction. arXiv preprint arXiv:2409.01854.

[52] Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., & Zhou, J. (2023). Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. arXiv preprint arXiv:2308.15363.

[53] Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[54] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv preprint arXiv:2305.14314.

[55] Newman, Mark EJ. "The structure and function of complex networks." SIAM review 45.2 (2003): 167-256.

[56] Fortunato, S. (2010). Community detection ingraphs. Physics reports, 486(3-5), 75-174.

[57] Lin, Dan, et al. "Analysis and mining of blockchain transaction network." Blockchain Intelligence: Methods, Applications and Challenges (2021): 41-71.

[58] Zhang Z, Li H, Zhang Z, et al. Graph meets llms: Towards large graph models[C]//NeurIPS 2023 Workshop: New Frontiers in Graph Learning. 2023.

[59] Heng W., Shangbin F., Tianxing H., Zhaoxuan T., Xiaochuang H., Yulia T.. Can Language Models Solve Graph Problems in Natural Language? NeurIPS, 2023.

[60] Jiabin T., Yuhao Y., Wei W., Lei S., Lixin S., Suqi C., Dawei Y., Chao H.. GraphGPT: Graph Instruction Tuning for Large Language Models. SIGIR 2024: 491-500.

[61] Yoon, M., Koh, J. Y., Hooi, B., & Salakhutdinov, R. (2023). Multimodal graph learning for generative tasks. arXiv preprint arXiv:2310.07478.

[62] Yang, C., Bo, D., Liu, J., Peng, Y., Chen, B., Dai, H., ... & Shi, C. (2023). Data-centric graph learning: A survey. arXiv preprint arXiv:2310.04987.

[63] Sanchez-Gonzalez, A., Godwin, J., Pfaff, et al. （2020, November）. Learning to simulate complex physics with graph networks. In International conference on machine learning （pp. 8459-8468）. PMLR.

[64] Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., et al. （2020）. Learning mesh-based simulation with graph networks. The Nineth International Conference on Learning Representations, 2021.

[65] Wu, T., Wang, Q., Zhang, Y., et al. （2022, August）. Learning large-scale subsurface simulations with a hybrid graph network simulator. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining （pp. 4184-4194）.

[66] Li, Zongyi, et al. "Geometry-informed neural operator for large-scale 3d pdes." Advances in Neural Information Processing Systems 36 （2024）.

[67] Li, Zongyi, et al. "Neural operator: Graph kernel network for partial differential equations." arXiv preprint arXiv:2003.03485 （2020）.

[68] Hao, Z., Wang, Z., Su, H., Ying, C., Dong, Y., Liu, S., ... & Zhu, J. （2023, July）. GNOT: A general neural operator transformer for operator learning. In International Conference on Machine Learning （pp. 12556-12569）. PMLR.

[69] Xie, Tian, and Jeffrey C. Grossman. "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties." Physical review letters 120.14 （2018）: 145301.

[70] Kipf, T. N., & Welling, M. （2016）. Semi-supervised classification with graph convolutional networks. The Fifth International Conference on Learning Representations, 2017.

[71] Choudhary, Kamal, and Brian DeCost. "Atomistic line graph neural network for improved materials property predictions." npj Computational Materials 7.1 （2021）: 185.

[72] Chen, Chi, and Shyue Ping Ong. "A universal graph deep learning interatomic potential for the periodic table." Nature Computational Science 2.11 （2022）: 718-728.

[73] Li, He, et al. "Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation." Nature Computational Science 2.6 （2022）: 367-377.

[74] Dai, Minyi, et al. "Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials." npj Computational Materials 7.1 （2021）: 103.

[75] Swanson, Kirk, et al. "Deep learning for automated classification and characterization of amorphous materials." Soft matter 16.2 （2020）: 435-446.

[76] Qu, H., & Gouskos, L. （2020）. Jet tagging via particle clouds. Physical Review D, 101（5）, 056019.

[77] Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. （2019）. Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics （tog）, 38（5）, 1-12.

[78] Gong, S., Meng, Q., Zhang, J., Qu, H., Li, C., Qian, S., ... & Liu, T. Y. （2022）. An efficient Lorentz equivariant graph neural network for jet tagging. Journal of High Energy Physics, 2022（7）, 1-22.

