大规模图计算与智能系统
——第十四期CCF秀湖会议报告

整理：郑卫国 彭 鹏 刘 钰 邹 磊

背景与意义

“图”作为支撑知识图谱、大数据融合、网络通信、金融风控等应用的基础数据模型，在越来越多的应用中被用来表达实体之间的复杂关联关系。工信部《“十四五”软件和信息技术服务业发展规划》明确强调要突破“并行图数据处理关键技术”；高德纳（Gartner）公司最新研究报告指出，到2025年，80%的数据分析任务都会用到图技术。随着图数据的规模越来越大、应用越来越广，图计算也面临诸多挑战。（1）数据规模增大：图数据规模的迅速增长，使现有的图计算系统在处理速度和存储能力上面临重大挑战；（2）复杂关系挖掘难：图数据中复杂的关联关系使挖掘和分析变得困难，亟须开发更高效的算法和工具以提取有价值的信息；（3）传统算法与图学习的结合：虽然传统图算法与图机器学习方向已有一些研究成果，但二者的有效结合及其在多学科、多领域的应用仍需深入优化；（4）基础模型构建的核心问题：图基础模型与文本类的语言基础模型存在根本区别，如何定义并构建适应不同应用场景的图基础模型是下一代图智能亟待解决的核心科学问题。因此，需要针对图数据的数据管理系统、图机器学习算法、图计算系统等进行深入研讨，为我国在下一代大数据与人工智能方面的研究布局提供参考思路。

第十四期CCF秀湖会议组织了来自数据库、系统结构、机器学习等领域的学者，以及在化学和脑科学等自然科学领域应用图技术进行科学发现的专家，以图数据为核心，共同探讨了“大规模图计算与智能系统”的研究进展与面临的挑战。与会专家达成的共识是：图作为一种独特的数据模态，具有极强的对复杂关联关系建模的能力，能将相关信息聚集到一起，在构建图计算系统时，结合在线、近线、离线三种业务场景至关重要；开发面向大规模图数据的混合事务/分析处理（Hybrid Transactional/Analytical Processing, HTAP）系统是亟待解决的挑战；传统的图算法与图学习已有大量的研究成果，但二者的结合以及在多学科、多领域的应用仍然有大量的工作需要优化和改进；图基础模型与文本类的语言基础模型存在根本区别，如何定义并构建图基础模型是下一代图智能亟待解决的核心科学问题。

专题报告

加拿大皇家科学院院士、滑铁卢大学教授M. 塔梅尔·欧苏（M. Tamer Özsu）在报告中对图计算研究给出了系统总览。他指出，虽然图数据无处不在，但是图计算研究分散为若干子领域，彼此联系不够密切，因此希望能从整体视角给出一个概览。他从图数据的类型、图动态性、图算法类型、图系统的负载类型等几个视角对图计算研究进行了概述，并提出了一些开放性问题，包括图处理的硬件支持、图的HTAP系统、兼容资源描述框架（Resource Description Framework, RDF）和属性图的通用图数

据库系统等。

俄罗斯工程院外籍院士、合肥工业大学教授吴信东在报告中围绕跨姓家谱系统——华谱系统，介绍了其主要功能和背后的知识图谱技术。目前存在四类计算范式：第一性原理的数学模型、逻辑符号推理、深度学习、生成式预训练大模型。四类计算范式各有千秋，优势互补是未来的发展方向。

香港中文大学教授于旭围绕几个前沿研究问题展开了讨论，包括使用关系数据库支持图计算负载、图结构学习、图与大语言模型的结合等。他最后提出了一些开放性问题，例如设计结合关系型数据库和图数据库的一体化系统、图分析任务与图查询任务的整合等。

欧洲人文和自然科学院外籍院士、上海交通大学教授林学民从图的发展历程和图论问题开始，指出图是跨不同领域的通用模型，并从内聚子图、网络弹性、子图枚举、二部图等多个领域系统化地介绍了图算法研究的进展，涵盖单CPU、FPGA、分布式等技术。

