基于图论的产业网络知识图谱挖掘与构建

李振军[1]，刘祖军[1]，王鹏[1]，杨斌[2]，李大中[3]，郭钰[3]，赵华[1]
1. 智慧足迹数据科技有限公司，北京 100000；
2. 中国联合网络通信有限公司研究院，北京 100000；
3. 联通数字科技有限公司，北京 100000

摘要
我国是全球产业规模最大、产业覆盖最全的国家，但受多种因素的影响，发现产业链的堵点断点、识别卡点、寻找代替通路、全面优化产业链势在必行。从数据底座构建、核心知识图谱挖掘、兼容传统产业链知识3个方面，阐述了基于图论的产业网络知识图谱的构建过程，以实现产业优化升级与模拟仿真。分析了产业网络知识图谱的应用场景和优势，并给出了其在集成电路行业的应用案例。

关键词
图论；产业图谱；知识网络

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023078

Construction of industry knowledge graph based on graph theory

LI Zhenjun[1], LIU Zujun[1], WANG Peng[1], YANG Bin[2], LI Dazhong[3], GUO Yu[3], ZHAO Hua[1]
1. Smart Steps Digital Technology Company Limited, Beijing 100000, China
2. China Unicom Research Institute, Beijing 100000, China
3. China Unicom Digital Technology Co., Ltd., Beijing 100000, China

Abstract
China has the largest industrial scale and the most abundant industry type among the world. However, due to the influence of many factors, it is necessary to discover the blocking and breakpoints of the industrial chain, identify the stuck points, find alternative channels, and comprehensively optimize the industry. Then, this article explained the construction process of the industry knowledge graph from three aspects: data base construction, core knowledge graph mining, and compatibility with traditional industry chain knowledge. Finally the application scenarios and advantages of industry knowledge graph was analyzed, and an application case was given in the integrated circuit industry.

Key words
graph theory, industry map, knowledge network

0 引言

产业经济是国民经济的命脉。我国是全球产业规模最大、产业覆盖最全的国家，但是近年来，受多种因素影响，多条产业链因为上游断供、技术封锁等外部原因而存在“卡脖子”风险。为应对此类风险，国家出台10余项产业链相关政策，提出“强链、补链、延链”等相关要求。在双循环、统一大市场、数字政府经济监测和“强链、补链、延链”等的指引下，发现产业链堵点断点、识别卡点、寻找替代通路、全面优化产业链势在必行。首要任务便是构建以完整生产要素为基础的产业网络知识图谱，从认知层面对产业链要素分布、发展逻辑、技术路线、参与主体等信息进行有机整合，为产业链优化升级打好基础、画好蓝图。进而借助有效的知识图谱进行产业优化升级与模拟仿真，针对性地保护产业网络弱势节点、优化冗余产业结构，以及制定保护性、预防性政策，以保证产业链正常运行。

随着各类产业规模的不断壮大、产业模式的推陈出新、产业要素的极大丰富，传统依靠产业背景知识和宏观经济知识人工构建知识图谱的方式已经无法真实反映日益复杂的产业网络的内在结构与运行方式。国际上，由于工业门类不齐全、跨国贸易导致的产业链分散化，其他国家尚无成熟技术进行全产业网络知识图谱的挖掘和构建。

作为全球产业覆盖最全的国家，我国对产业网络的研究方式还停留在人工构建知识图谱的阶段，目前尚无自动化、智能化解决从产业数据挖掘产业知识图谱的成熟技术，在产业知识图谱优化、模拟仿真的技术领域也基本处于空白状态。因此，针对产业网络进行智能化、自动化知识图谱挖掘是必由之路。

1 数据底座构建

产业内在逻辑关系十分复杂，尤其是局部环状结构已经完全超出简单的“产业链”的线性数据结构可以描述的范畴，故而传统的“产业链”概念已经不适用于描述产业内部结构。因此，需选取能够描述复杂拓扑结构且具备良好维度拓展性的图作为底层框架基础。图的拓扑结构复杂，十分适合描述复杂的产业网络内在逻辑关系。复杂拓扑结构的图往往能够兼顾平面图的图论基础和空间拓展性，相较产业链等线性数据结构有更好的描述性与兼容性，是后续知识图谱挖掘技术开发的坚实数据基础。