[79] Villar, S., Hogg, D. W., Storey-Fisher, K., Yao, W., & Blum-Smith, B. （2021）. Scalars are universal: Equivariant machine learning, structured like classical physics. Advances in Neural Information Processing Systems, 34, 28848-28863.

[80] Wu, Q., Zhao, W., Li, Z., Wipf, D. P., & Yan, J. （2022）. Nodeformer: A scalable graph structure learning transformer for node classification. Advances in Neural Information Processing Systems, 35, 27387-27401.

[81] Wu, Q., Zhao, W., Yang, C., Zhang, H., Nie, F., Jiang, H., ... & Yan, J. （2024）. Simplifying and empowering transformers for large-graph representations. Advances in Neural Information Processing Systems, 36.

[82] Miao, S., Lu, Z., Liu, M., Duarte, J. & Li, P.. （2024）. Locality-Sensitive Hashing-Based Efficient Point Transformer with Applications in High-Energy Physics. Proceedings of the 41st International Conference on Machine Learning, in Proceedings of Machine Learning Research 235:35546-35569

[83] Satorras, V. G., Hoogeboom, E., & Welling, M. （2021, July）. E （n） equivariant graph neural networks. In International conference on machine learning （pp. 9323-9332）. PMLR.

[84] Bogatskiy, A., Anderson, B., Offermann, J., Roussi, M., Miller, D., & Kondor, R. （2020, November）. Lorentz group equivariant neural network for particle physics. In International Conference on Machine Learning （pp. 992-1002）. PMLR.

[85] Ravanbakhsh, S., Schneider, J., & Poczos, B. （2017, July）. Equivariance through parameter-sharing. In International conference on machine learning （pp. 2892-2901）. PMLR.

[86] Weiler, M., Forré, P., Verlinde, E., & Welling, M. （2021）. Coordinate Independent Convolutional Networks--Isometry and Gauge Equivariant Convolutions on Riemannian Manifolds. arXiv preprint arXiv

[87] Kanwar, G., Albergo, M. S., Boyda, D., Cranmer, K., Hackett, D. C., Racaniere, S., ... & Shanahan, P. E. （2020）. Equivariant flow-based sampling for lattice gauge theory. Physical Review Letters, 125（12）, 121601.

[88] Boyda, D., Kanwar, G., Racanière, S., Rezende, D. J., Albergo, M. S., Cranmer, K., ... & Shanahan, P. E. （2021）. Sampling using SU （N） gauge equivariant flows. Physical Review D, 103（7）, 074504.

[89] Kaifeng Bi et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature 619, 533–538 （2023）.

[90] Hanzhu Chen et al. “SAC-KG: Exploiting Large Language Models as Skilled Automatic Constructors for Domain Knowledge Graph” ACL, 2024.

[91] Boci Peng et al. Graph Retrieval-Augmented Generation: A Survey[J]. ArXiv, 2024.

[92] D. Zheng, C. Ma, M. Wang, J. Zhou, Q. Su, X. Song, Q. Gan, Z. Zhang, and G. Karypis. Distdgl: Distributed graph neural network training for billion-scale graphs. CoRR, abs/2010.05337, 2020.

[93] B. Yang, S. W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of the International Conference on Learning Representations （ICLR） 2015, 2015.

[94] Chen J, Dong H, Hastings J, et al. Knowledge graphs for the life sciences: Recent developments, challenges and opportunities[J]. arXiv preprint arXiv:2309.17255, 2023.

[95] Walsh B, Mohamed S K, Nováček V. Biokg: A knowledge graph for relational learning on biological data[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020: 3173-3180.

[96] UniProt Consortium. UniProt: a worldwide hub of protein knowledge[J]. Nucleic acids research, 2019, 47（D1）: D506-D515.

[97] Gaulton A, Bellis L J, Bento A P, et al. ChEMBL: a large-scale bioactivity database for drug discovery[J]. Nucleic acids research, 2012, 40（D1）: D1100-D1107.

[98] Zhang N, Bi Z, Liang X, et al. Ontoprotein: Protein pretraining with gene ontology embedding[J]. arXiv preprint arXiv:2201.11147, 2022.

[99] Zhou H Y, Fu Y, Zhang Z, et al. Protein representation learning via knowledge enhanced primary structure reasoning[C]//The Eleventh International Conference on Learning Representations. 2023.

