AI TR
Tsinghua

人工智能之图数据库

Research Report of Graph Database

2020年第4期

清华大学人工智能研究院

北京智源人工智能研究院

清华-中国工程院知识智能联合研究中心

2020年7月

前　言

随着互联网+、社交网络、智能推荐等大数据的迅猛增长，大批 NoSQL 数据库已经成为互联网开发的新标配。对于大数据中关联关系的处理，图数据库的处理性能远超其他类型数据库，被广泛应用于金融、工业、政务、零售、电信和生命科学等各学科和工业领域，受欢迎程度遥遥领先。与此同时，图数据库也面临着底层设计和上层语言表达的多重挑战。

本期，我们选取图数据库作为 TR 报告的主题。报告围绕图数据库的基本概念、技术发展、产业应用、人才概况和热点趋势五大方面进行深入挖掘。其中基本概念、技术发展和产业应用章节由国内领先的图数据库团队即陈文光教授带领的清华团队完成。该团队在 OSDI、EuroSys、ATC 等顶级会议中发表过多篇相关论文，他们编写的图计算系统具有业界领先的性能，并在金融、互联网等多个领域得到实际使用。

此外，报告的人才概况和热点趋势章节依托清华大学唐杰教授自主研发的“科技情报大数据挖掘与服务系统平台”（简称 AMiner），以及第三方机构研报、媒体报道等公开资料，通过人工智能、大数据分析与挖掘、知识图谱、自然语言处理等技术，并结合文献计量学等情报学方法制作生成。

报告的数据来源与研究方法

1. 数据来源

本报告中与图数据库领域相关的人才数据来均自于 AMiner 系统。系统采用数据挖掘和社会网络分析与挖掘等技术，提供研究者信息抽取、研究者社会网络关系识别、研究者能力图谱、审稿人智能推荐等功能，提供研究者和研究领域的全面知识，为科研管理和服务提供有力支撑。平台自 2006 年上线以来，经过十多年的建设发展，已建立运作良好的数据采集及集成更新机制，收录论文文献超 3 亿，专利 1 亿，学者 1.3 亿，其中超过 50 万的学者经过了人工标注与审核吸引了全球 220 个国家/地区 1000 多万独立 IP 的访问，年度访问量 1,800 余万次。

2.学者及研究领域筛选方法

本次报告中的人才和技术篇采用大数据挖掘技术，对图数据库领域内的学者信息进行深入挖掘，参考 h-index、发表论文数、论文被引频次等指标，对学者信息进行筛选，比较和分析了图数据库领域人才在全球和国内的分布概况，领域的技术研究发展趋势，以及技术领先国家、机构趋势。

（1）由图数据库顾问组推荐期刊/会议列表和领域关键词，推荐的期刊/会议为数据管理国际会议（The ACM Special Interest Group on Management of Data, SIGMOD）、超大型数据库国际会议（International Conference on Very Large Databases, VLDB）、IEEE 国际数据工程会议（IEEE International Conference on Data Engineering, ICDE）、图形数据管理经验与系统国际研讨会（International Workshop on Graph Data Management Experiences & Systems, GRADES）、扩展数据库技术国际会议（International Conference on Extending Database Technology, EDBT）。领域关键词具体包括：图数据库（Graph databases）、属性图（Property graphs）、资源描述框架（Resource Description Framework, RDF）、图分析（Graph analysis）、ACID 事务属性（Atomicity, Consistency, Isolation, Durability, ACID transaction）、图匹配（Graph patterns）。

（2）通过 AMiner 大数据平台对 2000～2019 年发表在推荐期刊/会议的论文进行采集和清洗，并对论文作者信息进行深度挖掘；

（3）基于专家顾问推荐的领域关键词，根据论文作者的研究兴趣标签、作者名下的所有论文标题和摘要，筛选与图数据库领域相关，且 h-index 排名最靠前的 2,000 位研究学者；

（4）综合运用知识图谱、自然语言处理、可视化、文献计量学等技术手段，基于论文和学者数据，分析得出图数据库领域的技术研究发展趋势，以及技术领先的国家、机构趋势。

3.代表性学者画像

“学者画像”是 AMiner 平台的核心服务功能之一，其具体示例如图 1 所示。学者画像的特色在于除了提供专家学者如姓名、单位、地址、联系方式、个人简介、教育经历等个人基本信息之外，还利用团队多年的命名排歧相关技术基础，建立了较为完全的学者 — 论文映射关系，分析挖掘学者学术评价、研究兴趣发展趋势分析、学者合作者关系网络等信息。

图 1 代表性学者画像示例

4.领域热点话题

为了帮助读者了解图数据库领域的热点研究话题，本报告针对 AMiner 平台上收录的专家推荐的 100 篇必读论文（https://www.aminer.cn/search/pub?q=Cognitive%20Graph），采用主题生成模型（Latent Dirichlet Allocation, LDA）分析了这些论文的研究主题分布情况[1]。

[1] LDA 模型. [EB/OL]https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

目 录

图目录

表目录

1 概述篇

随着万物互联的 5G 时代到来，图数据库在人工智能、计算科学、生物信息、金融科技、社交网络等越来越多的领域发挥着举足轻重的作用。截至 2019 年 6 月，支付宝及其本地钱包合作伙伴已经服务超 12 亿的全球用户，中文网页数量达到 2.7 千亿，网页链接数量达到 12 万亿（2018 年），人脑神经突触链接数更是达到了百亿级别[1]。面对各种海量数据、尤其是对海量非结构化数据的存储，传统的信息存储和组织模式已经无法满足客户需求，图数据库却能够很清晰地揭示各类复杂模式，尤其针对错综复杂的社交、物流、金融风控行业，其优势更为明显，发展潜力巨大。

1.1 概念

图数据库（Graph Database）是一个基于图模型的在线数据库管理系统，具有图数据的创建（Create）、读取（Retrieve）、更新（Update）和删除（Delete）功能，简称 CRUD[2]。图数据库主要面向事务系统（On-Line Transaction Processing, OLTP）。另外，Neo4j、TigerGraph、ArangoDB、JanusGraph 等图数据库通常也会支持一些分析类的任务[3-4]。

1.1.1 图模型

图模型（Graph Model）是图数据的一种抽象表达，其中属性图模型（Labeled Property Graph Model, LPG）的使用最为广泛。以图 1-1 为例，图模型由顶点，以及连接顶点的边构成基础的图拓扑。除此之外，每个顶点和每条边均有自己的标签（Label），该标签定义了该顶点或边拥有的一个或多个属性。顶点、边、属性构成了属性图，其符合人们对客观事物的直观认识，在具体实现中，还分为强类型和弱类型、是否支持边标签、是否支持多标签等。

以王家卫的重庆森林电影为例，具体如下所示：

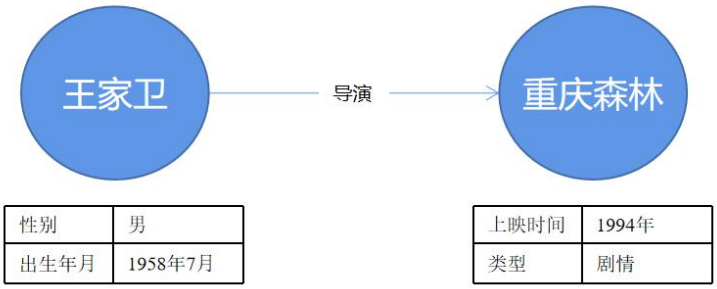


图 1-1 图模型实例

- 人物顶点“王家卫”，其属性包括“性别”为“男”，“出生年月”为“1958 年 7 月”等；
- 电影顶点“重庆森林”，其属性包括“上映日期”为“1994 年”，“类型”为“剧情”等；
- 导演边，从顶点“王家卫”指向“重庆森林”，边上属性为空。

上图构成了一个简单的图模型，如果有其他的关联关系比如演员、获奖情况、王家卫的其他电影等，同样也可以加入到这个图模型中。

另一类广为人知的模型是 RDF（Resource Description Framework）模型，它最早由 W3C 组织于 1999 年提出。RDF 用三元组（Subject, Predicate, Object）来表示实体的连接关系，每个元素有全局唯一的标识。目前 RDF 在知识图谱领域已经有比较成熟的工具链，它与属性图模型之间可以等价转换。

图模型的处理可以分为两类，一类是面向事务的联机事务处理（Online Transaction Processing，OLTP），主要解决实时增删查改的数据操作；另一类是面向分析的联机分析处理（Online Analytical Processing，OLAP），主要解决图上复杂迭代计算的效率问题。图数据库侧重 OLTP，需要满足 ACID 的事务特性，即原子性（Atomicity）、一致性（Consistency）、隔离性（Isolation）、持久性（Durability）。在复杂数据分析方面有所欠缺，典型操作为图上的局部计算；图分析引擎（Graph Analytical Engine/Graph Computing System）侧重 OLAP，典型操作为全图迭代计算。部分图数据库系统能同时支持 OLTP 和 OLAP，但由于数据访问模式和操作方式差距较大，会采用松耦合的方式，或者在设计上有所侧重。

OLTP 是局部图的简单操作，操作包括增删查改四种，每次操作涉及的数据仅为局部的小图，比如几十个顶点和边。OLAP 是全局图的复杂分析，操纵仅为

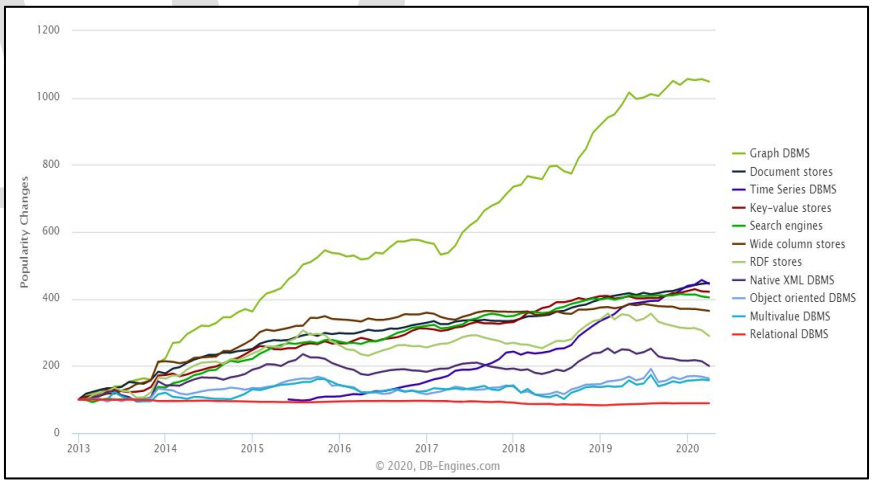
查询，但每次操作涉及的数据通常为全图的多次迭代，可能为上亿个顶点，上百亿条边。一个真实场景的应用需根据数据操作的模式来选择合适的底层系统。

本报告更侧重于图数据库的讨论，关于图分析引擎的讨论请参考 TR 系列的其他报告。

1.1.2 图数据库

图数据库是基于图模型，使用图结构进行语义操作的数据库，它使用顶点、边和属性来表示和存储数据，支持数据的增删查改操作。

根据全球知名的数据库流行度排行榜网站 DB-Engines（https://db-engines.com/en/ranking_trend）数据显示，图数据库的关注度增速远超其他类型的数据库[5]。更值得一提的是，全球最具权威的 IT 研究与顾问咨询公司 Gartner 在 2019 年的数据与分析峰会上预测 2020 年以后，全球图处理及图数据库的应用市场都将以每年 100%的速度迅猛增长[3]。



数据来源：DB-Engines 官网

图 1-2 图数据库的关注度

1.2 图数据库的历史发展

图数据库的起源可以追溯到 20 世纪 60 年代，引导式数据库（Navigational Database，比如 IBM 的 IMS）采用树状的结构来表示数据之间的分层关系，对图结构的支持可以通过虚拟顶点来完成。到 80 年代，支持属性图模型的图数据开始出现，包括 Logical Data Model 等[4]。

21 世纪初，商用图数据库开始崭露头角，比如 Neo4j 和 Oracle Spatial and Graph 等，并支持事务性 ACID。其中隔离性包括多个不同的隔离级别，从低到高分为未提交读（Read Uncommitted）、提交读（Read Committed）、可重复读（Repeatable Read）、序列化读（Serializable）。对事务的支持是数据库的标准配置，只有支持事务才能保证数据同时读写不会出现不可预知的错误。自图数据库支持事务后，其市场和应用有了爆发式的增长。

2010 年后，图数据库朝着多个不同的方向发展，包括支持大规模分布式图处理、支持多模态、图查询语言的设计、专用硬件的适配等（图 1-3）。

我们从如下几个方面来讨论分布式图数据库。一是高可用性（High Available），当某台服务器失去响应后，图数据库能否持续正常运行？一般需要多机热备技术来支持高可用性。二是读性能，即单台服务器的并发读不能满足性能需求，多机部署能否提高并发性，同时保证图数据库原有的事务特性。三是写性能，即能否将数据分片（Sharding）存储到多台服务器上，提高写事务的性能并保证水平扩展。图数据在模型上天然存在数据随机访问的特性，导致图数据库的系统实现一旦涉及分布式多机通信，性能就会断崖式下降。因此单机或分布式图数据库的选择，是个功能和性能的权衡问题。如果数据容量在单机系统上（目前商用服务器通常使用 256GB 内存，最大可达 4TB）可以得到满足，则不必使用分布式系统。

在近几年图数据库技术的介绍和宣传中，经常会提到一个词叫“原生图”（Native Graph），一般指的是跳过索引的邻居访问（Index-free Adjacency），需要对底层存储做不同于传统数据库的设计，是性能优化的一种方式。系统整体的设计和硬件特性、读写负载等均有关系，通常是个权衡的过程，无法在所有方面做到面面俱到，因此需要在具体的场景或评测程序中衡量。

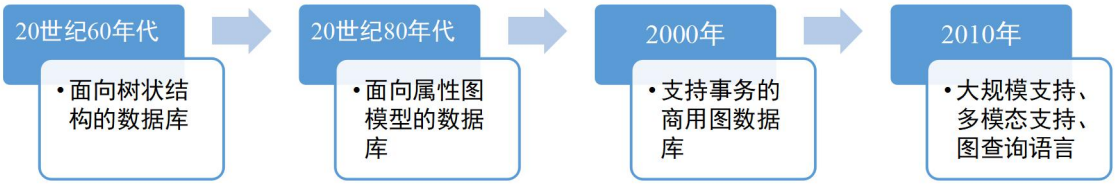


图 1-3 图数据库的发展史

1.3 图数据库的特征

1.3.1 优势

图数据库是面向关联关系图数据的存储数据库，在基于图的数据增加、删除、查询、修改等方面做了不同于其他数据库的设计。在图数据的操作抽象上，采用基于顶点的视角，比如顶点通过其所有出边访问其邻接顶点，这一类的操作也是图数据库系统设计的核心。

除了图数据库外，常见的数据库还有关系型数据库、键值对数据库等，图数据库的独特性可以体现为以下三个方面：

- **性能**

在关联关系的处理上，关系型数据库处理不可避免要用到表的 JOIN 操作，非常影响性能。而图数据库则是类似指针直接跳转访问，在典型查询上比关系数据库通常有 2 到 3 个数量级的性能优势。

- **兼容性**

现实中的项目通常是不断演进的，意味着数据内容甚至数据格式也会不断发生变化。在关系型数据库中，这意味着表结构的变化，或者多个新表的建立，对源数据的改动非常大。而在图数据库里，仅需添加新的顶点、边、属性，设置为对应的类型即可。从本质上说，一个表代表一个类型的数据，一个顶点代表一个特定的数据，意味着关系数据库更关注数据的类型，而图数据库更关注数据的个体，识别其关联关系。

- **直观性**

顶点和边的图模型相比于表模型更符合人的思维方式。比如我们面对面用纸笔交流介绍社交网络关系，会自然而然地使用点边的方式画出来，这正是图模型。

如果采用关系型数据，先将人物建表，再将关系建表，最后将数据进行映射，需要高度的抽象思维。在图数据上进行分析查询时，也可以直观地通过点边连接的拓扑，交互式找到想要的数据。因此有一种说法是：与关系型数据库相比，图数据关系才是真的处理“关系”的。

1.3.2 数据库横向对比

目前市面上存在各式各样的数据库，一般分为关系型数据库（SQL）和非关系型数据库（NoSQL，Not Only SQL），后者又分为键值存储数据库（Key-Value DB）、列存储数据库（Column Family DB）、文档型数据库（Document DB）和本文的主题—图数据库（Graph DB）。数据库在设计的实现上没有优劣高低之分，而是对不同的使用场景的适应性不同，比如高度结构化的数据在关系型数据库中能实现快速的逐行访问和数据一致性，海量的简单数据在键值对数据库上有最优的可扩展性，关联关系则在图数据库上有最好的模型和性能。

下表对五类数据模型的优劣分别做了简单介绍。

表 1-1 五类数据库对比

分类	模型	优势	劣势	典型系统
关系型数据库	表结构	数据高度结构化，一致性强，软件成熟度高	面向多跳的关联关系查询低效或不支持	MySQL Oracle
键值数据库	哈希表	查找速度快	数据无结构化，通常只被当作字符串或者二进制数据	Redis
列存储数据库	列式数据存储	查找速度快；支持分布横向扩展；数据压缩率高	数据插入效率偏低、按行的数据操作性能受限	HBase
文档型数据库	键值对扩展	数据结构要求不严格；表结构可变；不需要预先定义表结构	查询性能不高，缺乏统一的查询语法	MongoDB
图数据库	图结构	针对关联关系的建模、操作非常高效	高度结构化的数据处理能力不及关系型数据库	Neo4j、JanusGraph

1.4 图数据的未来挑战

大数据的到来，使得图数据库脱颖而出，在关联关系上的处理性能远超其他类型数据库，同时对图数据库的方方面面提出了更高的要求，既有底层的系统设计，也有上层的语言表达。举例如下：