在具体实践中，选取图的细分领域中的德布鲁因图[1]作为理论基础框架，针对性研发适配产业网络的数据结构。德布鲁因图是一种压缩节点关联关系形成新节点的有向图，在基因测序领域的基因数据拼装上有成熟的应用，能够支持超大规模测序数据的拼装工作，能够胜任超大规模数据集的数据结构底座。

首先，其压缩节点关系形成新节点是一种恰当描述产业网络关系的数据处理方式，尤其是相较链状线性结构，德布鲁因图能够支撑环状结构的描述。第二，德布鲁因图具备优秀且高效的节点增删特性。相较普通有向图的节点增删，德布鲁因图以删除点代替传统有向图需要进行边搜索的复杂环节，时间算法复杂度大大降低，更适合支撑超大规模产业网络的构建。第三，德布鲁因图具备良好的可拓展性，尤其是其对其他内存优化数据结构的兼容性。在德布鲁因图框架下，通过在节点上引入哈希表或FM-Index技术，可提升内存的使用率，提高其可覆盖产业网络的规模。

2 产业网络数据核心知识图谱挖掘

要构建自研的数据结构以描述超大规模真实产业网络，进而发掘产业网络知识图谱，除了上述构建复杂、增删节点缓慢等难点外，还有潜在情况导致对产业网络的调用失效[2]。产业网络在实际落地后，往往会经人工输入所需调取内容而进行目标子图的搜索。但是由于超大规模产业网络数据结构的复杂性，仅依靠人力难以精确搜索全量节点，对记忆模糊或者错误的节点名称通常会搜索失败。为了避免因这种问题影响产业网络数据结构的启动和调用，在实施中有必要植入自研的节点模糊搜索技术，支持搜索非完全正确的节点，并返回修正距离最近的节点结果。这种节点模糊搜索技术以Smith-Waterman算法为基础，进行了优化和二次开发。Smith-Waterman算法是一种基于动态规划策略开发的局部序列对比算法，可以针对某个序列搜寻与其相似度最高的片段，即搜寻编辑距离最短的片段。

借鉴Smith-Waterman的动态规划思想，可以将目标字段与其他字段形成的比较矩阵按照同字节的Index以树模型连接，按照树模型的分叉条件依次遍历Index对应值较大的比较矩阵，即可完成最短编辑距离的模糊搜索，且时间复杂度为$O(n)$，明显优于普通遍历匹配的复杂度$O(n^2)$。

将产业网络以标准化数据结构进行表达，可以得到有向图矩阵，该矩阵用横纵坐标Index描述节点的关联关系。以此为前提，以图论中的欧拉图理论为基础，研发产业网络矩阵的矩阵变精度分解算法。欧拉图作为图论中历史悠久的一类细分领域，有丰富的成熟定理与逻辑应用。根据欧拉图的定义，通过所有边且仅通过每条边一次的通路为欧拉通路，若欧拉通路首尾相连，则其为欧拉回路。具有欧拉回路的图被定义为欧拉图，而存在欧拉通路但不存在欧拉回路的图为半欧拉图。

在产业网络中，若忽略有向图特性而仅考虑连通性，产业网络矩阵则可变为对称矩阵。依照弗勒里算法，若连通图中存在欧拉环游，那么总能在多项式时间内找到一个欧拉环游回路。对于已找到的欧拉回路，总能通过矩阵的线性变换得到若干个分块对角矩阵。重复对分块对角矩阵进行欧拉回路的分解和线性变换，则可得到产业网络中全量变精度的子矩阵。如此便能够以变精度的欧拉回路定义产业网络中的重要节点群，即子产业；以欧拉回路的分解关系定义产业网络中的产业从属关系和细分关系。以无向欧拉图为理论基础在产业网络中发掘从属关系，定义产业知识图谱中的实体要素[3]。