[100] Lam H T, Sbodio M L, Galindo M M, et al. Otter-Knowledge: benchmarks of multimodal knowledge graph representation learning from different sources for drug discovery[J]. arXiv preprint arXiv:2306.12802, 2023.

[101] Huang K, Fu T, Gao W, et al. Artificial intelligence foundation for therapeutic science[J]. Nature chemical biology, 2022, 18（10）: 1033-1036.

[102] Huang K, Chandak P, Wang Q, et al. A foundation model for clinician-centered drug repurposing[J]. medRxiv, 2024.

[103] Wang Z, Zhang Q, Ding K, et al. Instructprotein: Aligning human and protein language via knowledge instruction[J]. arXiv preprint arXiv:2310.03269, 2023.

[104] You R, Yao S, Mamitsuka H, et al. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction[J]. Bioinformatics, 2021, 37（Supplement_1）: i262-i271.

[105] Kulmanov M, Hoehndorf R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms[J]. Bioinformatics, 2022, 38（Supplement_1）: i238-i245.

[106] Xiong B, Cochez M, Nayyeri M, et al. Hyperbolic embedding inference for structured multi-label prediction[J]. Advances in Neural Information Processing Systems, 2022, 35: 33016-33028.

[107] Alsentzer E, Li M M, Kobren S N, et al. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases[J]. medRxiv, 2022: 2022.12. 07.22283238.

[108] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine[J]. Scientific Data, 2023, 10（1）: 67.

[109] Carvalho R M S, Oliveira D, Pesquita C. Knowledge Graph Embeddings for ICU readmission prediction[J]. BMC Medical Informatics and Decision Making, 2023, 23（1）: 12.

[110] Bresso E, Monnin P, Bousquet C, et al. Investigating ADR mechanisms with explainable AI: a feasibility study with knowledge graph mining[J]. BMC medical informatics and decision making, 2021, 21（1）: 171.

[111] Sousa R T, Silva S, Pesquita C. Explainable representations for relation prediction in knowledge graphs[J]. arXiv preprint arXiv:2306.12687, 2023.

[112] Zhang B, Meroño Peñuela A, Simperl E. Towards explainable automatic knowledge graph construction with human-in-the-loop[M]//HHAI 2023: Augmenting Human Intellect. IOS Press, 2023: 274-289.

[113] Zhuang X, Zhang Q, Ding K, et al. Learning invariant molecular representation in latent discrete space[J]. Advances in Neural Information Processing Systems, 2023, 36: 78435-78452.

[114] Shao X, Yang H, Zhuang X, et al. scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network[J]. Nucleic acids research, 2021, 49（21）: e122-e122.

[115] Fang Y, Zhang Q, Zhang N, et al. Knowledge graph-enhanced molecular contrastive learning with functional prompt[J]. Nature Machine Intelligence, 2023, 5（5）: 542-553.

[116] Jiang P, Xiao C, Cross A, et al. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs[J]. arXiv preprint arXiv:2305.12788, 2023.

[117] Li D, Yang S, Tan Z, et al. DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature[J]. arXiv preprint arXiv:2405.04819, 2024.

[118] Fang Y, Zhang Q, Yang H, et al. Molecular contrastive learning with chemical element knowledge graph[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(4): 3968-3976.

[119] David M.Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[120] Feng Nie, Zhixiu Ye, Sifa Xie, Shuang Wu, Xin Yuan, Liang Yao, Jiazhen Peng, and Xu Cheng. "TIEG-Youpu’s Solution for NeurIPS 2022 WikiKG90Mv2-LSC."

[121] JieTang, Jimeng Sun, Chi Wang, and Zi Yang. "Social influence analysis in large-scale networks." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 807-816. 2009.

[122] JieChen, Tengfei Ma, and Cao Xiao. "FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling." In International Conference on Learning Representations. 2018.

[123] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. "Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks." In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 257-266. 2019.

[124] WillHamilton, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs." Advances in neural information processing systems 30 (2017).

[125] Bilinear — PyTorch 2.5 documentation

[126] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. "A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network." In Proceedings of NAACL-HLT, pp. 327-333. 2018.

[127] LiangYao, Jiazhen Peng, Shenggong Ji, Qiang Liu, Hongyun Cai, Feng He, and Xu Cheng. "Friend Ranking in Online Games via Pre-training Edge Transformers." In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2016-2020. 2023.