- **大数据的挑战**

在全民上网的时代，中国人口 14 亿，世界人口 75 亿，无论是社交分析还是资金转账，数据量都在十亿到千亿级别，而物联网的实体数更有两到三个数量级的增加。与其同时，这些数据在不断变化，不仅表现在数据量的持续增加，在数据丰富性上也不断提提升。

- **新硬件的挑战**

各式各样的新硬件层出不穷，包括 NVM、RDMA、FPGA、GPU 等，合理利用能大幅提升图数据库的功能和性能，从而对底层系统设计提出了更高的要求。

- **接口语言的挑战**

图数据库的发展还远没有关系型数据库成熟，因此各个学术机构及厂商都在各种探索的阶段。在接口语言方面，GQL 作为正在实施的图查询语言项目，尚需三到四年才能完善，那么需要学术机构及厂商在各自对图数据库定位和理解有更深的认识，才能做出有益的尝试。

- **数据建模的挑战**

图模型作为面向关联关系的强兼容性模型，同样需要大量的领域知识在现实场景到理论搭建桥梁，比如应该如何选择合适的数据，以及将那些实体抽象成顶点，哪些作为属性。另外项目通常不是一蹴而就，后期需要对模型进行扩展，对数据进行填充，这在模型建立之初应当予以考虑。

1.5 图数据库基准测试

图数据库的发展的成熟度还远不及关系型数据库，而图数据库的复杂程度又远超其他数据库，因此一个公认的基准测试方案显得尤为重要，近年来，对图数据库的评价方法逐渐达成了共识。图数据库的核心在于关联关系的处理，也就是

图模型中邻居的查询，本小节的评测讨论侧重于属性图模型，RDF 图也有类似的评测方法。

一个完备的基准测试应包含图数据的所有操作类型，在图数据中可以分为四类，包括本地查询（Local Queries）、邻居查询（Neighborhood Queries）、局部遍历（Traversal）、全局分析（Global Analytics）。本地查询是指查询只涉及单个顶点或单个边。举例来说，给定顶点 ID，通过索引查这个顶点对应的属性值，并进对应操作。这是最简单的查询方式，和其他数据库的查询无异。邻居查询是从某个顶点出发，沿着这个顶点的出边或入边，查询邻居顶点。该查询过程中，可以通过边和顶点的标签及其属性值进行过滤，筛选符合条件的结果。该查询方式也称为一度邻居查询。局部遍历是一个或多个顶点的多度邻居查询。通常在遍历的过程中，顶点和边上会有指定的限制条件，因此整个遍历过程中涉及的顶点数和边数不会太多，但遍历的条件可能会很复杂。全局分析其实就是图分析引擎的工作，需要对全图的所有数据做多次的迭代，最终得出想要的结果。

目前，图数据库的评测工具并不多见，主要包括针对核心部分的测试、数据导入导出测试、可靠性测试、可视化界面测试、ACID 事务测试等，各数据产品之间大同小异，本节将主要针对图数据库核心部分的测试进行介绍。

2013 年，Facebook 提出了 LinkBench，但并非针对大规模图数据库，也不再有人维护。部分图数据库厂家采用 KHop 的方式测查询性能，但该种方式覆盖的场景及其有限，比如没有读写并发的负载测试。国际上比较活跃并且相对权威的评测工具由 LDBC 提出，包括面向事务的测试标准 LDBC SNB，和面向分析的测试标准 LDBC Graphalytices，测试流程均包括了数据生成、性能指标生成、正确性验证。

LDBC SNB 模拟了一个社交网络的场景，数据包括人、博客、评论等，操作包括 29 个 Interactive 交互式操作任务和 25 个 Business Intelligence 分析任务来模拟实际社交网络的场景，比如发起新话题、查看最近回复、给评论点赞等，然后将操作按照合理的比例混合来并发地对图数据库进行访问，得出综合评分。29 个 Interactive 交互式操作任务又可以分为 14 个复杂只读查询任务（IC），7 个简单只读查询任务（IS）和 8 个事务型更新查询任务（IU），囊括了本地查询、邻居查询和局部遍历三种操作方式。Business Intelligence 分析任务其实是稍微复杂

点的局部遍历，相比较前面的 IC，其有更多的聚合操作，比如排序、求平均、取极值等。

LDBC Graphalytices 是图分析评测程序，测试了几个典型的图分析应用包括宽度优先搜索（Breadth-First Search），网页排序算法（Page Rank），弱连通分量算法（Weakly Connected Components），标签传播算法（Community Detection using Label Propagation）、单源最短路算法（Single Source Shortest Path）、局部聚类算法（Local Clustering Coefficient）。由于并行算法的结果并非是确定性的，LDBC Graphalytices 给出了每个算法结果正确性验证的方法，包括值正确（Exact match）、等价性正确（Equivalence match）、小数点正确（Epsilon match）。

以上两个评测程序的标准也在根据反馈不断更新中，同时有公布全网评测的结果。

2 技术篇

自 20 世纪 80 年代以来，随着信息技术的发展，关系型数据库技术日益成熟，但其局限性也日益突出：它能很好地处理所谓的“表格型数据”，却对技术界出现越来越多的复杂类型的数据无能为力[5]。90 年代以后，随着互联网+、社交网络、智能推荐等大数据的迅猛增长和其应用需求的推动，大量新型的非关系数据库出现，弥补了关系型数据库在处理海量关联数据时性能不足的缺点，而图数据库技术更是占据了其中的半壁江山。本章将重点对图数据模型、图数据的存储与管理、面向图数据的查询语言、图数据查询技术等相关技术进行简要介绍，并在最后列出了若干在业内具有一定知名度的图数据库，以作参考。

2.1 图数据模型

图模型（Graph Model）是图数据库表达图数据的抽象模型。目前主流图数据库采用的图模型主要包括资源描述框架（Resource Description Framework, RDF）和属性图（Property Graph）两种[6]。

2.1.1 RDF

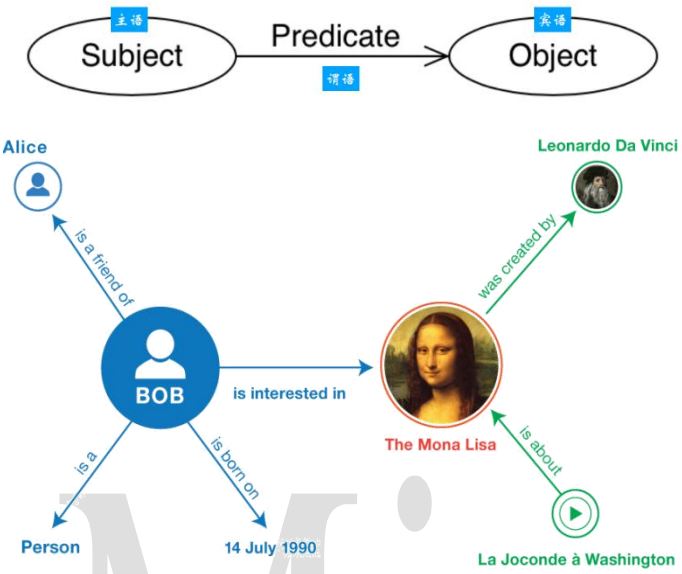
RDF 全称为资源描述框架，是由万维网联盟（World Wide Web Consortium, W3C）提出的一组标记语言规范技术，基于 XML 语法及 XML Schema 的数据类型以便更为丰富地描述和表达网络资源的内容与结构[7]。RDF 资源的三种表示形式分别为：URI、空白顶点（匿名资源）和 Unicode 字符串。

RDF 本质上是一个数据模型，它提供了一个统一的标准来描述 Web 上的资源，所谓资源可以指类（class）、属性（property）、实例（Instance）等等。RDF 在形式上表示为 SPO（subject, predicate, object）三元组（triple），即（主语/主体、谓语/属性、宾语/客体），用于描述具体的事物及关系，即实体以及实体之间的关系（图 2-1）。RDF 也可以表示为一张带有标记的有向图，图中有顶点和边，顶点对应实体，边对应关系或者属性，关系指的是实体之间、实体与属性之间的关系。

以图 2-1 为例，该图中的主谓宾三元组有：（BOB, is a, Person）、（BOB, is born on, 14 July 1990）、（BOB, is a friend of, Alice）、（BOB, is interested in, The

Mona Lisa）、（The Mona Lisa, was created by, Leonardo Da Vinci）、（La Joconde a Washington, is about, The Mona Lisa）。每个主谓宾三元组中的主语、宾语都是顶点，谓语都是边。

面向 RDF 模型的图数据库有 Virtuoso、Neptune 等。



图片来源：https://www.w3.org/

图 2-1 RDF 三元组实例

2.1.2 属性图

属性图是一个由顶点、边、以及顶点和边上的属性组成的图。顶点也称为节点（Node），边也称为关系（Relationship）。在属性图中，顶点和边是最重要的概念。以下我们将简要描述属性图的基本概念。

顶点（Vertex）是图中的实体。它们可以保存任意数量的属性（键-值对）。可以用标签标记顶点，表示它们在域中的不同角色。顶点标签还可以用于将元数据（例如索引或约束信息）附加到某一类顶点。

边（Edge）在两个顶点实体之间提供定向的、命名的、语义相关的连接。在有向图中，边由方向、类型、起始顶点和目标顶点组成。与顶点一样，边也可以具有属性。在大多数情况下，边具有定量属性，如权重、成本、距离、评级、时间间隔或强度。两个顶点可以添加任意数量或类型的关系，保证属性图模型的灵活性。

属性（Property）：顶点和边都可以有一个或多个属性，属性是一个键值对（Key/Value Pair），保存在顶点或边上。在实践中，一般每个顶点都会包含一个 id 或 name 属性作为主键。

标签（Label）：指示一组拥有相同属性类型的顶点，值一般不同。

路径（Path）：一组有顶点、边以链状首尾相连的集合叫做路径。

以图 2-2 为例。该图中从左向右有 3 个顶点，其标签分别为：Employee、Company、City，分别代表员工、公司、城市。Employee 顶点有 3 个用键值对表示的属性，表示该员工的姓名为 Amy Peters、出生日期为 1984-03-01、员工 ID 号为 1；从 Company 顶点有一条边指向 Employee 顶点，该条边的标签为 HAS_CEO，表示公司的 CEO 是 Amy Peters，同时该条边有一个属性表示了公司的创始时间为 2008-01-20；从 Company 顶点有一条边指向 City 顶点，该条边的标签为 LOCATED_IN，表示该公司位于该城市。

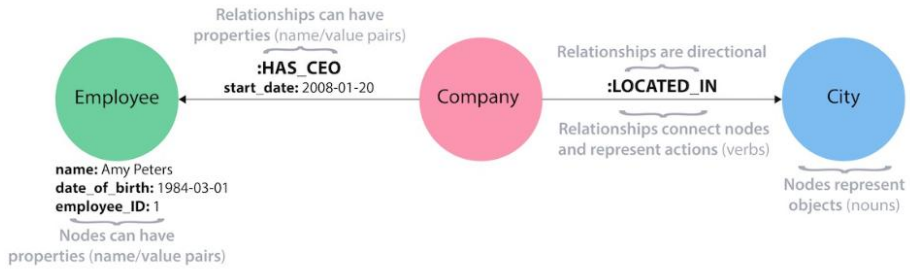


图 2-2 属性图实例[7]

主流的开源图数据库 Neo4j 和 JanusGraph 都采用属性图的数据模型。不同的是，Neo4j 使用原生设计的图存储，JanusGraph 使用非原生图存储，将图结构序列化存储到基于 BigTable Model 的键值对数据库（例如 Cassandra，HBase）中[8-9]。关于不同类型存储方法的比较，我们将在 2.2 节进行更具体地描述。

2.1.3 属性图与 RDF 模型的区别

属性图模型和 RDF 图模型这两种图模型的区别主要体现在：数据模型的结构、操作和约束这三个方面（表 2-1）。RDF 图模型的表达力强于属性图模型，是因为 RDF 的超图本质，一条三元组的谓语可在另一条三元组中做主语或宾语，但顶点和边上不能再定义属性[10]。总体说来，由于 RDF 图具有加强的逻辑理论背景，加之语义 Web 多年的标准化工作，其数据模型特性相对完善。属性图更

符合一般用户的直观，在顶点和边上直接定义属性，随着 Neo4j 等图数据库的应用，其获得了较强的用户认可度[11]。RDF 和属性图两种图模型都体现顶点和边的模型本质，在实践中可以相互转换，即 RDF 模型可以转为属性图模型，而属性图模型也可以转换为 RDF 模型。

表 2-1 RDF 图模型和属性图模型的区别

数据模型特性		RDF 图模型	属性图模型
结构	标准化程度	已由 W3C 制定了标准化的语法和语义[12]	尚未形成工业标准
	数学模型	三元组	有向标签属性图
	属性表达	通过额外方法，如“具体化”	内置支持
	概念层本体定义	RDFS[13]、OWL[14]	不支持
操作	查询代数	SPARQL 代数[15]	无
	查询语言	SPARQL[15]	Cypher[16]、Gremlin[17]、PGQL[18]、G-CORE[19]
约束	约束语言	RDF Shapes 约束语言（SHACL）	无

2.2 图数据存储

图数据库如何存储图，对存储效率和查询效率都至关重要。本节将对图数据常用的几种存储方式进行简要介绍。

2.2.1 链表

链表是一种在存储单元上非连续、非顺序的存储结构。数据元素的逻辑顺序是通过链表中的指针链接次序实现。链表是由一系列的结点组成，结点可以在运行时动态生成。每个结点包含两部分：数据域与指针域。数据域存储数据元素，指针域存储下一结点的指针。根据指针域是否连接了多个方向又可具体细分为单向链表、循环链表、双向链表等。

图数据库 Neo4j 是使用链表作为图数据存储结构的一个主要代表。每个顶点使用一个顶点记录来表示，每条边使用一个边记录来表示。如图 2-3 所示，每个顶点记录包括：（1）一个指向该点的第一条边的指针 nextEdgeID，（2）一个指向该点的属性的单向链表的指针 nextPropID，（3）点的标签 label，和（4）一些标志 flags；每个边记录包括（1）该条边所指向的两个顶点 firstVertex 与 second Vertex，（2）边的类型 relType，（3）该条边指向的两个顶点各自的边的双向邻

接表 firstPrev/NextEdgeID、secondPrev/NextEdgeID，（4）一个指向边的属性的单向链表的指针 nextPropID，以及（5）一些标志 flags。

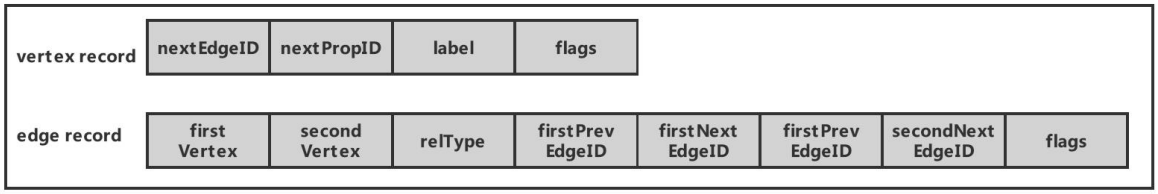


图 2-3 Neo4j 的顶点记录与边记录

下图展示了 Neo4j 对数据的物理存储模式：

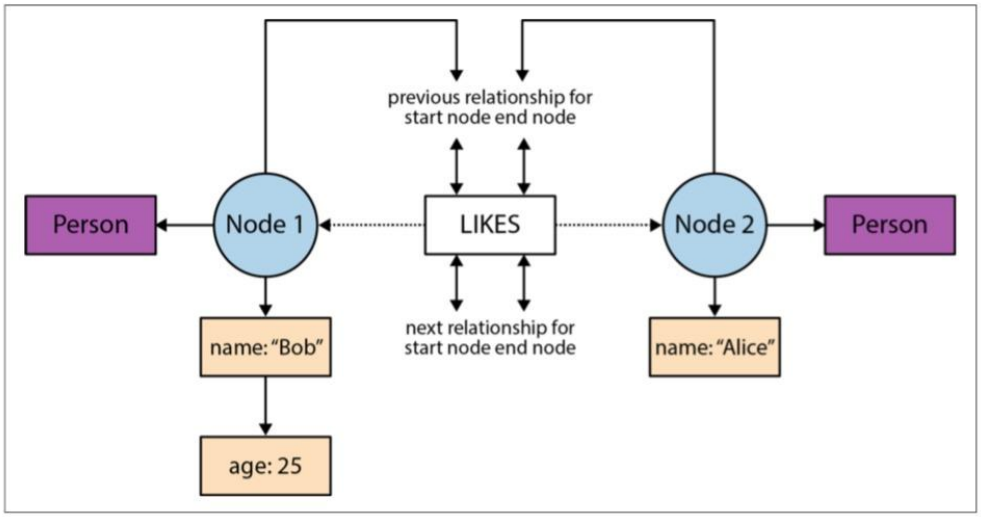


图 2-4 Neo4j 图数据库的物理存储模式

在图 2-4 中我们可以看到：Node1 和 Node2 两个顶点记录中，都有一个指针指向其标签 Person，都有一个指针指向对应的属性链表，同时还有一个指针指向了接连他们的一条边，该条边属于两个邻接表：Node1 的邻接表和 Node2 的邻接表，因此在该条边的边记录中，有两个指针指向了 Node1 的邻接表的前后两条边，同时有两个指针指向了 Node2 的邻接表的前后两条边。若要读取顶点的属性，我们从顶点指向的第一个属性开始遍历链表结构的属性。若要查找顶点的关系，我们从该顶点的边指针找到其第一条边（本例中的“LIKES”关系）。在这里，我们按照该顶点的双向邻接表进行遍历，直到找到我们感兴趣的关系（边）。找到所需边的记录后，我们可以使用边属性的单链表读取该边的属性，也可以检查边所连接的两个顶点的相关信息。

2.2.2 排序树

图数据库在设计和实现上继承并发展了关系型数据库的做法，例如使用 B+树、LSM 树等树形结构来存储图数据，并通过连接（Join）的方式来查询图数据。以 RDF-3X[20]、Virtuoso[21]为代表的众多 RDF 数据库就将主谓宾按照不同的排列顺序使用 B+树建立了不同的索引，并在不同查询中组合地使用不同的索引来获得最优的查询效率。