在前述迭代寻找无向欧拉回路的过程中，同时对每个经过分解的子矩阵的原有向图（即生成该对称矩阵的原矩阵）进行有向图欧拉通路的搜索，迭代分解产生若干有向半欧拉图。从同一矩阵分解的半欧拉图即该产业网络中的“产业链”，在知识图谱中即同级关系；“产业链”中的欧拉通路的节点即可被清晰定义为知识图谱中该细分“产业链”中的上下游关系。经过上述对产业网络的双矩阵迭代分解，就可以得到产业网络的知识图谱[4]。

在上述矩阵分解的迭代过程中，在对无向图矩阵进行欧拉回路分解的同时，可以对产业网络中的子产业进行优化。针对无向欧拉回路对应矩阵，若进行行列变换能使该矩阵不满秩，则该矩阵对应的子产业存在冗余产业，可对其进行去冗余优化。在对有向图矩阵进行半欧拉图分解的过程中，若存在不可继续拆分的半欧拉图，则该有向图对应的产业为不可替代“产业链”，需要被特殊保护。

3 兼容传统产业链知识

结合计算机科学与知识图谱科学的跨领域先进技术与算力，能够得到远超人力所能处理的超大知识图谱。但是，在构建产业知识图谱的实践中，仍不能忽略前人用传统方法论结合宏观经济学与产业背景知识形成的“产业链”知识图谱以及相关知识积累[5]。

以哈密顿图为理论基础的映射匹配技术，能够将传统“产业链”映射到产业网络知识图谱的关键节点链条，形成新技术与传统知识积累的融合升级，扩大知识图谱的应用范围。

哈密顿图是图论体系中应用广泛且理论成熟的细分领域，在路径规划等行业有广泛的应用。通过图中每个节点一次且仅一次的通路被称为哈密顿通路，而存在哈密顿通路的图则被称为哈密顿图。

为了将产业链映射到产业网络知识图谱中，首先可以通过前述节点模糊匹配技术将产业链的首尾节点映射至产业网络的基础节点[6]。通过产业网络知识图谱对基础节点的描述，可以快速定位到产业链在产业网络中所处的子矩阵。

寻找哈密顿圈是NP困难问题，仅能够通过遍历完成匹配映射，对于大规模产业链的匹配十分消耗时间与算力。因此需要对哈密顿圈搜寻算法按照目标进行定制化改进。

对于产业链向量，通过算法产生若干产业链补全向量，使每个产业链补全向量的维度等于前述产业网络的子矩阵的行秩。分别将产业链补全向量与产业网络子矩阵求积，产业网络子矩阵与产业链补全向量的正交部分会成为0，通过反推前述计算所得的若干积中的0元素的位置，则可将产业链中的关键节点映射至产业网络中的节点或节点群。按照已经定位的节点将产业网络矩阵持续分解，重复上述过程，直至全部的产业链补全向量与产业网络子矩阵的积无正交维度，则此产业网络矩阵被定义为不可分[7]。不可分矩阵可以直接映射至当前产业链。至此完成产业链对产业网络的映射匹配，得到产业链在产业网络中所处的位置。

对于已经成熟应用的产业链相关产品与服务，可以通过上述映射匹配技术将其引入产业网络中的模块，实现对有成熟应用市场的传统知识积累的向下兼容和融合补充。

4 产业网络知识图谱的应用与优势

产业网络知识图谱即产业研究数字化，可提升产业研究效率。通过搭建大数据平台，利用科技赋能产业研究，为投资者、企业、政府提供对应的功能和深度研究，降低产业研究中的信息不对称程度，解决各类主体产业研究中的痛点。

传统的产业研究存在以下问题。一是产业信息孤岛化，研究一条链的人力成本和时间成本很高；专业行业研究人员一般专精一条或两条链，很难掌握全产业链知识；二是传统产业研究覆盖的企业有限，一般只覆盖头部企业、上市企业等，很难做到对节点未上市企业、中小型企业等的全覆盖；三是数据实时更新困难，主要体现在数据量大，需要覆盖产业、行业、企业三大层面信息，缺乏前沿信息实时跟踪。产业网络知识图谱可提供所有产业链的节点信息，包括产业信息、行业信息和企业信息，并进行追踪更新，可帮助用户迅速了解各条产业链的基本情况，提高用户的学习效率。

