大数据与决策研究

（政策与技术跟踪专题）

2023 年第 28 期（总第 184 期）

广西壮族自治区信息中心
广西壮族自治区大数据研究院　　2023 年 9 月 1 日

编者按：图计算是基于图数据的分析技术与关系技术应运而生的计算技术。近年来，全球大数据进入加速发展时期，数据量呈现爆发式增长，图计算技术解决了传统的计算模式下关联查询的效率低、成本高的问题，在问题域中对关系进行了完整的刻画，并且具有丰富、高效和敏捷的数据分析能力。本期将介绍图计算技术相关情况。

本期要目

图计算技术概述

一、图计算技术定义

图计算简单来讲就是研究在海量数据中，如何高效计算、存储并管理图数据等问题的领域。作为一种处理和分析图结构数据的方法，图由节点（vertices）和连接节点的边（edges）组成，节点和边可以携带各种属性和标签。图计算的分析关键是关联行为的分析，旨在利用图结构的关联性和拓扑性质来进行复杂的数据分析和挖掘。

二、图计算技术特点

图计算技术的主要特点是高效地对具有复杂关联关系的数据进行深度计算。

（一）图的代表性结构。图提供了一种能够代表现实世界中绝大多数事物关联关系的独特的结构。与经典的表格或者矩阵不同，图上的节点和边并没有被赋予过多的权重，每个元素都依赖于其他元素并形成一种互联互通的关系，而这种关系是所有基于图的假设和预测的核心。在大数据计算中，通过分析图数据之间的关联性，能够高效地从噪声很多的海量数据中抽取有用的信息。

（二）图计算的高效性。图计算能够高效地对具有复杂关联关系的数据进行深度计算。例如在金融领域的信用卡套现场景中，通过深度分析个体和个体、个体和事件、事件和

事件间的关联性图计算便能够帮助银行从上亿点边规模的交易数据中精准且高效地识别出金融欺诈操作。

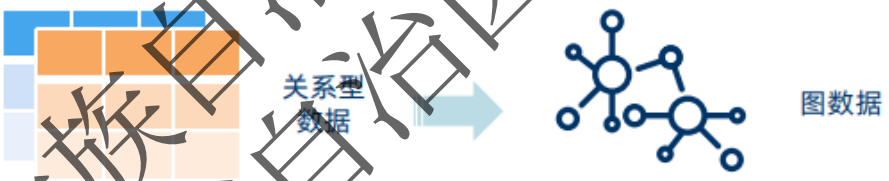
（三）图计算系统的深度性。图计算系统基于顶点和边的方式存储图数据和计算，能够建构任意复杂的网络和模型并存储大量的信息，进而完整且形象地映射分析人员想要研究的问题域。经典的表格结构的数据都能够用图数据来表示，但不是所有的图数据都能够用数组或表格的形式来表示。在对简单事物关系的数据进行计算时，列表型的数据尚且能够展现出高效的性能，然而一旦模型复杂度提升，例如金融领域中的交易数据，传统的列表数据模型的劣势将显现无疑。倘若在传统的关系型数据模式下进行分析和计算，复杂的业务场景将带来冗余的表之间的关联操作和频繁的数据通信，造成成千上万倍计算量的提升，系统性能大打折扣，极大降低了计算的效率。但是，在面对高度结构化的数据时，图计算的处理能力将不及基于传统数据模型的计算，这是由于在进行图计算的过程中存在着随机访问的问题。

三、图计算技术原理

图计算技术原理是基于图数据的分析和计算。图计算指代一切基于图数据的分析和计算。图计算的目标就是从图结构中挖掘出有价值的知识或规律，包括事件溯源、因果关系等。传统上，数据结构以表格形式居多，而图结构由一系列的点、边以及点和边上所具有的属性构成。关系型数据库在存储数据和数据之间的关系时，往往需要创建多张表来表示

数据和数据之间的关系不够直观且需要通过多表关联才能完成查询操作。而图数据可以直接在图中建立数据节点之间的边来表示数据节点之间的关系，数据在被放入数据库前，就已经做好关联。图数据既是简洁的、优雅的、高效的，又具有强大的扩展能力和个体间关系表征能力，因此图计算尤其适用于大数据背景下的复杂关联关系的分析计算。

不论是分子间的结构、还是神经结构、又或者是交通网络和能源网络，一切充满关联的事物都可以用图结构来表示。图计算的核心在于如何将数据建模为图结构，以及如何将解决问题的步骤变换为图结构上的操作和计算问题。当实际问题涉及到关联分析时，图计算往往能够使得问题的解法很自然地表示为一系列对图结构操作和计算的过程。