在树形结构的基础上，很多图数据库提出了各种压缩表示方法来减少空间开销，从而增加查询效率。例如，Sparksee[22]混合地使用了 B+树和 Bitmap（位图）。Sparksee 将点与边统称为 object，并使用唯一的 ID 来标识。对每一个属性名，都有一个对应的 B+树将点或边的 ID 映射到相应的属性值，相应地从每一个属性值到 ID 的映射则由 Bitmap 来维护。标签、点、边之间得相互联系也通过 B+树类似地进行映射。Sparksee 对每一个点存储了两份 Bitmap，对应该点的入边与出边。同时，Sparksee 使用两个 B+树来维护边到入点、出点的连接关系。Bitmap 的使用使 Sparksee 能支持一些 bit 级别的操作，如要获取同时拥有属性 A 与属性 B 的 ID 时，可以通过对两个属性的 Bitmap 进行 AND 的操作来进行筛选以加速处理。

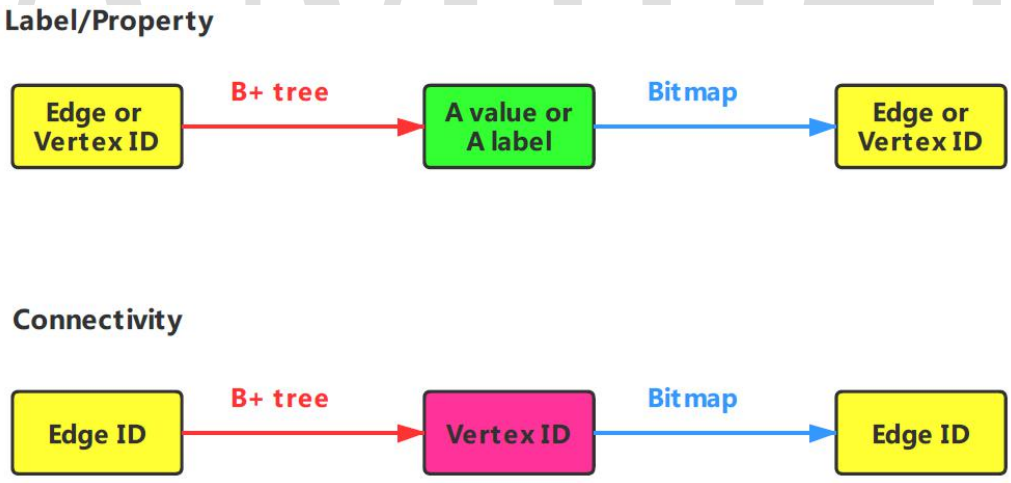


图 2-5 Sparksee[22]的映射关系

如图 2-5 所示，给定一个 ID，可以通过一个 B+树获取其标签值，根据一个标签值，可以通过 Bitmap 来获取所有属于该标签的 ID；给定一条边的 ID，可以通过两个 B+树分别获取其连接的入点与出点；给定一个点的 ID，通过其对应的两个 Bitmap，可以分别获取其连接的入边与出边。

2.2.3 哈希表

与树形结构类似，哈希表（Hash table，也叫散列表）也是一种常用的索引结构；不同之处在于，哈希表的查找更快（常数级复杂度），但是无法支持范围查询。

图数据库 ArangoDB[23]混合地使用了哈希表和链表来存储图数据，如图 2-6 所示。ArangoDB 使用 JSON 文档对点与边进行存储，每个文档有一个唯一 ID 为_key。点文档中只存储点的信息而不包含任何边相关的信息，因此在增加/删除边的时候不需要对点文档进行修改；边文档中存储了边的信息，其中每个边文档都有两个独特的属性_from 与_to，记录了与边相连的点文档的 ID。当需要查找一个点时，根据该点对应的_key，可以直接对 vertex index 进行哈希索引，找到该点所在的点文档。当需要查找某个点的边时，根据_from 与_to，对 edge index 进行哈希索引，可以找到相关的边文档链表。当需要遍历一个点的所有邻居点时，根据该点对应的_key，作为_from 从 edge index 中找到该点的邻边，然后根据邻边的边文档得到其对应的_to 属性，最后将_to 视为_key，从 vertex index 中寻找邻居顶点。

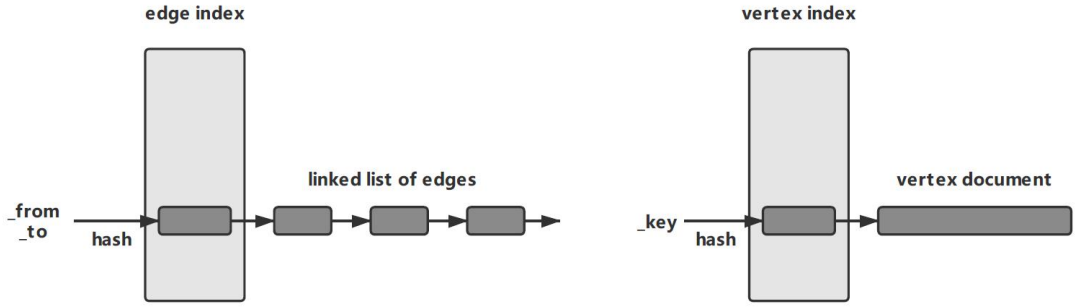


图 2-6 ArangoDB[23]的哈希索引

2.2.4 NoSQL 数据库

也有很多图数据库在存储上直接使用了 NoSQL 数据库，下面是几个典型的例子。

- **键值对存储**

键值对（Key-value, KV）存储数据库是用来存储、检索和管理关联数组的数据存储范式，是应用范围最广，也是涉及产品最多的一种 NoSQL 存储数据库[24]。

在数据被存储为键值对集合时，具有高性能和高伸缩性的优点，键值的确切形式取决于特定的系统或应用程序，它可以是简单的（例如，URI 或哈希），也可以是结构化的（值的结构通常是无模式的）。另外，也可以对键值存储施加一些额外的数据布局，构造无模式的值。

HyperGraphDB 的基本构建块是 atoms，以 KV 方式存储[25]。如图 2-7 显示了 HyperGraphDB 使用 KV 进行数据存储的示例，每个 atoms 都有一个强编码的 ID，以减少其冲突。HyperGraphDB 的顶点和超边都是 atoms，他们有唯一的 ID。每一个超边 atom 使用一个列表存储它所连接的所有顶点的 ID；顶点 atom 和超边 atom 存储了其类型 ID 和与之相关的值 atom 的 ID（如属性值）；值 atom 是一个可递归的结构，可以存储值 atom 的 ID 与二进制数据。

HyperGraphDB 维护了一个关联索引，将点 ID 映射到所有包含该点的超边，因此可以有效地从一个点遍历其超边。同时，通过类型索引可以获取某个类型 ID 的所有顶点与超边。最后，值索引可以快速找到包含了某个值 ID 的所有 atom。

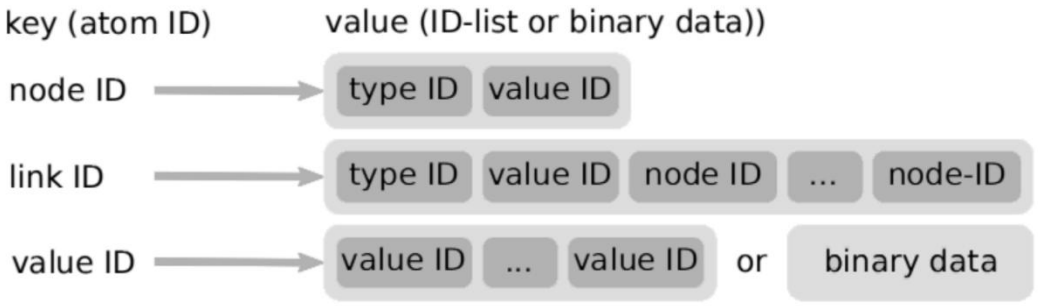


图 2-7 HyperGraphDB[25]的键值对存储图示

- **文档存储**

文档存储（Document stores）支持对结构化数据的访问，文档存储一般用类似 Json 的格式存储，存储内容为文档型，以封包键值对的方式进行存储。一般情况下，应用对要检索的封包采取一些约定，或者利用存储引擎将不同的文档划分成不同的集合，以管理数据。文档存储是基于键值对存储的，其存储结构比键值对更复杂。

与键值存储不同，文档存储关心文档的内部结构。这使得存储引擎可以直接支持辅助索引，从而允许对任意字段进行高效查询。支持文档嵌套存储的能力，使得查询语言具有搜索嵌套对象的能力，XQuery[26]就是一个例子。MongoDB[27]通过支持在查询中指定 JSON 字段路径实现类似的功能。

图 2-8 展示了 OrientDB[28]使用文档存储来表示顶点和边的方法。每个顶点对应的文档由顶点的属性、所有入边/出边对应的文档 ID、轻量级边对应的顶点文档 ID 这几部分组成；每条边对应的文档则由边的属性、出发/到达顶点的文档 ID 组成。

图 2-8 OrientDB[28]的文档存储图示

● 宽列存储

宽列存储（Wide-Column stores）结合了键值存储和关系表的不同特性，如图 2-9 所示。一方面，一个宽列存储将键映射到行，每一行可以有任意数量的单元格，每个单元格构成一个键值对；每行包含了单元键到单元值的映射，因此可以有效地实现二维键值对存储（行键与单元键）。另一方面，宽列存储的是一个表，因此单元键构成了列名，但与关系型数据库不同，宽列存储中同一列不同行的单元格中，列名与格式可能会有所不同。

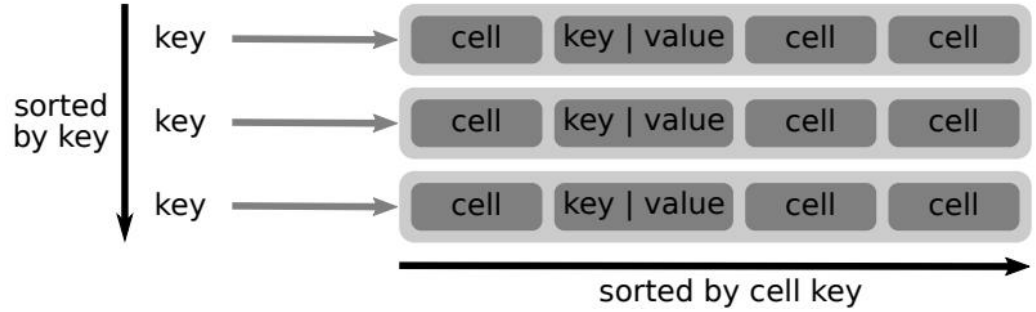


图 2-9 宽列存储示例[5]

图 2-10 是 Titan[29]和 JanusGraph[30]用宽列存储图数据的图示。每一行的键是顶点 ID，顶点的属性和与该顶点相关的出边/入边则存储在这一行的不同单元格中（通过不同的单元键予以区分）。

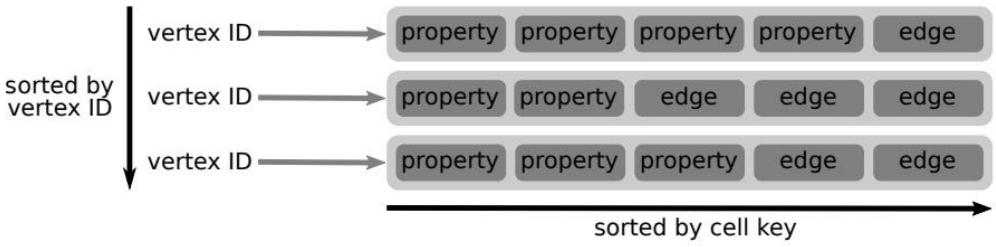


图 2-10 Titan[29]/JanusGraph[30]的宽列存储图示

2.3 图数据查询

按照查询范围的递增顺序，面向图数据的查询可以大体分成单点查询、邻居查询、路径遍历、子图匹配和全图分析这几类。

- 单点查询

单点查询涉及单个顶点或者边。例如查询一个点或者一条边的所有标签或者所有属性值、查询一个点或边的某个给定的属性、获取图中的所有标签、判断一个点或者一条边是否具有某个给定的标签。单点查询常用于社交网络工作负载[31-32]（例如，获取新浪微博某个用户的昵称、注册时间等信息）和一些基准测试[33]（例如，测量顶点查找时间与吞吐量）。

- 邻居查询

邻居查询表现为检索给定顶点的所有边、检索某条边连接的顶点；并且可以对邻居查询增加限制条件来获取更精细的结果，如查询具有给定标签的边。例如，查询新浪微博中某位大 V 的所有粉丝、查询某位用户所感兴趣的话题、查询某位用户的朋友，都可以对应到查询本地邻居[31-32]。

- 路径遍历

路径遍历查询可以探索邻居查询之外更大的图数据，路径遍历查询通常从单个顶点或者一个规模较小的顶点集合开始，遍历图的一部分数据，我们通常称遍历起始的顶点或顶点集合为根顶点或锚点。在查询时用户可以增加限制条件来规定需要检索与遍历的图数据。例如查询一个人的朋友的朋友的朋友（三度朋友）、查询一个人的三度朋友中喜欢打排球的人等。路径遍历查询是一个非常常用的图数据库查询任务，因此常用于性能基准测试[33-35]。

- 子图匹配

子图匹配要求从整个图中找出所有与给定的查询模板匹配的子图。子图的形状可能是一条链，也可能是一个环，亦或是更复杂的树和图。查询模板对应的子图可能包含较为具体的限制条件，例如一个具体的顶点；也可能是较为宽泛的限制条件，例如某类顶点，甚至没有任何限制条件。

- 全图分析

全图分析查询通常称为 OLAP（实时分析处理），这类查询通常会涉及整个图的查询（涉及所有顶点与所有边，但不一定涉及每个属性）。由于全图分析在不同的领域有着广泛地应用，如危险检测[36]、计算化学[36]等，因此不同的基准测试[35,37-38]通常会加入这一类查询。很多图数据库都会支持全图分析，图计算系统更是特别注重于解决全图分析查询，如 Pregel[39]、GraphX[40]、Gemini[41]等。全图分析常用于全局模式匹配[42-43]、最短路径[44]、最大流量或最小割[45]、最小生成树[46]、图直径、最远距离、连通分支、PageRank[47]等。一些对全图的遍历查询（如寻找所有不限长度的最短路径）属于全局分析查询的范畴。

可以看到，与关系模型以及其它 NoSQL 数据模型相比，基于图数据模型的查询具有更高的复杂度，这就对查询语言提出了更高的要求。图数据库尚且没有业界统一认可的查询语言。查询语言按照编写的逻辑，可以分为描述式和命令式两种。

描述式查询语言只表达要达到怎样的目标，而不关心底层系统通过怎样的逻辑来实现，对上层用户相对友好，SQL 就是典型的描述式查询语言。在图数据库领域，针对属性图的代表性描述式查询语言是 Cypher，由 Neo4j 首次提出，其开放版本为 OpenCypher[48]。针对 RDF 图的代表性描述式查询语言是 SPARQL，几乎所有支持 RDF 模型的数据库都会支持 SPARQL。

命令式查询语言是通过有序的一句命令来告诉图数据库如何去执行，通常需要对程序设计有一定的了解，执行的效率也较高。有些图数据库如 Neo4j、Virtuoso、TuGraph 等会开放较为底层的 API（使用 C++/Java 等或是厂商自己定义的程序设计语言），让用户通过实现类似关系型数据库中“存储过程”的方式，作为高层查询语言的实现无法尽善尽美时的辅助手段。Gremlin 是基于 Scala 的函数链式语言，可以归纳为高层命令式查询语言，但同时带有少量描述式语言的特性。

本节接下来的内容将围绕主流的图数据库查询语言展开，并在最后简单介绍一些图数据查询优化相关的技术。

表 2-2 图查询语言

图查询语言	易用性	性能	举例
描述式	强	一般	Cypher、SPARQL
命令式	弱	好	Gremlin

2.3.1 Cypher

Cypher 是 Neo4j 提出的图查询语言，它允许用户从图数据库中存储和检索数据。Neo4j 想让查询图数据变得易于学习、理解和使用，同时也融合了其他标准数据访问语言的强大功能，这同时也是 Cypher 的目标。

Cypher 语法提供了一种可视化的逻辑方式来匹配图中顶点和关系的模式。它是一种受 SQL 启发的声明性语言，用于使用 ASCII-Art 语法描述图中的可视模式。它允许我们声明想要从图数据库中选择、插入、更新或删除什么，而不需要精确地描述如何做到这一点。通过 Cypher，用户可以构建表达性强且高效的查询来处理所需的创建、读取、更新和删除功能。

OpenCypher 项目为 Cypher 提供了开放语言规范、技术兼容性工具包和解析器、规划器和运行时的参考实现。它由数据库行业的几家公司支持，允许数据库实现者和客户免费受益、使用和贡献 OpenCypher 语言的开发。Cypher 是一个描述性的图查询语言，语法简单，功能强大。和 SQL 很相似，Cypher 语言的关键字不区分大小写，但是属性值、标签、关系类型和变量是区分大小写的。

支持 Cypher 的图数据库包括 Neo4j[49]、RedisGraph[50]、AgensGraph[51]、TuGraph[52]等。

Cypher 的操作语句有：

MATCH：匹配图模式，是从图数据库中获取图信息的基本方式
WHERE：用于给图模式添加约束或者过滤
CREATE、DELETE：创建和删除顶点或者边
SET、REMOVE：使用 SET 设置属性值或给添加标签，使用 REMOVE 移除

以查询 Bob 的朋友的朋友为例，使用 Cypher 进行查询时，其查询语句为：

```
MATCH (person:Person)-[:knows*2]-(friend:Person)
WHERE person.name = 'Bob'
RETURN friend.name
```