大规模图系统的最新研究

清华大学教授陈文光从属性图数据库与查询语言出发，围绕几个查询示例展开讨论，揭示了查询语言与API编程的巨大差距，发现了图更新带来的在线近线一致性问题。查询语言的自动优化有巨大空间，在线近线一致性是应用的强需求。

华中科技大学教授郑龙指出了复杂场景的高能效统一加速的重要性，对构建一个灵活、高效的领域通用图计算系统给出了建议，例如领域通用图计算指令集架构及加速器、流水可软件定义的编译优化等。

北京大学教授邹磊以其团队构建的图数据库系统gStore为例，从图查询建模、图存储、图数据库查询执行与优化等方面展开讨论，提出了几个面向图数据库系统的开放性问题，例如如何设计囊括子图匹配、正则路径、其他复杂查询结构的统一查询计划表示形式，以及基于它的高效查询计划算法等。

武汉大学教授洪亮提出智能风控是数字金融的核心，图表达了金融大数据中的内在关联，是对金融风险建模的关键技术。图系统在金融风控方面发展的未来趋势包括增强特征提取能力、提升泛化学习能力等。

华为公司曾立博士指出图与大语言模型（Large Language Model, LLM）的协同是产业界和学术界当前探索的热点，是未来发展的方向。

蚂蚁集团洪春涛博士从支付风控、花呗反套现、团伙挖掘等方面展开讨论，他指出：（1）图作为数仓是可行且有优势的；（2）图在中短期有很多应用空间可供挖掘；（3）长期来看大流行需要大应用。

图学习与智能技术最新进展

北京大学助理教授贺笛介绍了基于图双连通性、子结构技术和图同态的图神经网络模型设计方法及其表达能力，提出未来的图神经网络模型设计应从模型结构、高效算法、新的理论三个维度进行研究。

东北大学教授张岩峰围绕图神经网络训练的执行模式和图神经网络的训练加速技术两个方面展开讨论，从数据驱动的训练模式加速、CPU-GPU异构训练加速、分布式训练加速三个方面总结了多种加速方法。

北京大学助理教授张牧涵从ChatGPT这类自然语言大模型能否出现在图领域出发，探讨了一种基于广泛数据训练的图基础模型，它能以单一模型适应广泛的下游任务。图基础模型面临的挑战包括特征/标签空间不对齐、任务类型多样、图的上下文学习等。OFA（One for All）模型使单个图神经网络（GNN）模型能够处理不同的分类任务和不同的图数据集，可以解决所有挑战。

中山大学副教授王桢从“猜你喜欢”场景的特征生成与选取、社交电商中的好友推荐切入，展示了通过优化邻接关系对表格型数据的特征交互及异构图连边预测进行建模，用于新场景的快速冷启动。

大模型时代下图技术的机遇与发展

浙江大学教授陈华钧指出，语言理解和知识处

理是实现认知智能的两个核心任务。通过增加模型规模和提升语料表示可以提升模型能力。

北京邮电大学教授石川指出，基础模型已经应用在语言、视觉和语音等领域，大语言模型具备理解、生成、逻辑、记忆等人工智能的核心基础能力。然而大模型无法有效解决图任务，图模型也缺乏大模型的知识涌现和任务泛化的能力，因此未来可以设计和实现图基础模型。

上海交通大学教授严骏驰团队的吴齐天博士从封闭世界假设下的学习和开放世界假设下的学习两个角度，讨论了不同条件下扩散过程与GNN、Graph Transformer模型的联系，提出了如何扩展到大数据、如何量化不确定性等是需要解决的挑战。

香港科技大学（广州）助理教授李佳由LLM解决图算法任务的局限性引入GraphWiz（GraphWiz构建了第一个图算法推理训练数据集），并提出了两点挑战：LLM允许的最大输入长度限制了目前可处理的最大图尺寸；GraphWiz和GPT-4目前都难以解决困难的图算法任务，比如NP-完全问题。