数据库	图数据库	关系型数据库
理论基础	图论	关系模型
数据存储方式	数据节点	二维关系表
关系存储方式	边	主键外键方式
关系查询方式	图查询、简洁	表连接
关联查询速度	快	慢

图 1　图数据库和关系型数据库对比

（来源：《2022 中国图计算技术及应用发展研究报告》）

图计算技术模型

为提高图计算系统分析计算图数据的能力，图计算系统需要针对图数据和图计算的特点，设计并实现支持频繁迭代操作、细粒度并行计算和通信开销小的全新计算模型，即图计算模型。按照计算对象，图数据计算模型可以分为节点中心计算模型、边中心计算模型、路径中心计算模型和子图中心计算模型四类。

一、节点中心计算模型

（一）同步节点计算模型。同步节点计算模型基于图计算局部性差、节点计算量小、并行性差异大的特点提出，将图算法的每一次迭代转换为图中每一个节点执行一次超步（superstep）运算。

（二）异步节点中心计算模型。异步计算模型与同步计算的迭代设计相同，仍为 BSP 三步操作模型，但是在接收上一轮超步计算的消息时，不再是由邻居节点推送更新的数据，而是由计算节点采用“拉”的方式选择性地读取邻居节点的消息。

（三）节点中心计算模型。GAS 计算模型沿用同步节点中心计算模型中超步的概念，并且通过划分大度数节点在单个计算节点内实现并行计算。

二、边中心计算模型

边中心计算模型将图数据以边列表为核心数据结构并维护源节点列表，每次迭代的计算操作更新目的节点列表（Uout），其记录在边列表上的计算操作产生的对每条边的目的节点进行更新的消息序列。边中心计算模型将图算法的迭代计算转换为可在边列表上顺序执行，避免了随机读写数据对内存资源的高要求，从而解决了节点中心计算模型面临的资源受限和通信开销过大的难题。边中心计算模型的流式顺序计算特点使得在全局图数据上的计算可以分块实现，顺序访问存储在硬盘上的数据，降低了图数据分析计算对内存容量的要求，即可在单机上实现对大规模图数据的分析处理。此外，边中心计算模型将每次迭代计算生成的目的节点更新消息序列进行重排序，获得按源节点合并排序的更新消息流，则更新消息数不大于节点数，大大简化了消息更新同步的开销。

三、路径中心计算模型

路径中心计算模型以路径为计算单元，即从源节点出发到目的节点的边序列。路径中心计算模型将图数据组织为前向边遍历树（forward-edge traversal tree）和后向边遍历树（reverse edge traversal tree），从而将图计算转换为在树上的迭代计算。路径中心计算模型基于前向遍历树和后向遍历树的每次迭代运算分为两步，消息分发和信息收集。

四、子图计算模型

子图中心计算模型完成图划分后，在多个子图上并行执行迭代图计算，一次超步运算执行两步操作：（1）子图并行执行用户定义计算操作，并输出计算结果；（2）包含相同节点的子图间更新节点信息。步骤（2）可在所有子图的步骤（1）操作结束后同步执行，或者在保持数据一致性前提下异步执行。子图中心计算模型通过子图划分方法，将图算法转换为多个子图上的迭代计算，成功减少了计算时的通信开销和迭代操作次数。

图计算模型	任务调度	数据划分	并行性	系统实现	优势	局限
节点中心	同步/异步	节点序列子集	高	分布式/单机	模型实现简单且易于算法迁移；计算并行性高，可采用同步或异步调度，适用于各类算法	计算节点之间通信开销大；完成图计算所需迭代次数多；计算并行性受数据一致性限制
边中心	同步/异步	边序列分块	中	单机	模型实现对设备资源要求低；数据存储、分块和读写访问更加简单；数据访问顺序执行，易于维护数据一致性	计算并行性受边列表分块限制；图算法迁移复杂，且适用范围小
路径中心	同步	子树	中	分布式/单机	数据查找访问简单快捷；基于两步操作，图算法实现简单	构建遍历树的初始化开销大；数据一致性实现复杂；模型实现困难，且数据存储复杂
子图中心	同步/异步	子图	低	分布式/单机	超步运算之间通信开销小；完成图算法的迭代次数少	计算并行性受限；基于子图的数据划分困难；实现用户可自定义的子图划分复杂