查询的图模式为从 Bob 出发，查询其边属性为“knows”的二度邻居，即为 Bob 的朋友的朋友，返回其姓名即可。

使用 SQL 时需要嵌入查询：

```
SELECT friend_name
FROM friend
WHERE
    name in (select friend_name from friend where name=’Bob’)
```

首先需要从关系型数据库的 friend 表（列名为 name、friend_name）中查询 Bob 的朋友作为一个集合，然后再将 friend 表查询该集合中的每个人的朋友作为返回结果。

可见，对图数据库进行查询时，使用 Cypher 比 SQL 的表达能力更强、查询更高效、更容易理解与使用。

2.3.2 Gremlin

Gremlin 是 Apache ThinkerPop 框架下的图遍历语言。Gremlin 可以是声明性的也可以是命令性的。虽然 Gremlin 是基于 Groovy[53]的，但具有许多语言变体，允许开发人员以 Java、JavaScript、Python、Scala、Clojure 和 Groovy 等许多现代编程语言原生编写 Gremlin 查询。Gremlin 是图遍历语言，其执行机制是在图中沿着有向边进行导航式的游走。这种执行方式决定了用户使用 Gremlin 需要指明具体的导航步骤，所以 Gremlin 是过程式语言。与受到 SQL 影响的声明式语言 SPARQL 和 Cypher 不同，Gremlin 更像一种函数式的编程。

支持 Gremlin 的图数据库包括：JanusGraph[30]、InfiniteGraph[54]、CosmosDB[55]、DataStax Enterprise(5.0+)[56]、Amazon Neptune[57]。

同样以查询 Bob 的朋友的朋友为例，使用 Gremlin 进行查询时，其查询语句为：

```
g.V().has("name","Bob").out("knows").out("knows").values("name")
```

查询语句十分的简单，但是明确指出了查询时的每一个步骤，首先对图 g 中的顶点集合 V()，找到其中的“name”为“Bob”的顶点，随后通过两个 out("knows") 找到其朋友的朋友，最后通过返回 values("name")查询得到朋友的朋友的姓名。

2.3.3 SPARQL

RDF 是一种用于表示 Web 中的信息的定向、标记的图数据格式。而该规范定义了 RDF 的 SPARQL 查询语言的语法和语义。SPARQL 可用于表示不同数据源之间的查询，无论数据是作为 RDF 本机存储还是通过中间件作为 RDF 查看。SPARQL 包含查询所需和可选图模式及其连接和析取的功能。SPARQL 还支持通过源 RDF 图进行可扩展的值测试和约束查询。SPARQL 查询的结果可以是结果集，也可以是 RDF 图。

SPARQL 查询的基本单元是三元组模式，多个三元组模式可构成基本图模式(basic graph pattern)。SPARQL 支持多种运算符，将基本图模式扩展为复杂图模式(complex graph pattern)。