北京海致星图科技有限公司沈游人博士从知识库层面阐述了知识图谱与大模型融合的可行性，指出复杂知识库的构建、图上检索能力的增强、全面的评价标准和正确性保证等是需要解决的挑战。

跨学科的图智能技术交流与合作

北京大学教授唐淳展示了化学中的图连接形式，它们跨越种属却结构相似，在多重序列对比之后基于进化信息的蛋白质结构进行建模，并提出蛋白质结构预测的大数据方法。

东南大学教授郑文明展示了情感特征的结构化表示，在情感特征的图表征模式中，图的邻接矩阵与情感活动模式密切相关。他提出了一个建设性问题：在基于脑电和功能性近红外光谱技术（fNIRS）的双模态情感识别中，如何将基于多通道脑电信号构建的图和基于fNIRS信号构建的图进行合并，建立异构模态的联合图表征模式。

浙江大学研究员张强提出，图在AI4Science中大有可为，图是科学数据的有效表示，图（谱）是专家知识的结构化总结，图（谱）加大模型能进行功能预测、从头设计，充当科研助手，引导科学发现。

观点集萃

图计算系统

图计算的核心技术优势在于其对现实世界中复杂关联关系有更直观的建模能力。不同于关系数据库将数据分散存储于多个表中的方式，图模型通过节点和边的形式自然地将相关信息聚合在一起，更贴近现实世界的结构和关系（洪春涛）。为了有效管理和操作图模型，我们需要识别并归纳出通用的操作算子，进而构建一个统一的系统进行综合管理（于旭）。

在设计针对图计算系统的HTAP方案时，需要采取一体化设计方法。在图事务处理层面维护适合更新的图数据结构，例如邻接列表；在分析层面维护内存中的压缩稀疏行（Compressed Sparse Row, CSR）结构，考虑通过运用分布式一致性协议实现事务与分析的同步机制（邹磊）。此外，为了充分发挥系统的潜力，图计算系统应该根据具体的业务场景，实现在线、近线和离线三种处理模式的紧密协作（陈文光、洪春涛）。

为了进一步提升图计算系统的性能，可以从以下几个方面进行考虑。（1）考虑到GPU等新型硬件的引入，GPU强大的并发计算能力是突破大规模图计算性能瓶颈的重要手段之一。然而由于图数据本身的稀疏性，GPU并非图计算的天然适配者，尤其是GPU等硬件设备上大规模图的访存优化是需要解决的重要挑战之一。同时，设计面向图计算的全新原生加速器，为进一步提高图计算的性能提供新的契机（郑龙）。（2）针对GNN的应用，我们可以结合GNN的典型层数特点，探索更高效的并行化策略。通过优化并行计算模式，能够充分利用现代多核处理器的计算能力，加速GNN的训练和推理过程（张岩峰）。（3）针对图数据库的查询优化问题，建立一个全面的优化机制至关重要，包括从存储、索引

到查询优化引擎的各个环节。特别是对于复杂图查询语言，需要设计一种统一的图查询计划优化表达范式，以及研究一种针对统一表达范式的优化机理，以确保图查询的高效执行（邹磊）。

当前，关系数据库工具已经相当成熟，但在图系统领域，工具链的发展仍有待加强。图处理技术尚未深入到基本操作的层面，因此，需要对图中的基本操作进行深入的提炼和总结（吴信东）。展望未来，我们可以考虑汇聚众人的智慧，共同开发一套图计算软件栈。然而，要实现这一目标，需要在图应用中找到共性的、基础的关注点，即图查询、图分析和图学习的交集部分。只有这样，才能构建一个真正强大且实用的图计算软件栈，推动图处理技术的进一步发展（程学旗）。我们还需要构建一个对用户更加友好的图计算系统，包括与现有的大数据平台无缝整合，以及将其部署在云计算平台上等，以满足用户多样化的需求并提供便捷高效的服务（彭鹏）。图计算虽然没有像大模型研究在学术界和工业界那样的影响力，但是近年来正以一种“润物细无声”的方式，慢慢融入到各个领域的数据管理分析方面，包括金融风控、电信网络分析，以及自然科学领域等（邹磊）。