面向企业，产业网络知识图谱可辅助市场分析、明晰企业定位、构建供应体系等。帮助企业了解所属产业的全貌，明确企业所处的节点和地位，分析企业所处市场的整体竞争格局，了解现阶段主要/潜在竞争对手，审视自身的优

劣势所在，及时进行战略转型。根据企业在产业链的覆盖情况，寻找存在关联性、互补性的企业，实现互惠互利、协同发展。

面向政府，产业网络知识图谱可辅助梳理产业现状、制定产业规划、开展精准招商。帮助政府纵观产业链的概况和全景，分析和对比地区各产业的企业发展情况；通过了解产业链上中下游节点，以及对应的市场规模、关联企业等数据信息，总结归纳本地企业在目标产业链的分布情况及具体特征，结合整链的价值传输逻辑，发现本地的优势、薄弱之处，从而定位关键突破环节、产品及技术，确定产业组合、重点项目；参考产业链各节点提供的企业综合实力，并深入分析企业投资能力及意向，结合地区产业和项目规划，确定重点培育企业和招商目标企业，进行“强链”和“补链”，打造全要素完备的产业集群。

目前，基于图论的产业网络图谱挖掘与构建技术已应用于产业图谱绘制，逐渐完成了动力电池、涤纶长丝等120余条产业链图谱的构建（示例如**图1~图4**所示）。例如服务某大型产业集团，实现对张江高科园区的产业洞察、企业监测等功能；服务某地工信部门，监测评估当地主导产业发展情况；服务某地商务部门，针对当地九大主导产业，进行强链、弱链、补链分析，并提供精准招商功能及服务。

5 案例分析

以集成电路产业网络知识图谱建设为例。近年来，集成电路行业主流的横向水平分工模式面临着严峻挑战，多个国家纷纷开始尝试构建自给自足的芯片供应链。为积极应对当前全球形势变化，提升产业链的自主可控水平，特建集成电路产业网络知识图谱。

一要逐步建立全面准确的全球集成电路行业数据库，通过数据解读各国（地区）产业链供应链分工、行业上下游供需变化、国际贸易流向变化、重大技术变革、相关产业政策等，进一步掌握企业的战略发展方向、主营业务（产品）与技术水平、全球业务布局情况、重大经营行为（投资并购/扩产停产等）、经营业绩情况、供应链信息等。二要梳理行业内、企业间的关联关系，构建知识图谱，包括行业图谱、企业图谱等。三是针对重大事件，模拟风险传导效果，评估重大事件对企业、产业内、产业间的关联影响，并辅助制定应对策略[8]。

（1）构建产业数据库

汇聚政府数据、协会数据、市场机构数据、企业数据、舆情数据等，统一治理形成覆盖全行业的高频数据库。包含1万家以上企业、100多万条数据，涵盖每家企业的基本信息、地区分布、上市企业财务指标、上市企业销售信息、设备招投标、企业投融资、重大项目（审批备案）、新闻舆情等高频数据。

（2）构建产业网络知识图谱

集成电路产业链涵盖设计、制造、封测、原材料、设备等、分销等一级环节，光刻机、光刻胶、EDA软件等40余二级环节[9]。

通过自研数据结构来描述超大规模真实产业网络，进而发掘产业网络知识图谱，流程如**图5**所示。

（3）构建企业关系图谱

以单家集成电路企业为例。第一，建立企业社交关系和企业复杂关系网络，涉及企业之间、股东之间、知识产权、主要管理人员之间、法律诉讼原被告等100多种关系[10]；第二，建立多种分析模型，包括企业图谱、供应链图谱、多节点关联等。