同样以查询 Bob 的朋友的朋友为例，使用 SPARQL 进行查询时，其查询语句为：

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
WHERE
{
   ?s foaf:name "Bob" .
   ?s foaf:knows ?fr .
   ?fr foaf:knows ?friend .
   ?friend foaf:name ?name
}
```

首先为<http://xmlns.com/foaf/0.1/>定义了一个前缀为 foaf，这样就不必每次都以完整的名字来引用。SPARQL 的查询语句中每一个条件都是一个（主 谓 宾）的三元组，其中变量以?为前缀，因此该查询中，需要满足的条件分别是：?s 的

名字为 Bob、?s 的朋友为?fr、?fr 的朋友为?friend、?friend 的名字为?name，最后返回?name 作为结果。

2.3.4 GQL

2019 年 6 月，隶属 ISO/IEC 联合技术委员会的全球诸多国家性标准机构开始就标准图查询语言 GQL 项目提案进行表决，有七个国家派出专家参与这项为期四年的项目，有望在 2023 年形成图查询语言的国际标准。

GQL 的全称是“图查询语言”，这种新语言将由监管 SQL 标准的同一个国际工作组开发和维护。GQL 高度依赖现有的语言。GQL 项目是自 SQL 之后的第一个 ISO/IEC 国际标准数据库语言项目。GQL 的主要灵感源自 Cypher（现在实现的版本有十多个，包括六款商业产品）、Oracle 的 PGQL 和 SQL 本身，以及用于只读属性图查询的 SQL 新扩展。

目前，GQL 标准将包括以下两到三个部分。通过引用 SQL/Framework 和 SQL/Foundation（ISO/IEC JTC1 9075:2016 第 1 部分和第 2 部分）的规范进行合并：

(1) 一些标量数据类型；

(2) 标量类型的操作、函数和谓词；

(3) 事务模型(隔离级别、提交、回滚等)；

(4) 安全模型；

(5) 客户端模型和会话模型

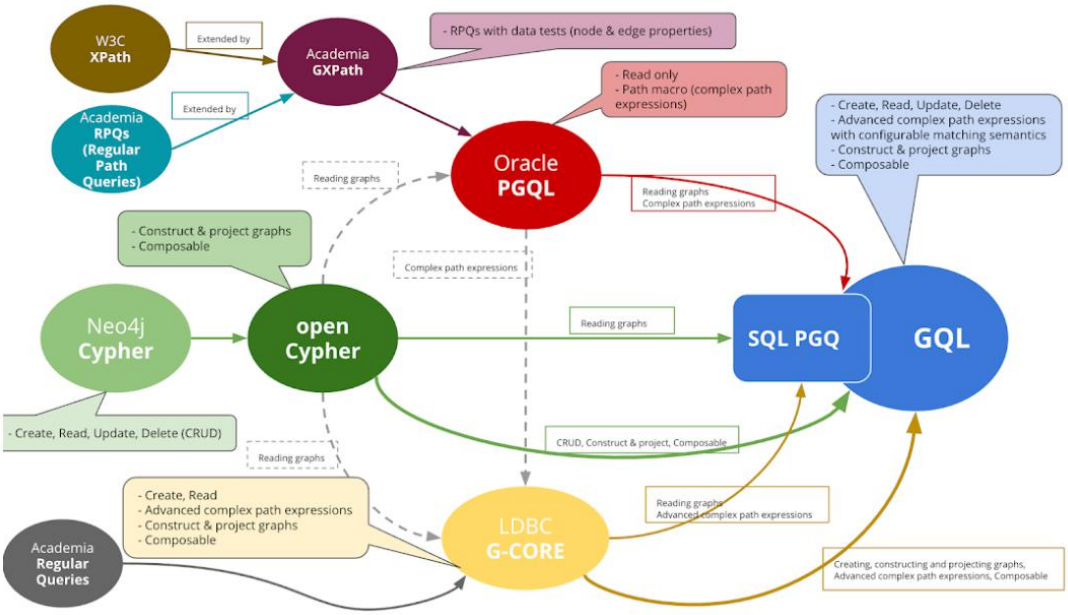
通过引用 SQL 标准中的现有规范，GQL 消除了重做大量详细基础工作的需要：

(1) SQL/PGQ (SQL 中的属性图查询)和 GQL 标准所需的功能；

- 最初集成到 SQL/PGQ:2020
- 随着 GQL 项目的进展分成单独的部分，由 GQL 和 SQL/PGQ 引用
- 图模式匹配

(2) GQL 特定功能

- 图数据类型：顶点/边缘/路径
- 不包含在 SQL:/PGQ:2020 中的图操作
- 图模拟 SQL/模式



注：各个图数据库厂商以及学术界推出的

图 2- 11 目前已有的面向图数据的查询语言示意图

2.3.5 其他查询语言

以下将对 PGQL（Oracle）、GSQL（TigerGraph[57]）、G-CORE（LDBC）等其他图数据查询语言进行简要介绍。

PGQL（属性图查询语言）是一种用于属性图数据模型的类似 SQL 的查询语言，它基于图模式匹配的范例，允许用户指定然后针对图中的顶点和边进行匹配的模式。与 SQL 一样，PGQL 支持分组（GROUP BY）、聚合（例如 MIN、MAX、AVG、SUM）、排序（ORDER BY）和许多其他熟悉的结构。此外，PGQL 还为可达性分析等应用程序提供了常规的路径表达式。本质上，PGQL 是一种图模式匹配查询语言。PGQL 查询描述由顶点和边组成的图模式。当根据属性图计算查询时，将返回与模式匹配的所有可能的子图。

GSQL 是一种用于进行图分析的高级且功能完备的查询语言。GSQL 具有类似 SQL 的语法，可以减少 SQL 程序员的学习成本，同时也提供 NoSQL 开发人

员首选的 MapReduce 用法，使用 MapReduce 的方式，可以实现大规模的并行计算。

G-CORE 是一种用于属性图数据库的图查询语言，G-CORE 是由 LDBC 图查询语言工作组设计的，该工作组由来自产业界和学术界的成员组成，旨在为图从业者带来这两个领域的精华。核心操作为图模式（JOF）、路径模式（RPQs）、聚合、子查询（Exists+SetOp）、图和路径构造。

2.3.6 查询优化

- 遍历顺序

在进行较为复杂的查询时，图数据库需要能够自动根据进行的查询选择出较优的遍历顺序，从而减少需要的数据访问量：对于采用了关系型数据库存储和查询方法的系统而言，这个目标对应于选择出较优的 Join 顺序[42-43]；对于很多使用原生图存储的系统而言，这个目标就是如何找到较优的遍历顺序。

- 辅助索引

为了选择出较优的遍历顺序，图数据库需要能够估计出符合条件的顶点/边有多少，从而估算出不同顺序的代价。为了达成这一点，通常图数据库需要建立一些具有统计性质的索引作为辅助。

除此之外，图数据库也可以针对特定查询建立专门的辅助索引。例如，对于一种特定的子图匹配，我们可以预先计算得到所有符合条件的子图，并在后续图的结构更新时同步地更新索引，从而缩短查询时间。

- 通信优化

对于分布式的图数据库，由于网络较 CPU 和内存通常更可能成为查询的瓶颈，因此通信相关的优化是其中重中之重。一些图数据库通过使用异步的查询处理模型，或是使用硬件的能力（如 RDMA）等，来增强网络的使用效率；一些系统则尝试根据实际负载的特点，动态地在不同节点之间迁移并且复制一些热门的数据来减少查询需要的通信。

2.4 常见图数据库

本节搜集整理了图数据库领域的若干资源，并对数据库的一些基本属性进行对比。

2.4.1 Neo4j

Neo4j 是一个流行的图数据库，它是开源的。最近，Neo4j 的社区版已经由遵循 AGPL 许可协议转向了遵循 GPL 许可协议[49]。尽管如此，Neo4j 的企业版依然使用 AGPL 许可，Neo4j 基于 Java 实现，兼容 ACID 特性，其主要优势是生态相对完善。

2.4.2 ArangoDB

ArangoDB 是由 ArangoDB GmbH 开发的一种免费的开源本机多模型数据库系统。数据库系统通过一个数据库核心和统一的查询语言 AQL（ArangoDB 查询语言）支持三种数据模型，兼有键/值对、图和文档数据模型，提供了涵盖三种数据模型的统一的数据库查询语言，并允许在单个查询中混合使用三种模型。基于其本地集成多模型特性，可以搭建高性能程序，并且这三种数据模型均支持水平扩展。

2.4.3 Virtuoso

Virtuoso Universal Server 是一个中间件和数据库引擎的混合体，它将传统关系数据库管理系统（RDBMS）、对象关系数据库（ORDBMS）、虚拟数据库、RDF、XML、自由文本、Web 应用服务器和文件服务器功能结合在一个系统中[21]。Virtuoso 不是为上述每个功能领域都提供专用服务器，而是一个“通用服务器”；它支持实现多个协议的单线程服务器进程。免费开源版本的 Virtuoso Universal Server 也称为 OpenLink Virtuoso。

2.4.4 Neptune

Amazon Neptune 是 Amazon.com 托管的图数据库产品[57]。它用作 Web 服务，并且它于 2017 年 11 月 29 日宣布是 Amazon Web Services 的一部分。Amazon

Neptune 支持流行的图模型属性图和 W3C 的 RDF，以及它们各自的查询语言 Apache Tinker Pop、Gremlin 和 SPARQL，包括其他 Amazon Web Services 产品。

2.4.5 JanusGraph

JanusGraph 是一个可扩展的图数据库，可以把包含数千亿个顶点和边的图存储在多机集群上[30]。它支持事务，支持数千用户实时、并发访问存储在其中的图。该数据库是 2016 年 12 月 27 日从 Titan fork 出来的一个分支，之后 TiTan 的开发团队在 2017 年陆续发了 0.1.0rc1、0.1.0rc2、0.1.1、0.2.0 等四个版本。

Janus Graph 项目启动的初衷是“通过为其增加新功能、改善性能和扩展性、增加后端存储系统来增强分布式图系统的功能，从而振兴分布式图系统的开发”，JanusGraph 从 Apahce Tinker Pop 中吸收了对属性图模型（Property Graph Model）的支持和对属性图模型进行遍历的 Gremlin 遍历语言。

2.4.6 TigerGraph

2012 年，TigerGraph 在硅谷成立，由华人科学家许昱博士创立，深耕大数据图分析领域。TigerGraph 是一款“实时原生并行图数据库”，既可以部署在云端也可以部署在本地，支持垂直扩展和水平扩展，可以对集群中的图数据自动分区，遵循 ACID 标准，并且提供了内置的数据压缩功能。它使用了一种消息传递架构，这种架构具备了可随数据增长而伸缩的并行性。

2.4.7 TuGraph

TuGraph 由清华大学团队于 2016 年开发，属于国内自主研发的商业图数据库，采用 C++自底而上做了完整的设计。TuGraph 设计的理念是性能优先，支持大数据量和高吞吐率，同时支持高效的在线事务处理（OLTP）和在线分析处理（OLAP）。通过三年的迭代，TuGraph 在 ACID 事务支持、Cypher 查询语言、可视化交互、多数据源导入、在线热备、用户审计等方面均达到了很高的可用性。

2.4.8 常见图数据库对比

图数据库对比信息如表 2-3 所示。希望能对读者朋友更好地了解图数据库有所帮助，同时也欢迎读者补充，如有好的意见或建议还请与 AMiner 编者联系，或者登录 https://www.aminer.cn/获取更多资料。

表 2-3 常见图数据库对比

数据库	版本号	许可证	语言	说明
Amazon Neptune	5.1 (2018.9)	专有	未公开	Amazon Neptune 是一个完全由亚马逊网站管理的图数据库。支持图模型属性图和 W3C 的 RDF，以及查询语言 Apache TinkerPop Gremlin 和 SPARQL。
ArangoDB	3.3.11 (2018.6.28)	专有	C++, JavaScript, .NET, Java, Python, Node.js, PHP, Scala, Go, Ruby, Elixir	ArangoDB 是由 ArangoDB 公司开发的 NoSQL 原生多模型数据库系统。该数据库系统支持三种重要的数据模型（键/值、文档、图），包括一个数据库核心和统一的查询语言 AQL（ArangoDB 查询语言）。
JanusGraph	0.5.1 (2020.3.25)	Apache 2	Java	开源，可扩展，分布在 Linux 基金会下的多机集群图数据库中；支持各种存储后端（Cassandra，Apache HBase，Google Cloud Bigtable，Oracle berkaleydb）；通过与大数据平台（Apache Spark，Apache gigig，Apache Hadoop）集成，支持全球图数据分析，报告和 ETL；通过外部索引存储（Elasticsearch，Apache Solr，Apache Lucene）支持地理，数字范围和全文搜索。
Neo4j	4.0.3 (2020.3)	GPLv3 社区版；企业和高级版是商业和 AGPLv3 版	Java, .NET, JavaScript, Python, Go, Ruby, PHP, R, Erlang/Elixir, C/C++, Clojure, Perl, Haskell	开放源码，支持 ACID，具有企业部署的高可用性集群，以及基于 Web 的管理，包括完整的事务支持和可视化顶点链接图浏览器；可以通过大多数编程语言使用其内置的 REST Web API 接口访问，以及专有的带有官方驱动程序的 Bolt 协议。
OpenLink Virtuoso	8.2 (2018.10)	开源版本是 GPLv2；企业版本是专有的	C, C++	支持 SQL 和 SPARQL 的多模型（混合）关联式资料库管理系统（RDBMS），用于对 SQL 表和/或 RDF 图所建模的数据进行声明性（数据定义和数据操作）操作。还支持索引 RDF-Turtle、RDF-N-Triples、RDF-XML、JSON-LD，以及从包括 CSV、XML 和 JSON 在内的许多文档类型中映射和生成关系(SQL 表或 RDF 图)。可以部署为本地或嵌入式实例（在 NEPOMUK 语义桌面中使用）、单实例网络服务器或无共享弹性集群多实例网络服务器。
TigerGraph	2.4 （2019.4）	未开源，开发者版支持单机单用户单图非商业免费	C++	TigerGraph 是一款“实时原生并行图数据库”，既可以部署在云端也可以部署在本地，支持垂直扩展和水平扩展，可以对集群中的图数据自动分区，遵循 ACID 标准，并且提供了内置的数据压缩功能。它使用了一种消息传递架构，这种架构具备了可随数据增长而伸缩的并行性。

数据库	版本号	许可证	语言	说明
TuGraph	1.10.0 (2020.3)	商业闭源	C++	TuGraph 是性能优先的国产自主研发的图数据库，主要特点是单机大数据量，高吞吐率，以及灵活的 API，同时支持高效的在线事务处理（OLTP）和在线分析处理（OLAP）。

3 产业应用篇

图数据库的应用原理是查询和分析连接数据，对海量数据建立关联，并通过多样及快速的方法对数据进行分析与挖掘。此外，与其他类型数据库相比，图数据库的操作更为便捷、数据更加直观、存储模式灵活、应用场景丰富，是未来处理复杂数据关系的技术趋势。

本篇从实际用例（Use Case）和解决方案（Solution）出发，以数据的关联特征与问题的相似性为基础进行归类和展开，挑选其中 9 个典型的图数据库应用场景进行介绍，具体场景如图 3-1 所示。图中所介绍的场景均具有一定的通用性，且不局限于某个特点的行业。

图 3-1 图数据库应用场景

● 反欺诈

欺诈无处不在，我们经常会听到“金融欺诈”“电商欺诈”“电信欺诈”“医保欺诈”等，都可能造成巨额的经济损失，甚至威胁到社会安全。传统的反欺诈方案，主要是针对独立的业务实体进行分析。而当今的欺诈者的手段也越来越复杂和隐蔽，他们会很耐心地，长时间地通过伪造众多身份，制造虚假交易来维护一个的欺诈网络，伺机作案。

反欺诈的挑战主要体现在数据量大，数据分散，无法做到实时分析。利用图数据库技术可以将分散的数据建立联系，高效地处理分析海量数据，并及时反馈

分析结果。反欺诈场景适用的行业和部门有金融，保险，电信，医疗，公共安全，情报等。

为了维护医保安全，遏制“医患合谋”套利，相关单位需要通过数据分析来识别医保欺诈行为和医保欺诈团伙。固定的一群病人不定期、频繁地让一个医生进行诊疗，就可能为疑似骗保事件。下面是一个图数据库在“反医保欺诈”的应用案例（图 3-2）。

通过图 3-2 方案，对医院目录，药品目录，诊疗目录，交易明细等数据进行建模和分析，就可以做到对疑似骗保行为和疑似医保欺诈团伙的有效识别和预警。