[128] https://ogb.stanford.edu/docs/leader_linkprop/#ogbl-collab

[129] Jin, Y., Yan X., Liu, S. and Wang, X. A unified framework for combinatorial optimization based on graph neural networks. arXiv:2406.13125, June 2024.

[130] Yue, P., Jin, Y., Dai, X., Feng, Z., and Cui, D. Reinforcement learning for scalable train timetable rescheduling with graph representation. IEEE Transactions on Intelligent Transportation Systems, 25(7): 6472 – 6485, 2024.

[131] Liu, S., Yan, X. Yan, and Jin, Y. End-to-end Pareto set prediction with graph neural networks for multi-objective facility location. Evolutionary Multi-Criterion Optimization (EMO2023).

[132] Dai E, Zhao T, Zhu H, et al. A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability[J]. Machine Intelligence Research, 2024: 1-51.

[133] Dong Y, Ma J, Wang S, et al. Fairness in graph mining: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(10): 10583-10602.

[134] Li H, Wang X, Zhang Z, et al. Out-of-distribution generalization on graphs: A survey[J]. arXiv preprint arXiv:2202.07987, 2022.

[135] Li Q, Li J, Sheng J, et al. A survey on deep learning event extraction: Approaches and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[136] T. H. Nguyen, K. Cho, and R. Grishman, “Joint event extraction via recurrent neural networks,” in Proc. NAACL HLT, 2016, pp. 300–309.

[137] X. Liu, Z. Luo, and H. Huang, “Jointly multiple events extraction via attention-based graph information aggregation,” in Proc. EMNLP, 2018, pp. 1–10.

[138] T. Zhang, H. Ji, and A. Sil, “Joint entity and event extraction with generative adversarial imitation learning,” Data Intell., vol. 1, no. 2, pp. 99–120, May 2019.

[139] Zhao X, Deng Y, Yang M, et al. A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers[J]. ACM Computing Surveys, 2024, 56（11）: 1-39.

[140] Zhang, Wen, Chi-Man Wong, Ganqiang Ye, Bo Wen, Wei Zhang, and Huajun Chen. "Billion-scale pre-trained e-commerce product knowledge graph model." In 2021 IEEE 37th International Conference on Data Engineering（ICDE）, pp. 2476-2487. IEEE, 2021.

[141] David M.Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022.

[142] Feng Nie, Zhixiu Ye, Sifa Xie, Shuang Wu, Xin Yuan, Liang Yao, Jiazhen Peng, and Xu Cheng. "TIEG-Youpu's Solution for NeurIPS 2022 WikiKG90Mv2-LSC."

[143] JieTang, Jimeng Sun, Chi Wang, and Zi Yang. "Social influence analysis in large-scale networks." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 807-816. 2009.

[144] JieChen, Tengfei Ma, and Cao Xiao. "FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling." In International Conference on Learning Representations. 2018.

[145] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. "Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks." In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 257-266. 2019.

[146] WillHamilton, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs." Advances in neural information processing systems 30 (2017).

[147] Bilinear — PyTorch 2.5 documentation

[148] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. "A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network." In Proceedings of NAACL-HLT, pp. 327-333. 2018.

[149] LiangYao, Jiazhen Peng, Shenggong Ji, Qiang Liu, Hongyun Cai, Feng He, and Xu Cheng. "Friend Ranking in Online Games via Pre-training Edge Transformers." In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2016-2020. 2023.

[150] https://ogb.stanford.edu/docs/leader_linkprop/#ogbl-collab

[151] Liang, Lei, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu et al. "KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation." arXiv preprint arXiv:2409.13731 (2024).

[152] Jie Tang, Jimeng Sun, Chi Wang. and Zi Yang. Social Influence Analysis in Large-scale Networks. In KDD, pages 897-816, 2009.

免责声明

本白皮书所含数据和观点仅反映编制组在发布本白皮书时的判断，编制组已尽最大努力确保信息的准确性、完整性和可靠性，但不作任何形式的保证。在任何情况下，本白皮书中的信息或表述均不构成投资建议，且编制组对本白皮书中的数据和观点不承担法律责任。此外，编制组保留在未另行通知的情况下对本白皮书所载信息进行修改的权利，读者应自行关注相关变动。

全国智能计算标准化工作组（SAC/SWG32）
浙江省杭州市余杭区中泰街道科创大道之江实验室
0571-58005197
IComputingSWG@zhejianglab.org