单节点企业关系图谱：将企业的股东、供应商、下游客户、高管、对外投资、分支机构、历史股东、历史高管、历史对外投资、历史法人，以树形结构展现出来，企业主要人

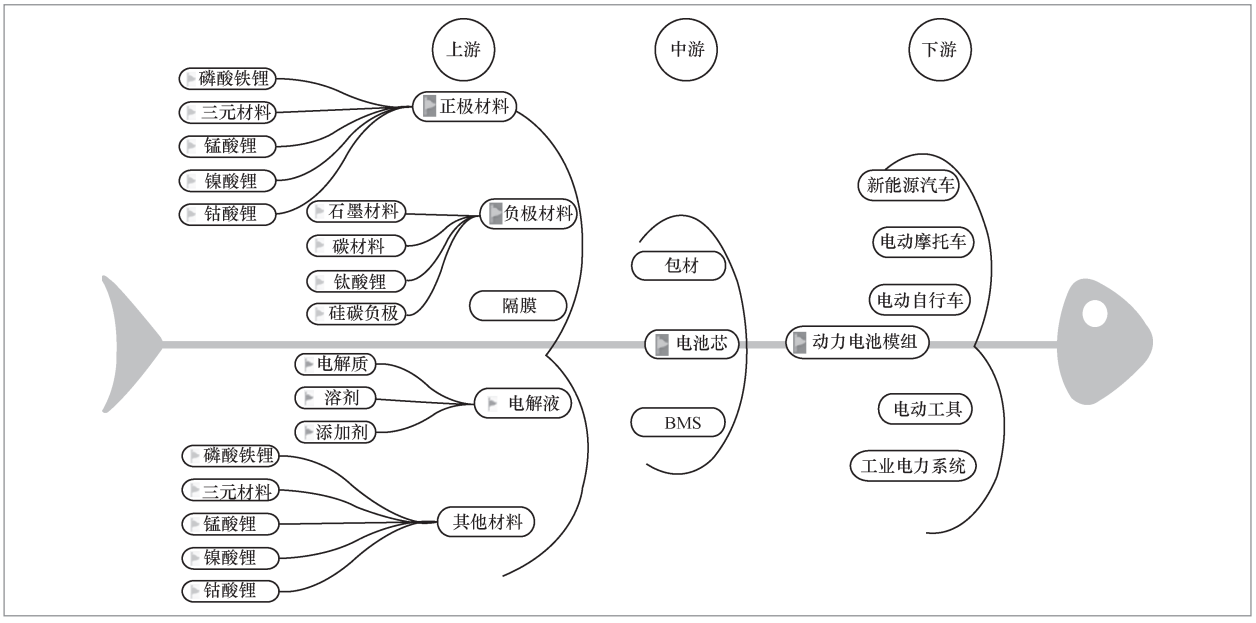

图 1　动力电池产业图谱

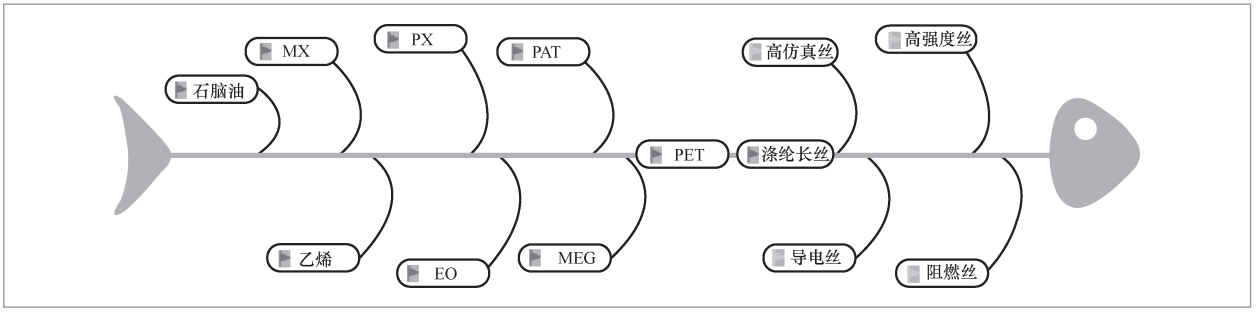

图 2　涤纶长丝产业图谱