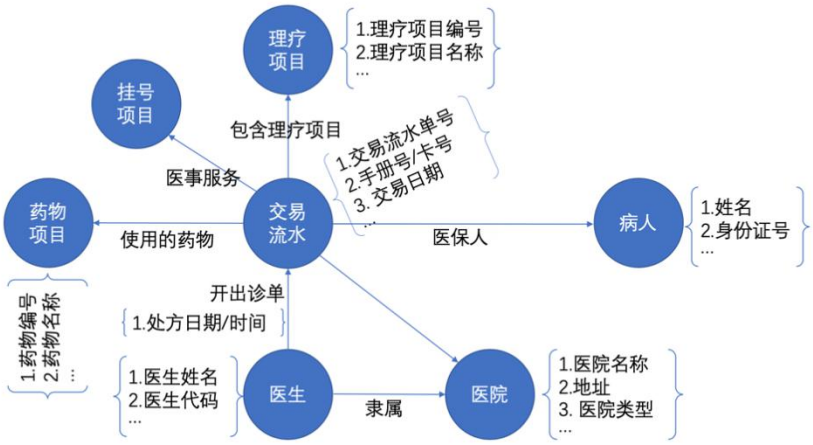


图 3-2 反医保欺诈方案的图数据建模示意图

● 推荐引擎

推荐引擎是电子商务平台在激烈竞争环境中的制胜法宝，精准及时的推荐，需要将商品、客户、库存、供应商、物流和网络舆情等数据有效地关联在一起。传统的推荐引擎是静态地针对孤立的历史数据进行离线分析，数据往往滞后一天，无法做到精细地建模。

图数据库在实时推荐引擎方案上的优势具体表现为：

(1) 整合复杂多源数据，来自商品，客户，库存，供应链等；

(2) 深链接分析，可以完成 3 到 10 跳（hop，可理解为层或度）联系的遍历和查询；

(3) 实时响应。

推荐引擎适用的行业和部门有零售、餐饮、广告、媒体出版及影视等。

以下是某图数据库供应商构建深链接推荐引擎（图 3-3）和实时推荐引擎（图 3-4）的方案，以供参考[57]。

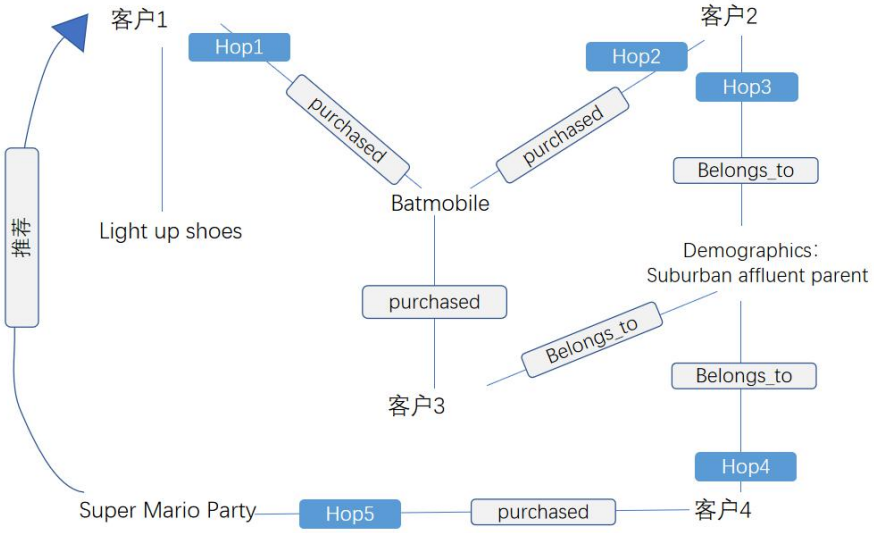


图 3-3 图数据库深链接推荐引擎方案示意图[57]

图 3-4 图数据库实时推荐引擎方案示意图[57]

● 知识图谱

知识图谱（Knowledge Graph, KG）的概念最早由 Google 提出。它的本质是一个图结构的语义网络，顶点是实体或概念，边是这些概念间的语义关系。它获取信息并将其集中到一个本体（Ontology）中，本体可以是人、概念、组织等。并应用推理器来推导新知识。它也可以被看成一个数据库，可以提高搜索引擎基于语义的数据的查询结果（图 3-5）。

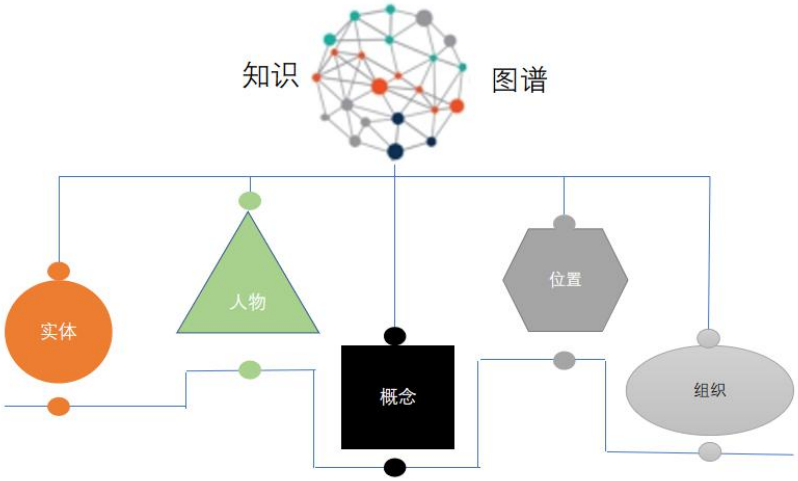


图 3-5 知识图谱将数据中的信息提炼并集中到一个实体中

大量的企业和组织已经开始使用知识图谱。主要的目的是从各种数据孤岛中获取大量数据并为其增值，以更有意义和更智能的方式使用。

和传统的基于关键词的搜索相比，基于图的搜索（Graph-based Search）的优势：

(1) 搜索结果更精准和更丰富；

(2) 搜索的速度更快，更实时；

(3) 搜索体验像在进行语言交流，更加智能。

知识图谱适用的行业有能源，机械制造，教育培训，政府机关，咨询等。

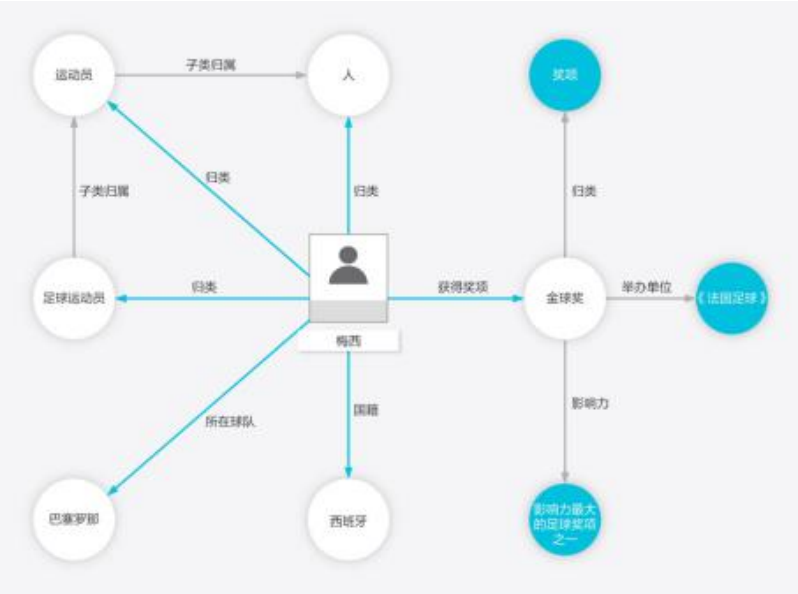


图 3-6 图数据库快速建立知识图谱实例

图 3-6 描述的是用图数据库建立的针对足球明星的知识图谱实例。案例中，用户可以方便、高效地查询感兴趣的信息。例如，哪位巴塞罗那的球员曾经获得“法国足球”举办的赛事的奖项？

● 身份和访问管理

身份和访问管理（IAM）的目的是制定一系列规则来管理，某人（例如管理员、销售、客服、用户）可以访问某种资源（例如文件，合同，网络设备，库存）。

对于大型企业和组织来说，存在以下问题和挑战：

(1) 人员识别和访问授权的数据是高度互联和复杂的；

(2) 人员，部门，规则和资源增长。权限查询的性能下降，无法满足正常的客户体验；

(3) 动态变化的组织结构，业务规则和外部环境。

图数据库既可以存储复杂，密集关联的访问控制结构和数据，支持等级结构（Hierarchical）和图结构的数据模型，还可以方便地支持自底向上的权限访问控制，例如，给定特定的访问资源（文件，服务器等），可以方便设置谁可以访问，谁有权利改变设置等。

身份和访问管理场景适用的行业和部门有政府机关，军事，公共安全，大型企事业，科研机构，医疗机构，高校等。

下图是北欧和亚洲手机运营商 Telenor 的资源访问管理数据模型图，该方案成功地帮助 Telenor 提升了身份和访问管理。

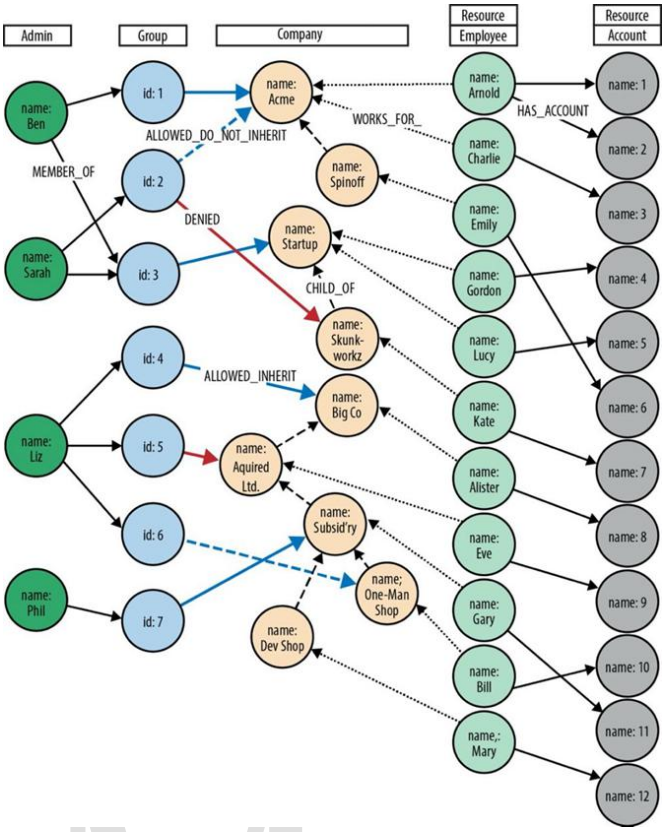


图 3-7 Telenor 的资源访问管理数据模型图

● 主数据管理

主数据（Master Data）是企业多系统共享的，描述核心业务实体的数据，例如客户，供应商，账户和组织部门等数据。和交易数据相比，主数据变化缓慢。主数据的互联性和共享性，为企业的商业分析提供了机会和优势。

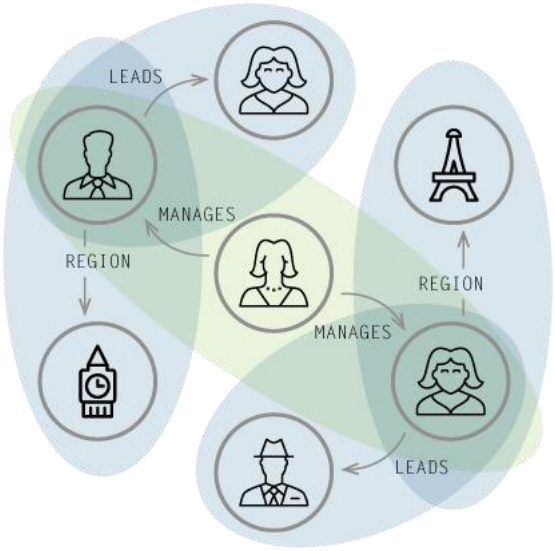


图 3-8 主数据示例图

主数据管理的挑战：

(1) 复杂的层级结构的数据集，查询成本高，效率低；

(2) 实时存储和查询困难；

(3) 动态变化的结构难以管理。

图数据库可以有效解决主数据管理问题，具体表现在：

(1) 主数据建模更容易；

(2) 相比关系数据库的方案，大大减少人力资源（建模师，架构师，DBA 等）；

(3) 图模型方便整合多源数据，例如来自 CRM，库存系统，财务系统，销售系统的数据。

主数据管理场景适用的行业和部门有能源，机械制造，航空航天，化工，大型企业等。下面的例子，描述一个企业的组织架构管理。

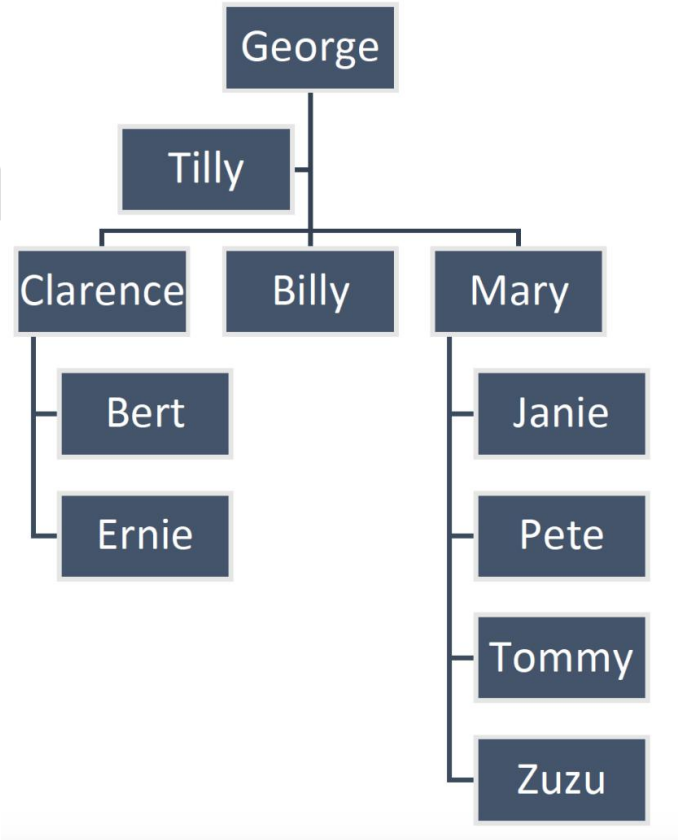


图 3-9 主数据层级图，描述人员的汇报和管理关系

随着人员的增长，用关系型数据库来查询和维护会变得越来越复杂和昂贵。例如，当有员工得到了提升，所有的关系都要进行重置和调整。而且现实世界的人员架构，不是简单的层级架构（图 3-9），实际更接近于图 3-10 的表现方式，是一个典型的企业组织架构图，描述的是人员的汇报和管理关系。

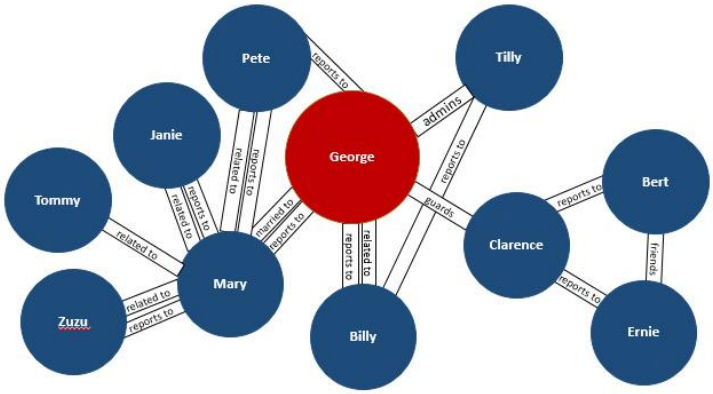


图 3-10 现实世界的人员汇报和管理关系

● 网络和 IT 设备管理

无论你的业务处在防火墙的哪一边，图数据库都是设计，存储和查询网络，IT 设备数据的很好选择之一。

网络和 IT 设备管理的挑战体现在：

(1) 复杂，紧急网络事故的定位和抢修；

(2) 设计缺陷，潜在风险的分析；

(3) 快速增长的物理，虚拟顶点数，制度规则升级等对管理的挑战。

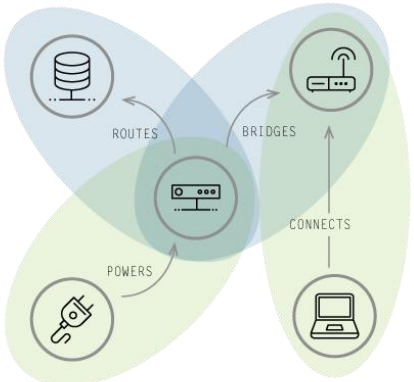


图 3-11 最能直观地表示网络和 IT 设备的拓扑结构

网络本身就是最自然和直观的图结构。图数据库可以用来：

(1) 存储网络 IT 设备的配置信息，例如 IP，端口，路由等；

(2) 帮助日常实时报警；

(3) 分析网络设计的潜在缺陷；

(4) 分析网络运营故障模式。

网络和 IT 设备管理场景适用的行业和部门具体有互联网，云计算，信息化充分且具有大量的网络和 IT 设备需要维护的单位和部门。

图 3-12 直观地展现费马科技为某大型互联网企业设计的网络设备拓扑和报警管理的应用方案。

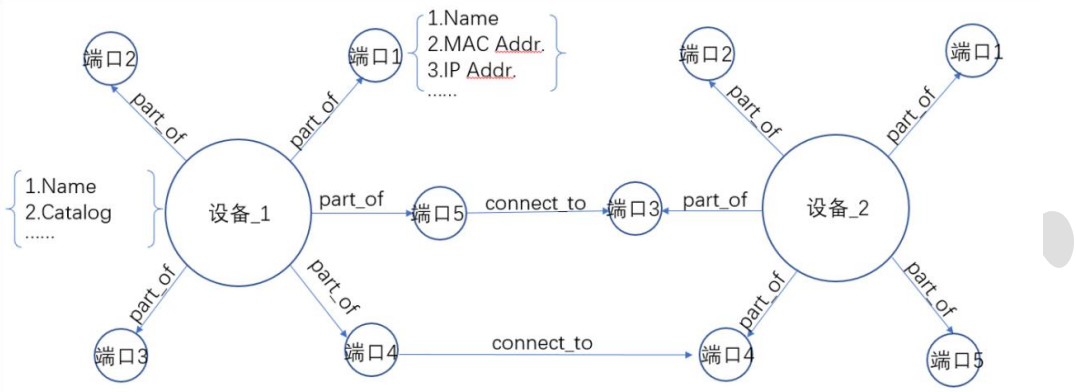


图 3-12 某企业网络设备拓扑和报警管理应用方案的示意图

● 地理空间分析

很多企业的管理者希望拥有和地理空间数据相关的产品和应用。地理空间数据可以帮助企业发现新的市场和收入来源，实时地监控和避免欺诈行为。企业还希望以尽可能小的代价来重构数据模型和系统架构，把地理空间数据接入已有的或未来的应用中，同时实现在线交易（OLTP）和在线分析（OLAP）任务。

传统的关系型数据库在处理地理空间分析的挑战：

(1) 数据模型不利用位置数据的实时分析和查询；

(2) 地理空间数据分析依赖于第三方索引；

(3) 地理空间数据的 SQL 查询十分复杂。

图数据库的建模理念就是将周围环境的人和事物构建成相互联系的实体，并深入挖掘实体之间的关系。地理空间数据是对图数据建模的最直观补充。图数据库是处理地理空间数据最自然的选择。

地理空间分析场景适用的行业和部门有交通运输、旅游、气象、采矿、地质、水利、新能源（风能，太阳能）等。以下两个例子，都是用图数据库管理地理空间数据，实现移动电商推荐和出租车实时定位功能。

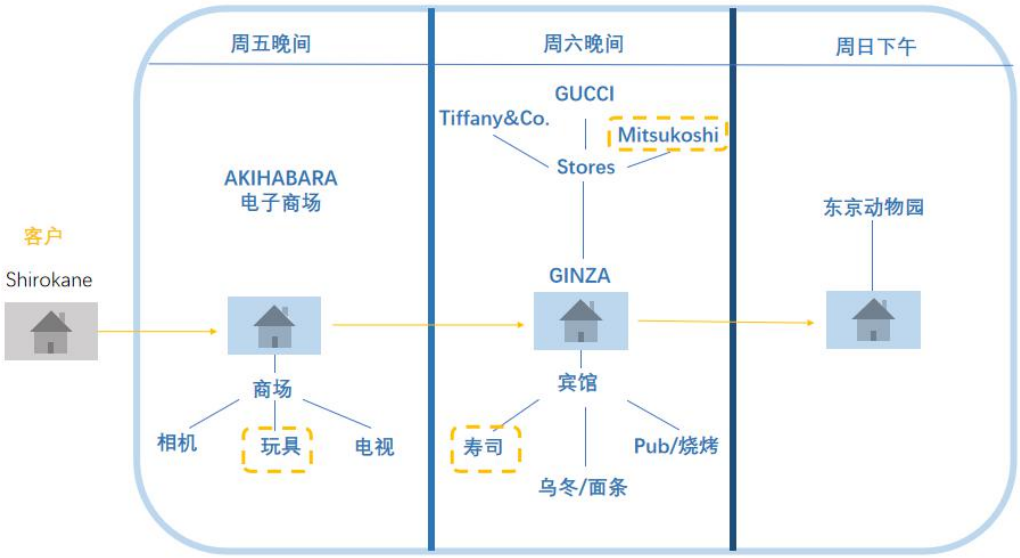


图 3-13 客户的地理空间数据分析在移动商业推荐上的应用示例

图 3-14 出租车实时定位

● 时序数据分析

时序数据，是一系列由时间索引的数据点。金融交易、库存变化、服务器日志、IoT 传感器采集等都是这样的数据类型。随着互联网数据和 IoT 设备数据雪崩式地增长，时序数据的分析和挖掘成为企业面临的新的挑战。

时序数据分析场景适用的行业和部门有工业制造、能源环保、医疗健康及 IoT 等。

图数据库企业和专家在如何有效地对时序数据进行分析上，也进行着创新和探索，并取得一定成绩。下面，我们看一个图数据库在电网 IoT 时序数据上的应用例子（图 3-15）。

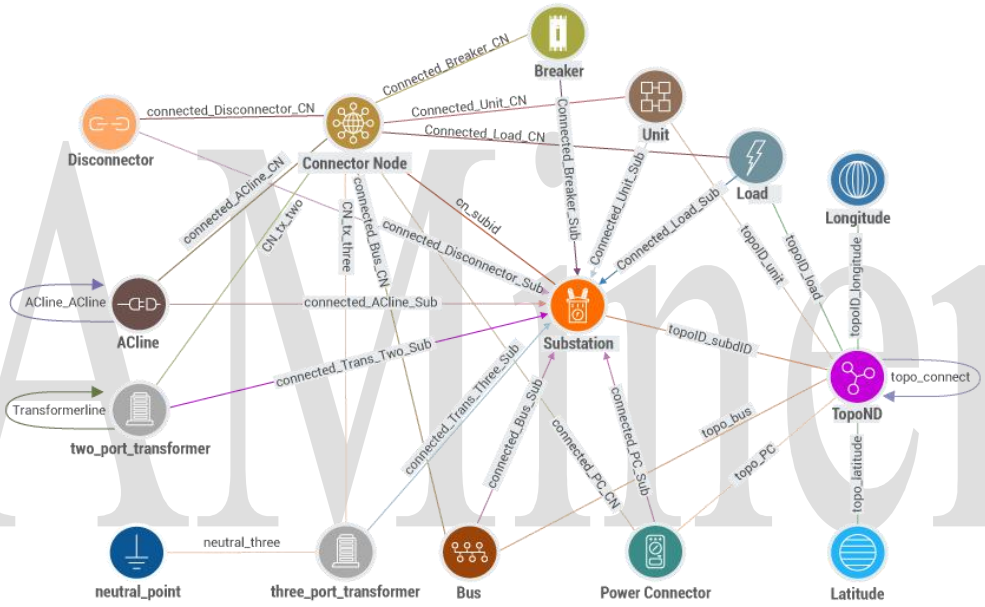


图 3-15 电网 IoT 传感器的时序数据图模型示例

上图中，能源电力企业部署大量的设备，无数的传感器，时刻不停地产生着海量的时序数据。企业需要监控和分析各个设备的电流量，查找瓶颈，及时上报故障和问题。要想满足电网的平衡，企业需要收集各个环节的基础设备的信号，并对发电、用电和输电进行统一管理。这时候需要实时深链接分析（Real-time Deep Link Analysis）。深链接分析指的是 3 到 10 跳（Hop，可理解为层或度）顶点的遍历和查询能力。图数据库在这一场景具备天然的优势。

● 社交网络

社交网络天然具备图数据结构，无需再为关系型数据建立图模型。无论是针对已经声明的社交关系（例如微信朋友圈），还是根据行为推理出潜在社交关系，图数据库都可以很好地为企业的创新社交网络应用做出贡献。

社交网络场景适用的行业和部门有互联网、广告、媒体出版、零售、公共安全等。这里我们举一个国际事件来展示图数据库在社交网络上的应用。

有报道称，某黑客团体，利用在社交网络工具推特（Twitter）上发推文（Tweets）的方式操作美国 2016 年的大选。在美国政府介入调查之际，大量的推文和账号已经被删除，如何恢复和分析数据？如何找到黑客团队的运作方式？他们如何渗透到日常美国人的在线对话中并试图影响公众舆论？图数据库就成为回答这些问题的有力工具。美国新闻媒体 NBC News 就使用图数据库供应商 Neo4j 的产品和技术，得出结论。

其中某图数据库供应商的方案体现为：

(1) 构建图模型后，展示了推文、用户（有些已经被识别为黑客账号）、主题、标签、源应用程序和链接之类实体之间的关系；

(2) 中心度算法用来测量顶点的中心度；

(3) 社区发现算法揭示了频繁交互的用户网络；

(4) 发现最具影响力的潜在黑客；

(5) PageRank 算法确定在群中最有影响力的账号。

黑客还会留下其他的线索，例如一般的用户会用移动设备发推文，而黑客一般会用 Web 客户端。根据推文数量暴涨的时间规律，也可以发现黑客所在的时区信息。

4 人才篇

在大数据时代，图数据库技术不断迭代更新，覆盖人群和应用场景逐渐扩大，在图数据库道路上的众多学者专家们也在不断探索与研究。本篇将对领域学者的分布情况和代表性学者进行简要介绍。

图数据库领域学者筛选的具体方法如下：首先，通过 AMiner 大数据平台挖掘图数据库领域学术会议及期刊：数据管理国际会议（The ACM Special Interest Group on Management of Data, SIGMOD）、超大型数据库国际会议（International Conference on Very Large Databases, VLDB）、IEEE 国际数据工程会议（IEEE International Conference on Data Engineering, ICDE）、图形数据管理经验与系统国际研讨会（International Workshop on Graph Data Management Experiences & Systems, GRADES）、扩展数据库技术国际会议（International Conference on Extending Database Technology, EDBT）的近 10 年论文，提取论文中所有学者信息，以此分析学者的分布情况。然后从中选出与图数据库领域关键词相关度最高的 2,000 位领域活跃学者，再按照学者的 h-index 进行排序，最后对其中排名靠前的部分学者进行简要介绍。领域关键词由图数据库顾问组给出，具体包括：图数据库（Graph databases）、属性图（Property graphs）、资源描述框架（Resource Description Framework, RDF）、图分析（Graph analysis）、ACID 事务属性（Atomicity, Consistency, Isolation, Durability, ACID transaction）、图匹配（Graph patterns）。

4.1 学者情况概览

4.1.1 全球学者概况

- **学者地图**

学者分布地图对于进行学者调查、分析各地区竞争力现况尤为重要，图 4-1 为图数据库领域全球顶尖学者分布情况。其中，颜色越趋近于红色，表示学者越集中；颜色越趋近于绿色，表示学者越稀少。从地区角度来看，欧洲、北美洲、东亚是图数据库领域学者分布最为集中，从国家角度来看，前十国家的学术数量如图 4-2 所示。

图 4-1 图数据库全球顶尖学者分布

- **国家对比**

根据 AMiner 平台数据分析不同国家在“图数据库”领域的技术发展差距，具体分析方法为根据论文作者的国家信息，将论文分类到各个国家中，从而统计出每个国家的论文发表数量、人才数量以及论文引用数等。图 4-2 展示了该领域前 10 个国家论文发表数量和人才数量的总体对比情况。美国的论文数量和人才数量位于全球第一，遥遥领先于排位第二的中国，随后为德国、英国、法国等欧洲国家。

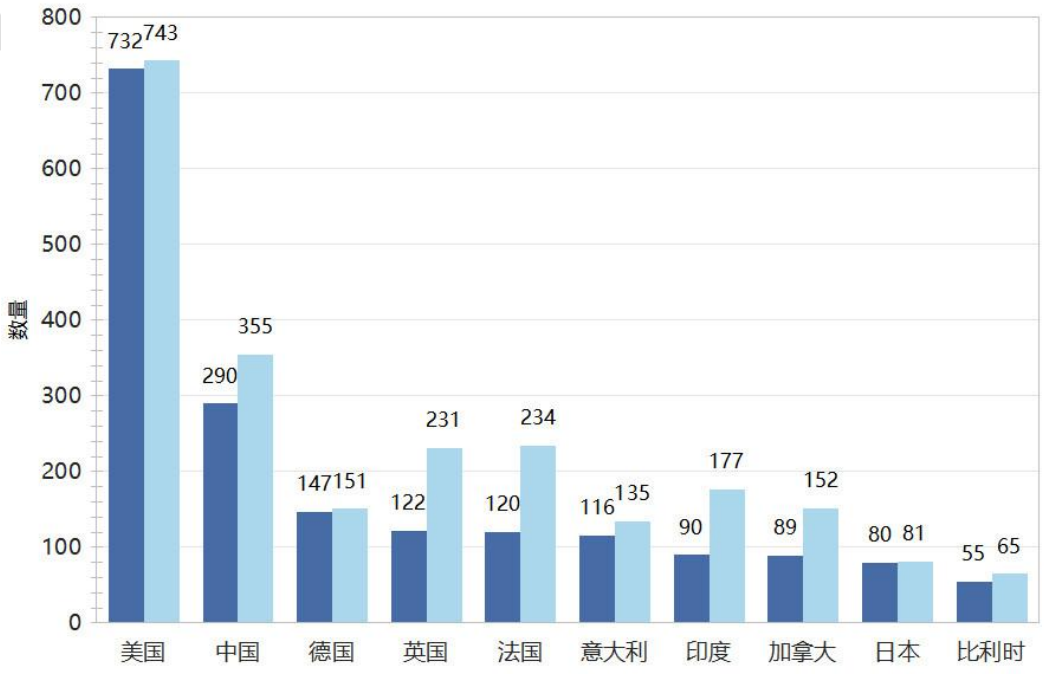


图 4-2 图数据库领域 Top 10 国家论文发表数量和人才数量对比

- **学者 h-index 分布**

图数据库学者的 h-index 分布如下图所示，大部分学者的 h-index 都在 10 以下，其中 h-index 小于 10 的人数最多，有 588 人，占总学者数量的 59.51%（图 4-3）。由此可见，在图数据库领域，世界级科研领军人物极度稀缺。

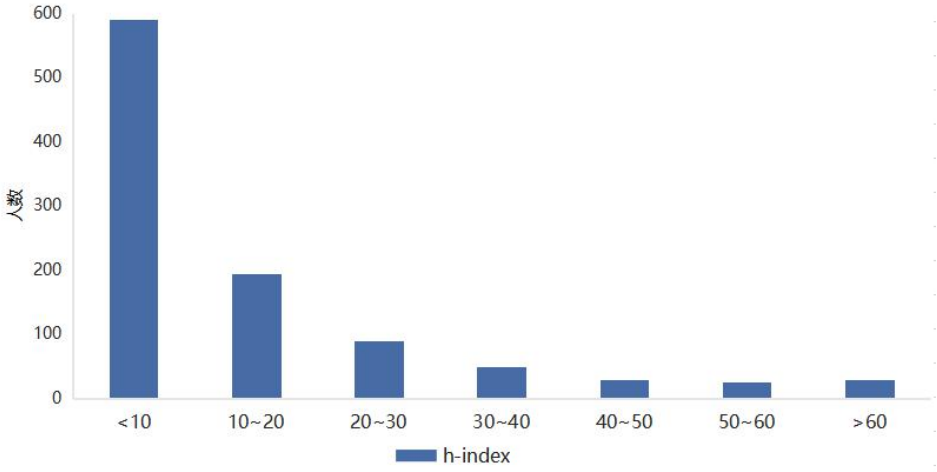


图 4-3 图数据库领域学者 h-index 分布

- **人才迁徙**

AMiner 可以对图数据库领域的学者的迁徙路径进行分析，如图 4-4 所示。从中可以看出，美国图数据库领域人才的流失和引进相对比较均衡，作为图数据库领域人才流动大国，人才输入和输出都大幅度领先，且从数据来看人才流入大于人才流出。中国、英国、德国和法国都落后于美国，中国和英国有轻微的人才流失现象，而法国有少量的人才流入。

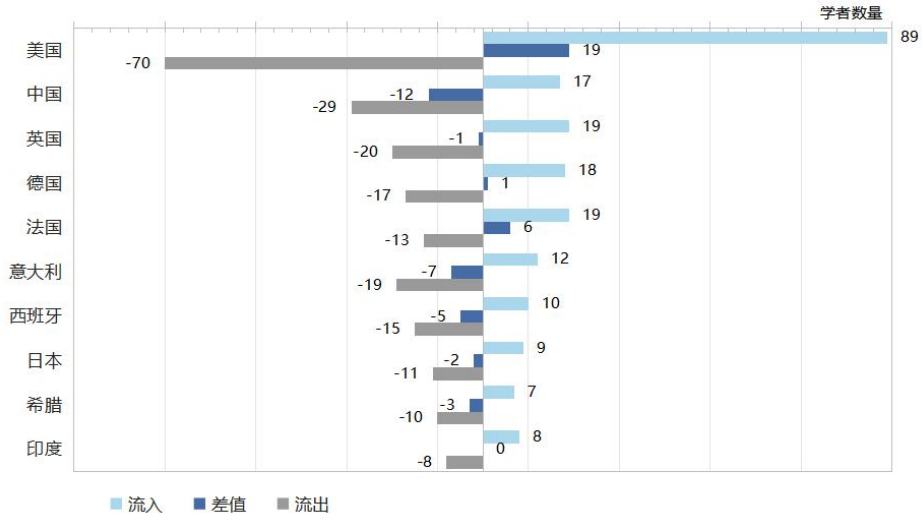


图 4-4 图数据库全球学者迁徙图

- **机构对比**

通过 AMiner 平台挖掘论文中的作者单位信息，将论文映射到各个单位机构中，统计每个机构的论文发表数量、学者数量以及 h-index，并按照论文发表数量从高到低对机构进行了排序，列出其中论文数量排名前五的机构，如图 4-5 所示。

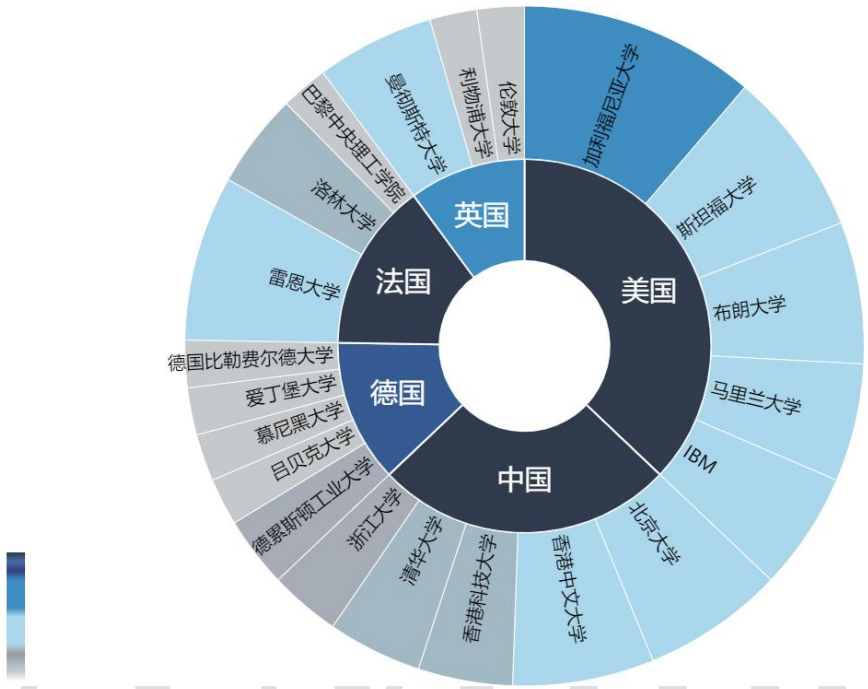