图 2　图计算模型对比

（来源：《人工智能之图计算》）

图计算技术典型应用场景

目前，图计算技术已初步应用于医疗、金融、社交分析、自然科学以及交通等领域，众多互联网公司以及很多年轻的人工智能领域创业公司也都开展了图计算相关的业务。

一、图计算技术在医疗行业的应用

图计算的出现使得对病人的智能诊断成为可能。对病人开具处方需要依据病人的病情特征与以往的健康情况，以及药物的相关情况。过去的医疗大多依赖于医生的个人经验与病人的自我描述，传统的数据处理系统无法一次性调出多个与病人情况、保险情况、药物情况相关的数据库——挑战在于信息必须由多个在线资源拼凑而成，包括列出疾病和治疗的电子病历、医疗保险或其他跟踪医疗服务的数据库、描述药物的数据库，在某些情况下，还有跟踪临床试验的独立数据库。该场景是经典的链接网络，每个节点之间具有相互依赖性。变量可包括患者年龄和性别、特定药物（或药物组合）的结果、特定剂量，给药时的疾病阶段和潜在的药物相互作用。传统的 SQL 数据库实际上不可能计算这样的问题，因为传统的纯软件图无法提供应用所需的深度嵌套的连接，而图分析系统的出现则使得这样的场景成为了可能。

二、图计算技术在金融行业的应用

在金融实体模型中，存在着许许多多不同类型的关系，

以及数十亿的结点和边。有些是相对静态的，如企业之间的股权关系、个人客户之间的亲属关系，有些则是不断地在动态变化，如转账关系、贸易关系等等。这些静态或者动态的关系背后，隐藏着很多以前我们不知道的信息。之前，我们在对某个金融业务场景进行数据分析和挖掘过程中，通常都是从个体（如企业、个人、账户等）本身的角度出发，去分析个体与个体之间的差异和不同，很少从个体之间的关联关系角度去分析，因此会忽略很多原本的客观存在，也就更无法准确达到该业务场景的数据分析和挖掘目标。而图计算和基于图的认知分析正是在这方面弥补了传统分析技术的不足，帮助我们从金融的本质角度来看这个问题，从实体和实体之间的经济行为关系出发来分析问题。利用图计算和图认知技术从交易本身出发，探查交易方的交易历史，跟踪交易的轨迹，追溯资金的流向，找出规则方法无法覆盖的新的洗钱模式，及时地更新现有的探查规则，可以大幅度降低误报率。利用图计算和图认知技术，完整刻画企业客户之间、企业与自然人之间的社会关系、经济往来关系，构建全方位的风险关联网络，实现风险要素的动态性和完整性呈现。当关联网络内某家企业发生信贷风险时，利用风险关联网络中风险客户的客户画像，经济行为轨迹等信息进行交叉关联分析，预测风险的传导路径和扩散范围，帮助银行采取有效措施，阻断风险传染源，进行风险隔离，从而提升风险管理的可靠性和准确率。

三、图计算技术在互联网行业的应用

随着数据的多样化，数据量的大幅度提升和算力的突破性进展，超大规模图计算在互联网大数据行业发挥着越来越重要的作用，尤其是以深度学习和图计算结合的大规模图表征为代表的系列算法。

（一）社交网络分析。社交媒体平台如微信、淘宝、腾讯 QQ 等拥有庞大的用户网络，通过图计算技术可以分析用户之间的关系、社交圈子、信息传播路径等，用于推荐系统、个性化推荐、社交网络挖掘等。

（二）网络安全分析。图计算技术可以帮助识别和预测网络攻击、恶意行为、异常活动等，通过分析网络设备、用户行为和网络流量之间的关系，提高互联网系统的安全性。

（三）搜索引擎优化。图计算技术可用于搜索引擎的链接分析，帮助发现网页之间的链接关系、网页排名等，改善搜索结果的质量和相关性。

（四）广告推荐。通过对广告内容、用户兴趣和广告投放位置等进行关联分析，图计算技术能够提供更精准的广告投放策略，提高广告效果和用户满意度。

（五）知识图谱构建。图计算技术可以用于构建和维护知识图谱，将各种数据源的信息以图的形式组织起来，帮助搜索引擎、虚拟助手等更好地理解和回答用户的查询。

（六）物流和路径规划。图计算技术可以应用于物流和路径规划领域，分析交通网络、物流节点之间的关系，提供

最优的路径规划方案，减少交通拥堵和运输成本。

（七）数据中心管理。对于大规模的数据中心，图计算技术可用于资源调度、容错分析、故障诊断等，提高数据中心的运行效率和稳定性。

（来源：《2022 中国图计算技术及应用发展研究报告》；《人工智能之图计算》）

编辑部地址：南宁市体强路 18 号广西信息中心 1412 号房

联系电话：0771-6113592

电子邮箱：dsjyjs@gxi.gov.cn

网　　址：http：//gxxxzx.gxzf.gov.cn/

扫描二维码获取
更多决策参考信息