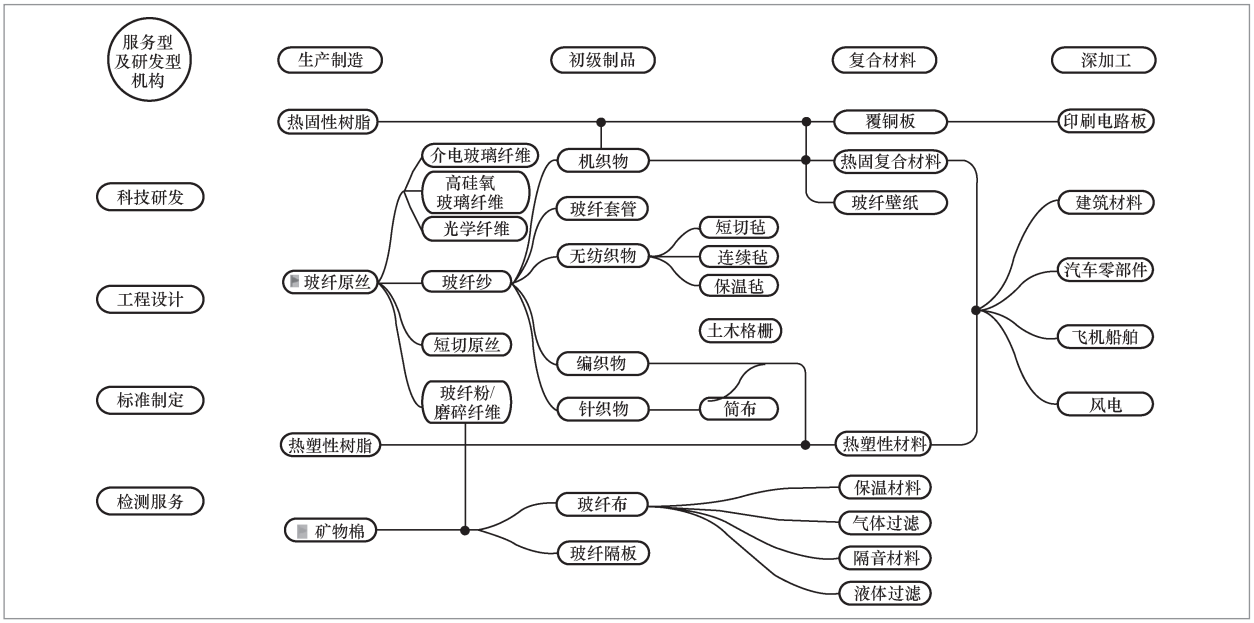

图 3　玻纤及复合材料产业图谱

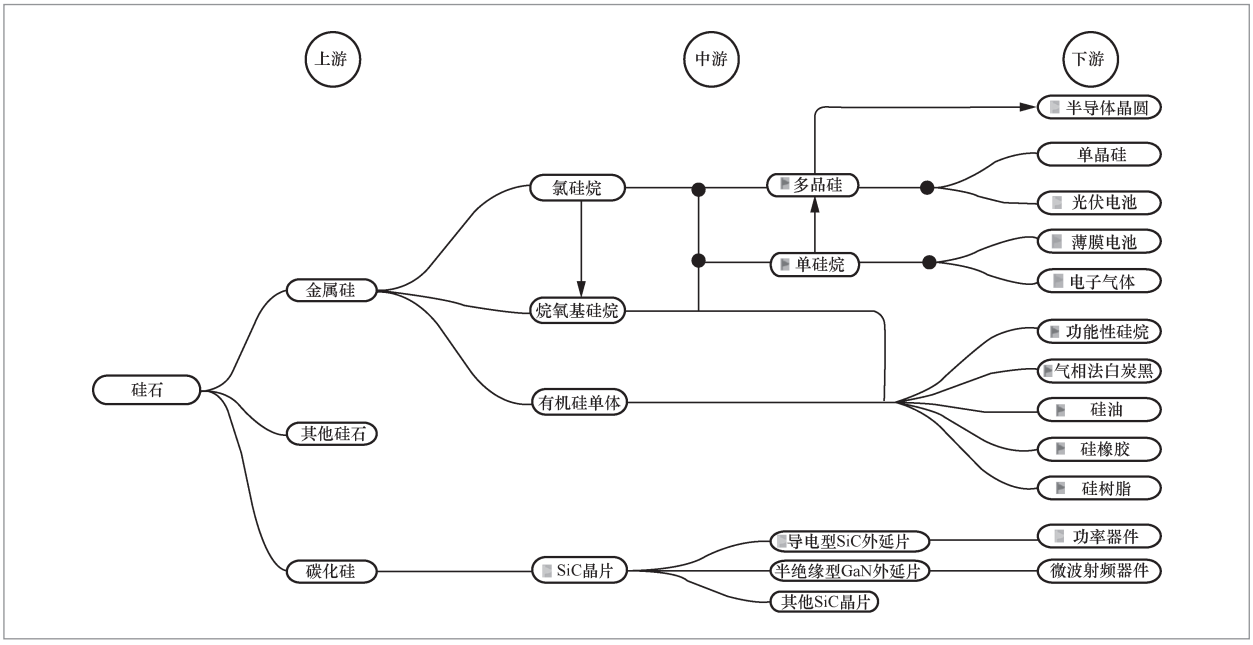

图 4 硅基新材料产业图谱

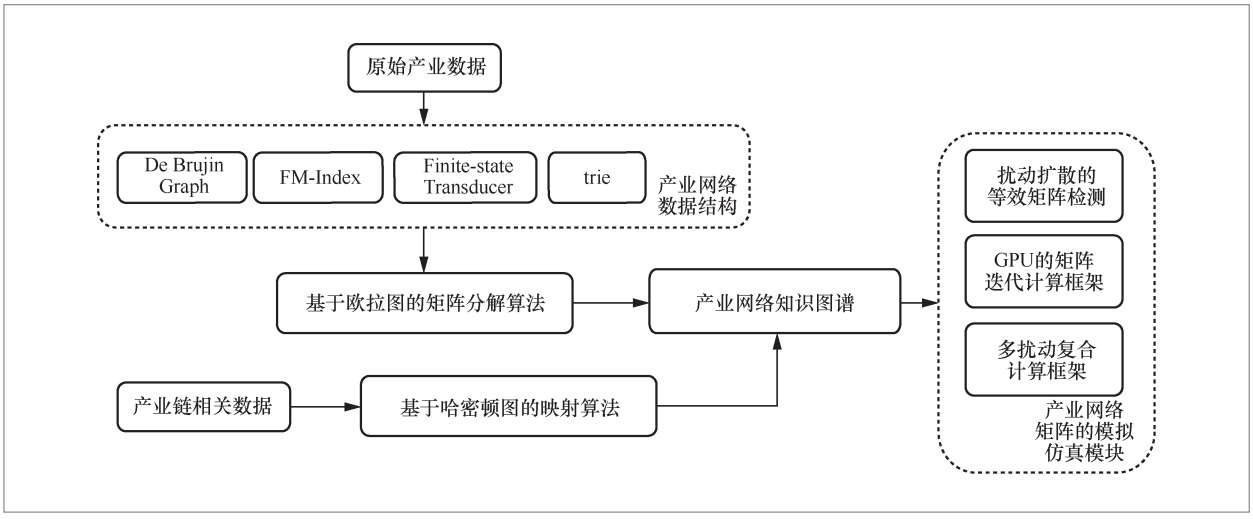

图 5 产业网络知识图谱挖掘流程

员和关联企业一目了然。点击企业节点可查看企业简介、产业标签、供应商、下游客户、股东、对外投资、主要管理人员等信息。

多节点关联探寻：在一个基于公司股权、高管任职、专利、招投标、涉诉、供应关联等关系信息形成的企业复杂关系网络中，探寻任意多个企业之间是否存在关联关系，根据关联路径挖掘目标企业谱系中是否存在异常关联，并根据关系路径长短，衡量企业间关系的密切程度。系统会随机找出20条路径，并按路径长短顺序排列。此功能需结合层级和关系使用，当图谱较复杂时，可通过减少层级和关系来简化图谱，找到较短路径。