图机器学习

图神经网络的表达能力是图机器学习中的基础问题之一，十分值得研究，传统机器学习理论不完全适用，而基于威斯费勒－莱曼（Weisfeiler-Lehman, WL）测试的表达能力体系无法面向多种图学习场景实现普适实用性。基于图双连通性、子结构计数和图同态等新的判定能力的图神经网络在特定任务上能力更强（贺笛）。图机器学习处理的数据包括显式图数据和隐式（广义）图数据，从GNN扩展到Graph Transformer是处理广义图数据的有效途径，图上的分布外泛化等也是重要问题之一（吴齐天）。

图学习领域同样受到了大语言模型的影响。在当前阶段，将大模型与GNN相结合是实现具有普适性、高精度图学习模型的最佳方案，其中基于文本属性图实现多任务、多数据语义对齐是目前的主要手段。然而，大模型与GNN的融合仍处于早期阶段，有很大的优化空间；而图提示学习方法尚缺乏方法论层面的指导，导致其实际效果波动较大。在应用方面，知识图谱与大模型的交叉结合是最常见、发展最快的方向，利用知识图谱实现检索增强生成取得了较好效果（张牧涵）。

图学习的基准（benchmark）构建面临一系列挑战，虽然存在OGB（Open Graph Benchmark）等常用的基准数据集，但仍存在数据集数量少、图数据规模小等问题；现有基准数据集在数据划分等方面不尽合理，导致基准对图学习效果的评估意义受限（张牧涵）。与图数据库系统相比，图学习需要领域专家定义预测任务，因此数据规模要足够大，质量要足够高。高质量的大规模图学习数据集是图机器学习领域未来发展的关键（石川）。

图基础模型是目前研究的热点，有很多问题尚待解决。例如，GNN和LLM相结合得到的图学习模型算不算“大”模型，仍然有待商榷，因为GNN本身属于小参数量模型。基于LLM的图基础模型面向图算法问题的能力仍然十分受限，在计算的准确性和支持的图规模等层面需要进行大幅度优化。从远景目标来看，图基座模型应当有能力从图中提取出通用架构解决多模态信息的问题，实现跨域结构知识迁移。然而，图数据结构知识迁移面临很大的技术挑战，目前还有广阔的研究空间。退而求其次，做好某个应用领域内的图基座模型具有更强的可行性，但图结构知识是否具有共性、不同领域的结构知识迁移如何实现还有待讨论（石川）。以化学大模型为例，如果模型结构不合适，仅靠增加参数和数据量无法起作用，需要在特定场景下选择合适的模型（贺笛）。

图算法

在图数据建模领域，现实世界中的大多数应用场景都是动态的，而现有的图算法研究大部分集中在静态图上。因此，未来的研究应当更多地关注动态图或流图领域。值得注意的是，动态图并不等同于流图。动态图的本质是操作流，它关注的是图

结构随时间的变化；而流图是数据流，它侧重于数据在图中的流动和处理。这两种图各有特点，但都对理解和分析现实世界中的复杂系统具有重要意义（M. Tamer Özsu）。

图不仅仅是一种数据结构，也可以是一种独特的数据模态。在许多业务分析中，图作为一种基础的、共通的表征形式，扮演着至关重要的角色。如果没有图这一强大的工具，许多复杂的业务分析任务将难以顺利进行。因此，深入研究和应用图数据模态，对于推动业务发展和创新具有不可估量的价值（曾立）。

在图数据查询与优化方面，与关系模型所依赖的明确代数相比，图模型尚缺乏一套通用的算子体系。因此，在设计或整合图算子时，必须确保这些算子相对于关系代数能够带来足够的性能增益，这样才能吸引用户的注意并促使他们从关系模型迁移到图模型。为了实现这一目标，需要精心制定查询规划，以便进行有针对性的、增量式的优化。只有这样，才能充分发挥图模型在数据查询与分析方面的潜力，为用户提供更加高效、灵活的数据处理解决方案（于旭、邹磊）。图上的很多任务十分困难，计算复杂度较高，新设计的算法运行效果比较好，但是复杂度与之前已有工作相比没有改善，可以考虑特定图上证明复杂度的界（魏哲巍）。面对图数据结构复杂、规模庞大、动态变化、属性丰富等特性，设计轻量级、并行化、语义感知的图算法十分关键，也是未来图算法研究的重要机遇（郑卫国）。