图 4-5 图数据库领域学术机构对比

从上图可以看出，美国、中国、德国、法国和英国拥有世界上最顶尖的科研机构。其中美国科研机构发表的论文数量最多，其中排名前三的科研机构分别为加利福尼亚大学、斯坦福大学、布朗大学。中国科研机构紧随其后，发表论文总数位居世界第二，其中排名前三的科研机构分别为北京大学、香港中文大学、香港科技大学。

4.1.2 国内学者概况

- **学者地图**

AMiner 选取图数据库领域国内专家学者绘制了学者国内分布地图，如图 4-6 所示。通过下图我们可以发现，珠三角地区在图数据库领域的人才数量最多，京津冀地区的也有较多的人才分布。相比之下，内陆地区图数据库产业人才较为匮

乏，这也从一定程度上说明了图数据库领域的发展与该地区的地理位置和经济水平都是息息相关的。

图 4-6 图数据库国内学者分布

- **中外合作**

中国与其他国家在图数据库领域的合作情况可以根据 AMiner 数据平台分析得到，通过统计论文中作者的单位信息，将作者映射到各个国家中，以中国与各国之间合作论文的数量的高低进行排序，其中合作论文数量前 10 的关系如表 4-1 所示。

表 4-1 图数据库领域中国与各国合作论文情况

序号	合作国家	论文数	平均引用数	引用数
1	中国-美国	38	103.03	3915
2	中国-新加坡	18	45.67	822
3	中国-加拿大	10	85.89	859
4	中国-澳大利亚	6	8.50	51
5	中国-日本	3	3.67	11
6	中国-意大利	2	6.00	12
7	中国-德国	2	2.50	5
8	中国-印度	1	4.00	4
9	中国-比利时	1	1.00	1
10	中国-巴西	1	0.00	0

从上表数据可以看出，中美合作的论文数、引用数、平均引用数、学者数遥遥领先，表明中美间在图数据库领域合作之密切；另外，中国与加拿大的合作论文数量虽然只有 10 篇，但平均数量达到 85.89 次，仅次于美国，说明在合作质量上达到了较高的水平。

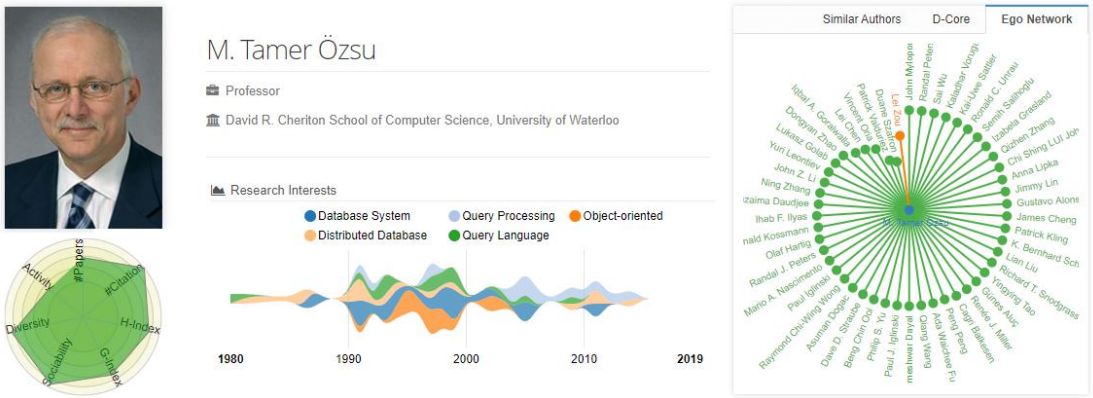
4.2 代表性学者及其论文解读

综合图数据库领域学者的 h-index 以及领域知名度与活跃度，我们收集整理了 M. Tamer Özsu、Peter Boncz、Jian Pei（裴健）、Haixun Wang（王海勋）、Frank Van Harmelen、Jeffrey Xu Yu（于旭）、Wenfei Fan、、Xuemin Lin（林学民）、Christian Bizer、Lei Chen（陈雷）、等十位领域高水平学者，通过“学者画像”的方式进行展示，另外还包括了学者的基本信息、AMiner 获奖信息、研究兴趣和相关代表性论文等，且学者排名不分先后。限于报告篇幅，不能逐一罗列所有学者，如有疏漏，还请与 AMiner 编者联系，或者登录 https://www.aminer.cn/获取更多资料。

- **M. Tamer Özsu**

教授

滑铁卢大学



M. Tamer Özsu 是滑铁卢大学计算机科学教授兼 David R.Cheriton 计算机科学学院院长。自 2007 年以来，他一直是 ACM 杰出讲师，是 CCS CAN 和 Inc CAN 的成员，在清华大学大数据软件国家工程实验室技术咨询委员会、香港科技大学工程学院咨询委员会、香港科技大学大数据研究所和香港大学技术专家咨询委员会多媒体软件工程研究中心工作。

M. Tamer Özsu 的研究方向是数据管理。在查询处理、事务处理和数据库集成等基础数据库技术方面也做了很多工作，主要的研究方向是：（1）数据库技术在非传统数据类型中的应用；（2）分布式并行数据管理。他是加拿大皇家学会（Royal Society of Canada）、美国科学促进协会（American Association for Advancement of Science）、计算机械协会（Association for Computing Machinery）、电气与电子工程师协会（Institute of Electrical and Electronics Engineers）、土耳其科学院（Science Academy of Turkey）当选成员和 Sigma Xi 成员。

相关论文精选

1.gStore: Answering SPARQL Queries via Subgraph Matching

Lei Zou, Jinghui Mo, Lei Chen, *M. Tamer Özsu*, Dongyan Zhao

PVLDB, no. 8 (2011)

论文链接:

https://www.aminer.cn/pub/53e9adbdb7602d97037c18e4/gstore-answering-sparql-queries-via-subgraph-matching

论文解读：由于 RDF 数据的使用越来越多，因此对 RDF 数据集进行 SPARQL 查询的有效处理已成为一个重要问题。但是，现有的解决方案有两个局限性：（1）它们无法以可扩展的方式用通配符回答 SPARQL 查询；（2）他们不能有效地处理 RDF 存储库中的频繁更新问题。因此，大多数人必须从头开始重新处理数据集。本文提出了一种基于图的方法来存储和查询 RDF 数据。（1）该方法没有像大多数现有方法那样将 RDF 三元组映射到关系数据库中，而是将 RDF 数据存储为大图。然后将 SPARQL 查询转换为相应的子图匹配查询；（2）为了加快查询处理速度，本文还提出了一种新颖的索引以及一些有效的修剪规则和有效的搜索算法。该方法可以以统一的方式回答确切的 SPARQL 查询和带通配符的查询；（3）文本中提出的维护算法可以有效处理 RDF 存储库的在线更新。

2.DistanceJoin: Pattern Match Query In a Large Graph Database

Lei Zou, Lei Chen, *M. Tamer Özsu*

PVLDB, no. 1 (2009): 886-897

论文链接:

https://www.aminer.cn/pub/53e9ac28b7602d97035e85c7/distancejoin-pattern-match-query-in-a-large-graph-database

论文解读：在对图数据进行子图搜索、最短路径查询、可达性验证和模式匹配时，模式匹配查询比子图搜索更具灵活性，比最短路径或可达性查询具有更多信息。本文解决了大数据图 G 上的模式匹配问题。具体来说，给定一个模式图（即查询 Q），我们希望查找所有具有与 Q 中相似连接的匹配，（在 G 中）为了显著减少搜索空间，我们首先通过图嵌入技术将顶点转换为向量空间中的点，然后将模式匹配查询覆盖到转换向量空间上的基于距离的多方联接问题。本文还提出了几种修剪策略和联接顺序选择方法来有效地处理联接处理。大量实验结果表明，该方法比现有方法的性能要高出几个数量级。

3.Processing SPARQL Queries Over Distributed RDF Graphs

Peng Peng, Lei Zou, *M. Tamer Özsu*, Lei Chen 0002, Dongyan Zhao

The VLDB Journal, no. 2 (2014)

论文链接：

https://www.aminer.cn/pub/56d87c63dabfae2eee44d017/processing-sparql-queries-over-distributed-rdf-graphs

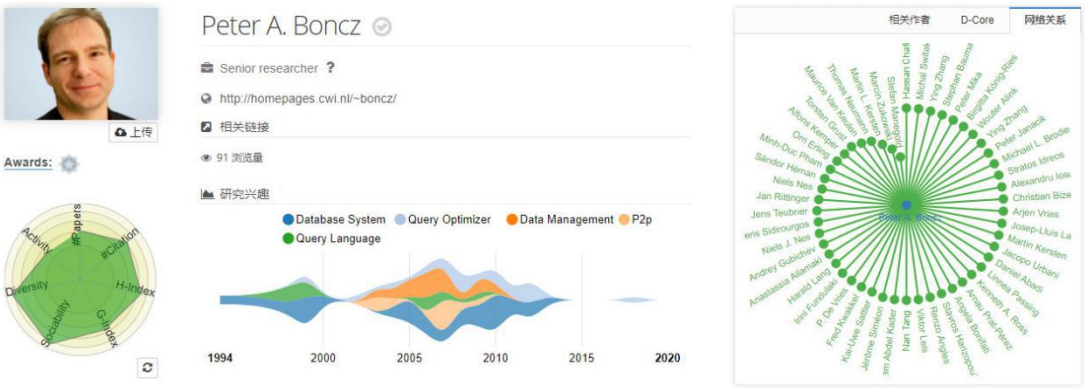
论文解读：本文提出了在分布式环境中通过大型 RDF 图处理 SPARQL 查询的技术。作者基于“部分评估和组装”框架，回答了 SPARQL 查询等同于在 RDF 图上查找查询图的子图匹配项。（1）针对分布图上子图匹配的性质，作者在 RDF 图的每个片段中引入部分答案；（2）提出了集中式组装和分布式组装；（3）在数十亿个三元组的真实 RDF 存储库和基准 RDF 存储库中进行的大量实验证实，该方法在系统性能和可伸缩性方面均是当前最优。

- **Peter Boncz**

教授

阿姆斯特丹自由大学

2018 AMiner Most Influential Scholar Award in Database



Peter Boncz，CWI 数据库架构研究组高级研究员，阿姆斯特丹自由大学教授。

Peter Boncz 的研究兴趣：数据库体系结构、计算机体系结构、自适应存储和查询处理、图形数据库和 RDF。Peter Boncz 曾担任 2014 — 2019 年 VLDB 捐赠基金董事会成员、与 Kenneth Salem 一起担任 2017 年 VLDB 的 PC 主席、CIDR2020 的 PC 主席、ACM SIGMOD 2019（阿姆斯特丹）总主席、《IEEE 数据工程公报》（2010 — 2012 年）副主编、VLDB 期刊副主编（2011 — 2017 年）。PVLDB 编辑委员会成员、PVLDB 副主编、PVLDB 主编。他还曾任 SIGMOD、VLDB、ICDE、EDBT、CIDR、CIKM（Area Chair）等主要数据库会议的 PC 成员。

相关论文精选

1.The Linked Data Benchmark Council: a Graph and RDF Industry Benchmarking Effort

Renzo Angles, *Peter A. Boncz*, Josep-Lluis Larriba-Pey, Irini Fundulaki, Thomas Neumann, Orri Erling, Peter Neubauer, Norbert Martínez-Bazan, Venelin Kotsev,Ioan Toma

SIGMOD Record, no. 1 (2014): 27-31

论文链接：

https://www.aminer.cn/pub/53e9ae5cb7602d97038770d2/the-linked-data-benchmark-council-a-graph-and-rdf-industry-benchmarking-effort

论文解读：链接数据基准委员会（LDBC）是一个欧盟项目，旨在为图形和 RDF 数据管理系统开发行业实力基准。它包括创建一个非营利性 LDBC 组织，组织行业参与者和学术界人才参与该项目，制定管理基准以及审计和发布正式结果。本文对 LDBC 项目进行了概述，包括项目目标和组织形式，并描述了用于基准测试开发的过程和设计方法。

2.Deriving an Emergent Relational Schema from RDF Data

Minh-Duc Pham, Linnea Passing, Orri Erling, *Peter A. Boncz*

WWW, pp.864-874, (2015)

论文链接:

https://www.aminer.cn/pub/5736977f6e3b12023e66600b/deriving-an-emergent-relational-schema-from-rdf-data

论文解读：该文主要描述了允许从 RDF 数据中检测“紧急”关系模式的技术。经过验证，在各种各样的数据集上，90％以上的 RDF 三元组结构具有可解释性。此外，本文还提出了一套针对语义挑战的技术解决方案，以给人们提供对这些紧急表、列和表之间关系具有逻辑意义的简称。该技术可以通过多种方式加以利用，例如，提高 SPARQL 系统的效率，或者在任何 RDF 数据集之上使用现有的 RDBMS 和基于 SQL 的应用程序。

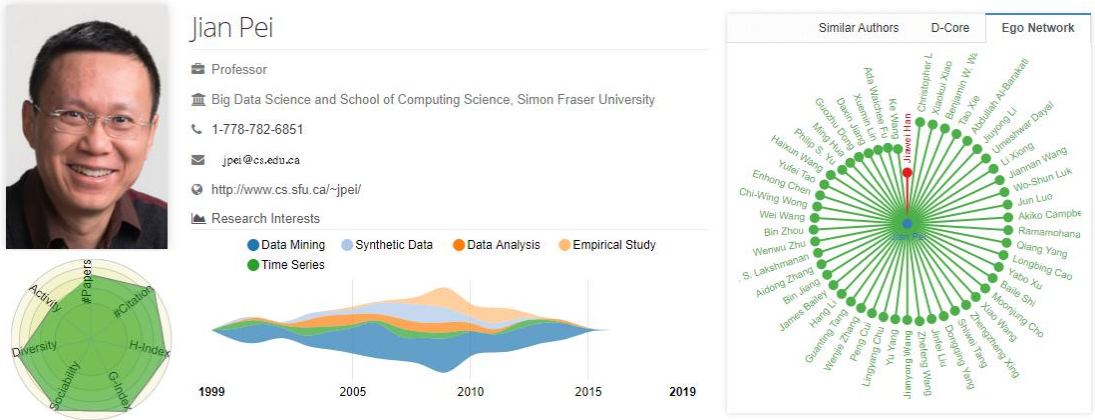
- **Jian Pei**

教授

加拿大西蒙弗雷泽大学

2016 AMiner Most Influential Scholar Award in Database

2016 AMiner Most Influential Scholar Award in Data Mining



研究领域是数据科学、大数据、数据挖掘和数据库系统。他曾任 IEEE 知识与数据工程交易（TKDE）主编（2013 — 2016 年），现任计算机协会（ACM）数据知识发现特别兴趣小组（SIGKDD）主席，以及许多高级会议的一般联合主席或项目委员会联合主席。

他与全球和本地行业合作伙伴保持广泛的行业关系。他是企业数据战略、医疗信息学、网络安全智能、互联网金融和智能零售业的积极顾问和教练。他的行业合作伙伴和客户包括《财富》全球 500 强企业和独角兽初创公司。

相关论文精选

1.Asymmetric transitivity preserving graph embedding

Mingdong Ou, Peng Cui, *Jian Pei*, Wenwu Zhu

KDD, (2016)

论文链接：

https://www.aminer.cn/pub/57aa28de0a3ac518da9896d8/asymmetric-transitivity-preserving-graph-embedding

论文解读：论文提出了一种全新的图嵌入算法，即 High-Order Proximity Preserved Embedding（HOPE），该算法既可扩展保留大型图的高阶邻近，还能够捕获非对称传递性。实验结果表明，HOPE 可以比现有算法更好地逼近高阶邻近度，并且在链接预测和顶点推荐等方面的表现优于现有算法。

2.Scalable mining of large disk-based graph databases

Chen Wang, Wei Wang, *Jian Pei*, Yongtai Zhu, Baile Shi

KDD, pp.316-325, (2004)

论文链接：

https://www.aminer.cn/pub/53e9a6d0b7602d9703007451/scalable-mining-of-large-disk-based-graph-databases

论文解读：该文提出了一种简单、有效的索引结构 ADI（邻接索引），以支持在无法保存到主存储器中的大型数据库上挖掘各种图形模式。此外，ADI 结构可以轻松地用于各种现有的图形模式挖掘算法中。例如，我们通过使用 ADI 结构来改编众所周知的 gSpan 算法。实验结果表明，ADI 可以在大型数据库上进行可伸缩的图形模式挖掘。在一组实验中，新的基于磁盘的方法可以挖掘具有一百万个图的图数据库，而原始的 gSpan 算法只能处理多达 30 万个图的数据库。而且，当两者都可以在主内存中运行时，新算法比 gSpan 更快。

3.On mining cross-graph quasi-cliques

Jian Pei, Daxin Jiang, Aidong Zhang

KDD, pp.228-238, (2005)