6 结果分析

集成电路产业网络知识图谱实现了对全行业的产业数据、企业数据的统一汇集、治理及展示，可帮助政府及企业用户快速认清行业全貌，理清本地家底，做好集成电路产业的运行监测工作，进一步针对重点招商环节，构建精准招商模型，辅助招商引资。此外还可用于研究突发事件对集成电路企业的影响。例如，针对荷兰宣布DUV光刻机出口限制事件进行实战训练。根据国内企业与荷兰ASML公司的供应关联，快速识别出直接受影响的晶圆厂，并向上游原材料及设备、下游应用领域逐级传导，模拟断供风险传导路径，评估此事件对国内产业链的总体影响。利用产业网络知识图谱可快速识别出受影响的企业，全程不超过8小时，而传统研究方式通常需要数天才能完成同类工作。

7 结束语

本文利用“图”这种基本且通用的“语言”和“高保真”的方式构建产业网络图谱，非常直观、自然、直接、高效地描述了产业节点、企业间的纷繁复杂的关系，有效地解决了产业数据量大、散、乱，关系复杂等难题，降低了用户产业链研究的学习成本和时间成本，并可进行产业优化升级与模拟仿真，针对性地保护产业网络弱势节点、优化冗余产业结构，并制定保护性、预防性政策，以保证产业链供应链正常运行。

参考文献:

[1] GRIGORCHUK R, LEEMANN P H, NAGNIBEDA T. Lamplighter groups, de Brujin graphs, spider-web graphs and their spectra[J]. Journal of Physics A: Mathematical and Theoretical, 2016, 49(20): 205004.

[2] LI X W, MA J M, YU J, et al. A structure-enhanced generative adversarial network for knowledge graph zero-shot relational learning[J]. Information Sciences, 2023, 629: 169-183.

[3] CHEN H, DENG W W. Interpretable patent recommendation with knowledge graph and deep learning[J]. Scientific Reports, 2023, 13: 2586.

[4] KANG S, SHI L, ZHANG Z Y. Knowledge graph double interaction graph neural network for recommendation algorithm[J]. Applied Sciences, 2022, 12(24): 12701.

[5] ZHOU Z W, TING Y H, JONG W R, et al. Knowledge management for injection molding defects by a knowledge graph[J]. Applied Sciences, 2022, 12(23): 11888.

[6] CHOI J. Graph embedding-based domain-specific knowledge graph expansion using research literature summary[J]. Sustainability, 2022, 14(19): 12299.

[7] ZHOU B, SHEN X W, LU Y Q, et al. Semantic-aware event link reasoning over industrial knowledge graph embedding time series data[J]. International Journal of Production Research, 2023, 61(12): 4117-4134.

[8] CHU Y F, DAI M M. Research on risk spread model of industrial chain[J]. Grey Systems: Theory and Application, 2014, 4(2): 328-338.

[9] WORLD L. Recent progress of the integrated circuit industry in China – overview of the manufacturing industry[J]. Journal of Microelectronic Manufacturing, 2019, 3(3): 1-8.

[10] Lithotechsolutions World. Current status of the integrated circuit industry in China[J]. Journal of Microelectronic Manufacturing, 2018, 1(1): 1-8.

作者简介

李振军（1978– ），男，智慧足迹数据科技有限公司总经理，主要研究方向为产业经济、大数据、人工智能。

刘祖军（1992– ），男，智慧足迹数据科技有限公司算法工程师，主要研究方向为协同过滤推荐、自然语言相关技术、电信信令处理相关技术。

王鹏（1992– ），男，智慧足迹数据科技有限公司产业研究员，主要研究方向为产业分析、精准招商、企业分析、战略新兴产业、大数据应用。

杨斌（1986– ），男，博士，中国联合网络通信有限公司研究院首席研究员，主要研究方向为图神经网络、推荐算法。

李大中（1976– ），男，联通数字科技有限公司技术研发部总经理，主要研究方向为人工智能、区块链、数据治理、大语言模型、图谱技术。

郭钰（1994– ），女，联通数字科技有限公司项目实施与交付工程师，主要研究方向为系统分析与设计、数学建模与算法研究、人工智能、大数据、隐私计算。

赵华（1979– ），女，智慧足迹数据科技有限公司副总经理，主要研究方向为经济分析、产业分析、经济管理。

收稿日期：2023-09-28