在图学习与图算法结合方面，目前，尽管有人尝试利用AI技术优化图算法，但该领域仍然有大量的工作需要优化和改进。例如，基于强化学习求解图算法时，推理成本往往较高，而且在拟合最优决策方面也面临着瓶颈（李友焕）。另一方面，“Transformer as Algorithm”的概念提出了一种可能性，即理论上可以用Transformer拟合任何算法并生成相应的解。然而，一旦涉及到神经网络，计算代价往往会变得更高。因此，在实际应用中，如何有效地结合图学习和图算法是一个需要仔细考虑的问题（贺笛）。此外，要尽可能地增强图神经网络的表达能力，使其能够拟合传统的图算法，同时避免预定义特征的需要。这样或许能够在保持算法灵活性的同时降低计算成本，从而为图数据的处理和分析开辟新的可能性（张牧涵）。

跨领域图计算

在情感识别的研究中，基于情感特征的图表征模式与情感活动模式具有良好的匹配关系，在对情感信息的感知中，可以使用图算法实现跨领域、多功能的合作；在人机交互领域，图算法可以用于开发更加智能的聊天机器人，使其能够理解和适应用户的情感状态，提供更加人性化的交互体验；在心理健康领域，图算法可以用于分析患者的言语和行为模式，帮助医生诊断和治疗情绪障碍。但人类情感的复杂性、异构模态图之间的信息传递与交互、特征的提取与融合等问题都是跨领域合作要面临的挑战（郑文明）。

化学结构中存在着多种图连接形式，图是科学数据的有效表示。在化学领域，构建化学元素知识图谱能预测分子属性，图算法可以预测化学反应的结果及形成原因。此外，图算法能够根据所需的分子结构倒推原料和反应步骤。图算法的应用不仅可以提高化学反应预测的准确性，还可以加速新材料的发现和合成路径的设计，对于推动化学科学的发展具有重要意义。在实际应用中，数据的质量和可用性、模型的泛化能力都是亟待解决的问题（唐淳、张强）。

跨领域学术合作在合作方式和预期成效方面面临一系列挑战。现在的自然科学发现越来越离不开计算机技术的赋能，然而由于自然科学与计算机科学对学术研究的定位和评价体系不同，如何真正实现高效的跨学科合作，是一项值得探索的问题（石川、邹磊）。

小结

图计算系统、图机器学习和图算法三者之间存在着紧密的联系，如图1所示。

图计算系统构成了支撑图模型高效管理与操

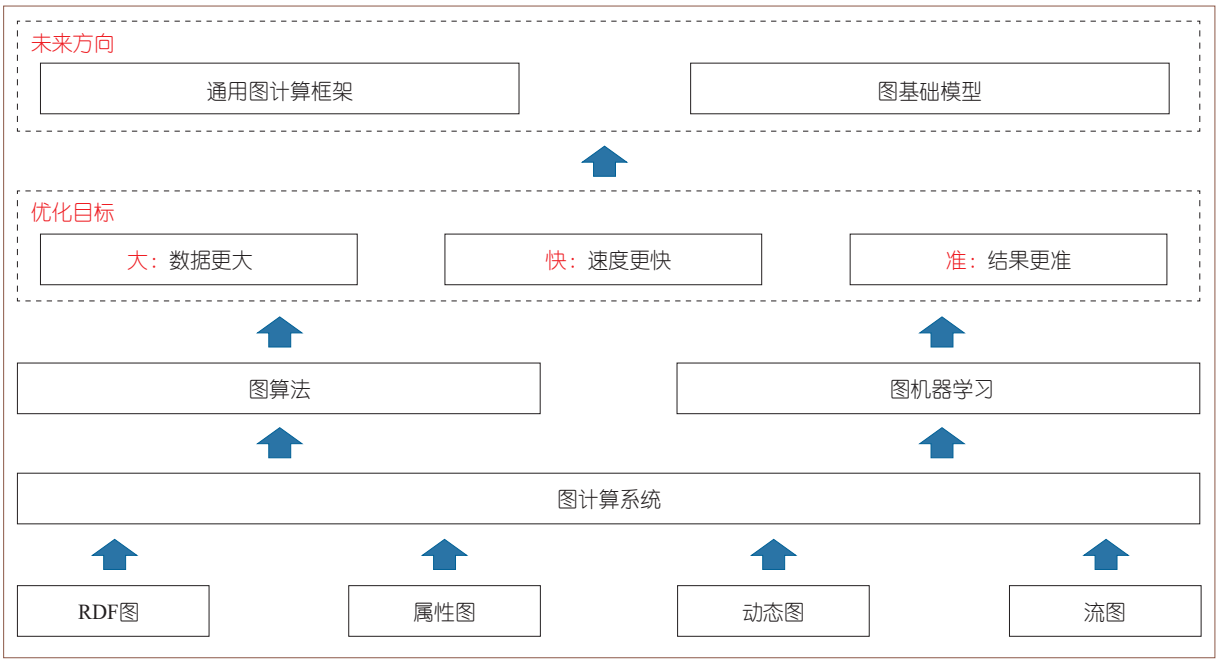

图1　大规模图计算与智能系统

作的基石，不仅为图数据的查询、分析和学习任务提供了强有力的平台，还确保了这些任务的流畅执行。图算法则是实现这些任务的关键工具，针对图数据特有的查询与分析需求提供了精确的方法和策略。图机器学习作为一个专门针对图数据的学习框架，充分利用图的结构信息挖掘知识、执行预测任务，其效能显著依赖于图计算系统所提供的强大计算资源和数据管理能力。特别是随着大模型技术的不断进步，图基础模型日益成为图机器学习领域研究的焦点，预示着未来这一方向上将涌现出更多的创新和突破。图计算系统、图机器学习和图算法三者的结合为理解和利用图数据中的复杂关系和模式提供了强大的工具和平台，不仅支持我们处理更大的图数据，还极大地提升了图处理的速率，并确保了图处理结果的精准性。

展望未来，我们面临几个关键的开放性问题：

1. 如何深入挖掘和精炼图中的基本操作，并在此基础上共同构建一个通用的图计算框架？

2. 图基础模型是否真实存在，且可被开发和训练？如果答案是肯定的，那么我们应该如何着手实现？

3. 除金融、电信网络等经典的图计算应用领域之外，图计算研究在自然科学领域的切入点在哪里？如何高效地开展和评估跨领域的研究合作？

这些问题不仅是对当前图计算研究边界的挑战，也是推动图技术在各个应用领域进一步发展的关键。我们期待图计算系统、图机器学习和图算法领域的专家学者能够一起构建相互合作、交叉融通的研究环境，并通力打造一系列有利于图研究领域发展的社区行动，例如图机器学习基准、图计算挑战赛和专门面向图的会议 / 专刊等。 ■

整理：郑卫国　彭　鹏　刘　钰　邹　磊

会议发起人：陈文光　邹　磊

特邀嘉宾（按姓名拼音排序）：

M. Tamer Özsu　程学旗　林学民　吴信东

参会嘉宾（按姓名拼音排序）：

陈华钧　贺　笛　洪春涛　洪　亮　李　佳　李友焕
刘　钰　彭　鹏　沈游人　石　川　唐　淳　王　桢
魏哲巍　吴齐天　严骏驰　于　旭　张牧涵　张　强
张岩峰　曾　立　郑　龙　郑卫国　郑文明

会议记录：庞　悦　苟向阳　王一君

（本文责任编委：**袁　野**）