论文解读：本文提出了一种有效的算法 Crochet，该算法从跨市场客户细分以及基因表达数据和蛋白质的联合挖掘等一些有趣的应用中演化而来，可以有效解决交叉图拟群的挖掘问题。

- **Haixun Wang(王海勋)**

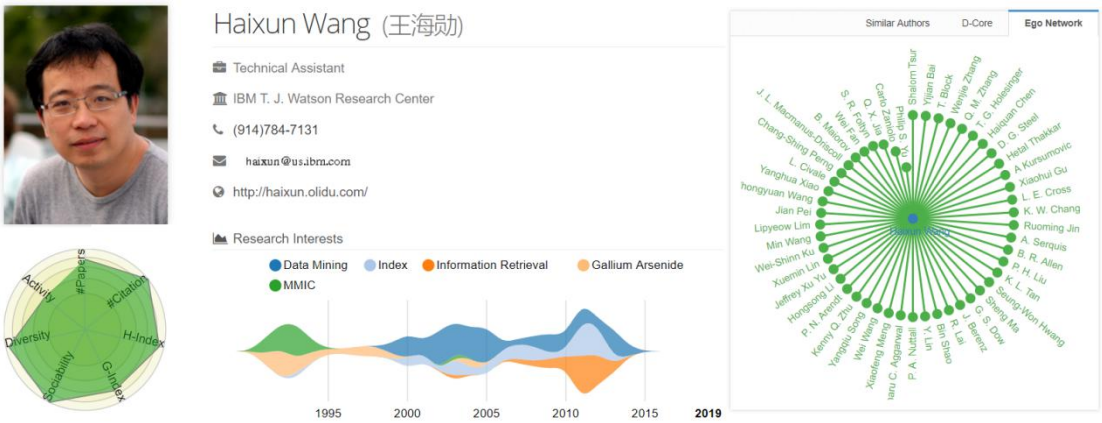
WeWork 副总裁

2018 AMiner Top 10 Most Influential Scholar Award in Database

2020 AI 2000 Most Influential Scholar Award Honorable Mention in Database

2016 AMiner Most Influential Scholar Award in Database

2016 AMiner Most Influential Scholar Award in Data Mining



王海勋的研究兴趣为：（1）文本分析、自然语言处理；（2）知识库、语义网络、人工智能；（3）数据库语言与系统，图数据管理。他是 TKDE（IEEE 数据工程事务）、KAIS（知识与信息系统）和 JCST（计算机科学与技术杂志）的编辑委员会成员。他是 ICMLA2011、WAIM 2011 的 PC 主席，曾担任 SIGKDD、SIAM DM、ICDM 等各种顶级会议的 PC 副主席和 PC 高级成员。目前在全球最大的联合办公空间公司 WeWork 担任技术工程副总裁。

相关论文精选

1.BLINKS: Ranked Keyword Searches on Graphs

Hao He, *Haixun Wang*, Jun Yang, Philip S. Yu

SIGMOD Conference, pp.305-316, (2007)

论文链接：

https://www.aminer.cn/pub/53e9aeebb7602d97039187bd/blinks-ranked-keyword-searches-on-graphs

论文解读：本文提出了一个图检索方案 BLINKS，该方法用于包含关键字 top-k 的图数据双层索引和查询处理搜索方案。BLINKS 搜索策略增加了可证明界限，同时还利用了双层索引来修剪和加速搜索。为了减少索引空间，BLINKS 将数据图划分为多个块：二级索引在块级别存储摘要信息以启动和引导块之间的搜索，并为每个块提供更详细的信息以加速块内的搜索。实验结果表明，BLINKS 比现有方法的性能提高了几个数量级。

2.ViST: a Dynamic Index Method for Querying XML Data by Tree Structures

Haixun Wang, Sanghyun Park, Wei Fan, Philip S. Yu

SIGMOD Conference, pp.110-121, (2003)

论文链接：

https://www.aminer.cn/pub/53e9a812b7602d9703157a36/vist-a-dynamic-index-method-for-querying-xml-data-by-tree-structures

论文解读：本文提出了一种用于搜索 XML 文档的新颖索引结构——ViST。与索引方法不同，ViST 索引方法将查询分解为多个子查询，然后将这些子查询的结果连接起来以提供最终答案，ViST 使用树结构作为查询的基本单位，以避免昂贵的连接操作。此外，ViST 在 XML 文档的内容和结构上都提供了统一的索引，因此与仅对内容或结构进行索引的方法相比，ViST 具有性能优势。ViST 支持动态索引更新，它仅依赖 B +树，而没有使用 DBMS 不能很好支持的任何专用数据结构。实验表明，ViST 在支持结构化查询方面具有可扩展和高效性。

3.Natural language question answering over RDF: a graph data driven approach

Lei Zou, Ruizhe Huang, *Haixun Wang*, Jeffrey Xu Yu, Wenqiang He, Dongyan Zhao

SIGMOD Conference, (2014)

论文链接：

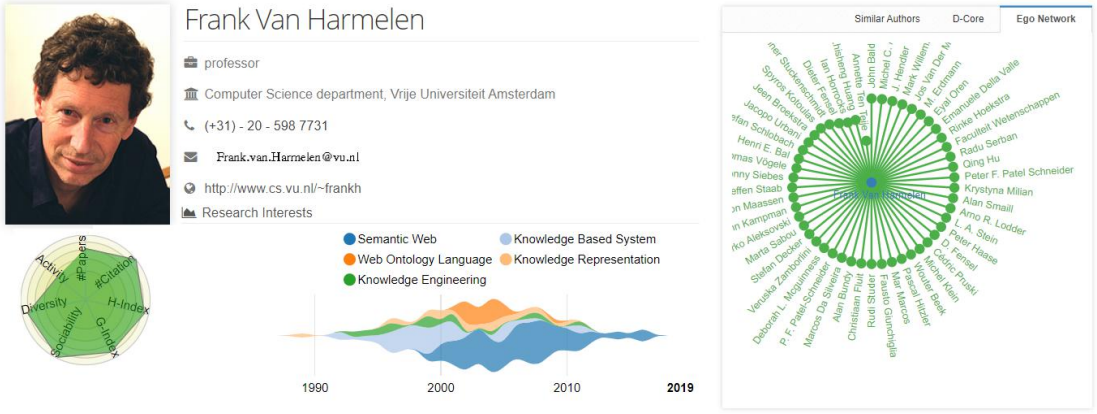
https://www.aminer.cn/pub/555048b745ce0a409eb70b87/natural-language-question-answering-over-rdf-a-graph-data-driven-approach

论文解读：RDF 问题/答案（Q/A）允许用户在 RDF 代表的知识库上以自然语言提问。为了回答一个国家的语言问题，现有工作采用两个阶段的方法：问题理解和查询评估。它们的重点是对问题的理解，以解决自然语言短语的歧义问题。最常见的技术是联合消歧。本文提出了一个用于从图形数据驱动的角度回答 RDF 存储库（RDF Q/A）上的自然语言问题系统的框架。该框架是一种以结构化的方式对自然语言问题的查询意图进行建模的语义查询图。在此基础上，将 RDF Q / A 简化为子图匹配问题。尤其当找到查询匹配项时，该方法还解决了自然语言的歧义问题。大量实验结果表明，与传统方法相比，该方法在精度和查询性能方面都有大幅提高。

- **Frank Van Harmelen**

教授

阿姆斯特丹自由大学



Frank van Harmelen，阿姆斯特丹自由大学计算机科学系的知识表示和推理教授，荷兰皇家科学和人文学会的成员，兼任中国武汉科技大学的客座教授。1989 年获得数学和计算机科学博士学位。在爱丁堡期间，他与 Alan Bundy 教授共同开发了一个基于逻辑的专家系统工具包，以归纳定理证明的证明规划。1990 年到 1995 年，他回到了阿姆斯特丹，在 Wielinga 教授领导的 SWI 系工作。1995 年，他加入了阿姆斯特丹自由大学的人工智能研究小组，现为阿姆斯特丹知识表示研究小组负责人。

Frank van Harmelen 对于语义网的发展起到举足轻重的作用。他是第一个欧洲语义网项目的合作伙伴（OWL），该项目为网络本体语言语义网奠定了基础。OWL 已经成为一个世界性的标准，不仅被广泛的商业应用，并且已经成为整个

研究团体的基础。他与人合著的 *Semantic Web Primer*，是该领域的第一本学术教科书，已被翻译成 5 种语言，在全世界传播。Frank van Harmelen 也是 Sesame 的设计师之一，Sesame 的是一个 RDF 存储和检索引擎，在学术界和工业界有着广泛的应用，下载量超过 20 万次。此项工作在 2012 年第 11 届国际语义网会议上获得了 10 年影响奖，这是该领域最具声望的奖项。

相关论文精选

1.Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema

Jeen Broekstra, Arjohn Kampman, *Frank Van Harmelen*

International Semantic Web Conference, pp.54-68, (2002)

论文链接：

https://www.aminer.cn/pub/53e9a22bb7602d9702b2e32e/sesame-a-generic-architecture-for-storing-and-querying-rdf-and-rdf-schema

论文解读：本文概述了 RDF 和 RDF Schema 作为一种通用的体系结构，以及它的实现和对该结构的初步应用经验。RDF 和 RDF Schema 是两个 W3C 标准，旨在通过机器可处理的语义数据丰富 Web。Sesame 是一种用于 RDF 和 RDF Schema 中高效存储和表达查询大量元数据的体系结构。Sesame 的设计和实现独立于任何特定的存储设备。因此，可以将 Sesame 部署在各种存储设备之上，例如关系数据库、三重存储或面向对象的数据库，而不必更改查询引擎或其他功能模块。Sesame 结构支持并发控制，RDF 和 RDFS 信息的独立导出以及 RQL 的查询引擎，RQL 是 RDF 的查询语言，其原生支持 RDF Schema 语义。

2.A Semantic Web Primer

Grigoris Antoniou, *Paul Groth*, *Frank Van Van Harmelen*, Rinke Hoekstra

The Computer Journal, no. 1 (2005): 126-126

论文链接：

https://www.aminer.cn/pub/53e99a52b7602d97022b849c/a-semantic-web-primer

论文解读：具有机器可读内容的语义 Web 的发展具有彻底改变万维网及其用途的潜力。语义 Web 入门指南提供了对该领域介绍和指南，描述了其关键思想、

语言和技术。本书适合用作教科书或供专业人士独立研究，它着重于本科水平的基本概念和技术，使读者能够自己继续构建应用程序，包括练习、项目说明以及对相关在线资料的带注释的参考。

本书内容：（1）对不同语言（OWL2 规则）的处理扩展了 RDF 和 OWL 的覆盖范围，独立于 XML 定义了数据模型，并包括 N3/Turtle 和 RDFa 的覆盖范围；（2）专门介绍 OWL2（新的 W3C 标准）；（3）涵盖了查询语言 SPARQL、规则语言 RIF 以及规则与本体语言和应用程序之间交互的可能性。

3.From SHIQ and RDF to OWL: the making of A Web Ontology Language

Ian Horrocks, Peter F. Patel-Schneider, *Frank van Harmelen*

J. Web Sem., no. 1 (2003): 7-26

论文链接：

https://www.aminer.cn/pub/53e9ac7bb7602d9703652774/from-shiq-and-rdf-to-owl-the-making-of-a-web-ontology

论文解读：OWL Web 本体语言是一种新的形式化语言，用于表示语义 Web 中的本体。OWL 具有描述逻辑和框架功能。OWL 还与 RDF（语义 Web 的 W3C 基础）共享许多特性。本文讨论了 OWL 的哲学和功能如何可以追溯到较旧的形式主义上，并通过对 OWL 的其他一些约束来进行修改。

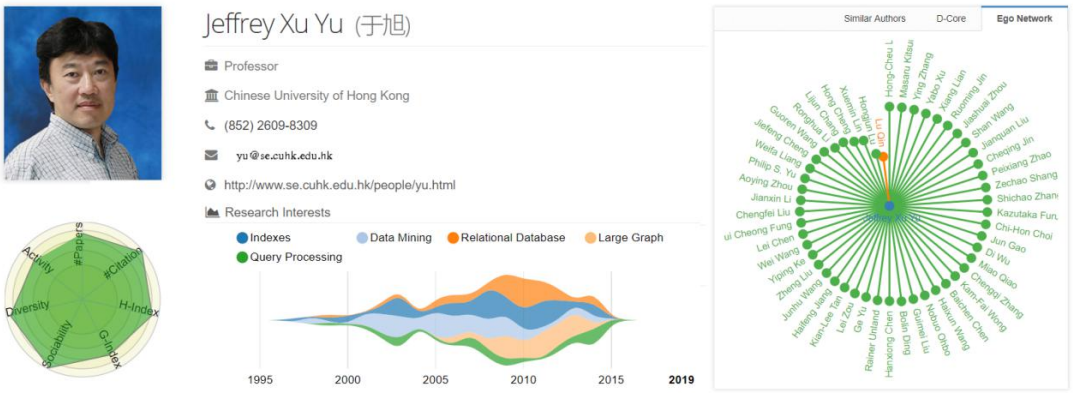
- **Jeffrey Xu Yu(于旭)**

教授

香港中文大学

2020 AI 2000 Most Influential Scholar Award Honorable Mention in Database

2018 AMiner Most Influential Scholar Award in Database



于旭，香港中文大学系统工程与工程管理系教授。

于旭的主要研究兴趣包括关系数据库中的关键字搜索、图挖掘、图查询处理以及图模式匹配。于旭在国际会议/研讨会上为 300 多个组织委员会和程序委员会提供服务，其中包括 APWeb'04，WAIM'06，APWeb/WAIM'07，WISE'09，PAKDD'10，DASFAA'11 的 PC 联合主席，ICDM'12，NDBC'13，ADMA'14，CIKM'15，Bigcomp'17，DSAA'19 和 CIKM'19，以及 APWeb'13 和 ICDM'18 的大会主席。于旭曾担任 ACM SIGMOD 执行委员会的信息总监和成员（2007～2011 年），IEEE 知识和数据工程事务的副主编（2004～2008 年），VLDB Journal 的副主编（2007～2013 年），亚太网络会议指导委员会主席（2013～2016 年）。目前，他还担任 *ACM Transactions on Database Systems*，*WWW Journal*，*the International Journal of Cooperative Information Systems*，*the Journal of Information Processing* 和 *Journal on Health Information Science and Systems* 的副主编辑。

相关论文精选

1.Graph Clustering Based on Structural/Attribute Similarities

Yang Zhou, Hong Cheng, *Jeffrey Xu Yu*

PVLDB, no. 1 (2009): 718-729

论文链接：

https://www.aminer.cn/pub/53e9a9ebb7602d970334d2d6/graph-clustering-based-on-structural-attribute-similarities

论文解读：图聚类的目标是基于各种标准（例如：顶点连通性或邻域相似性）将大图上的顶点划分为不同的聚类，图聚类技术对于检测大型图中的密集连接的非常有用。现有的许多图聚类方法主要集中在聚类的拓扑结构上，但在很大程度上

忽略了异性结构的顶点属性。在本文中，作者基于结构和属性的相似性，通过统一的距离度量，提出了一种新颖的图聚类算法 SA-Cluster。该方法将与属性关联的大图划分为 k 个簇，以便每个簇包含具有均一属性值的密集连接的子图，以此有效地提高自动学习结构相似度和属性相似度性能。大量实验结果表明，SA-Cluster 的收敛性可靠有效。

2.Graph Indexing: Tree + Delta

Peixiang Zhao, *Jeffrey Xu Yu*, Philip S. Yu

VLDB '07 Proceedings of the 33rd international conference on Very large data bases, pp.938-949, (2007)

论文链接：

https://www.aminer.cn/pub/53e9a80cb7602d97031518d0/graph-indexing-tree-delta

论文解读：本文基于图数据库的树特征，提出了一种全新的高效图索引方法。我们从三个关键方面来分析树作为索引特征的有效性和效率：特征大小、特征选择成本和修剪能力。为了获得比现有的基于图的索引方法更好的修剪能力，除了频繁的树特征（Tree）之外，我们还根据需要选择少量的判别图（Δ），而无须事先进行大量的图挖掘。研究证明，（Tree +Δ≥Graph）可用于索引，并且是解决图查询问题的最佳选择。事实证明：（1）利用（Tree +Δ）进行索引构建是有效的，（2）利用（Tree +Δ）进行图包含查询处理是有效的。实验结果表明，（Tree +Δ）具有紧凑的索引结构，在索引构建中实现了更好的性能数量级。此外，该索引方法的性能优于 gIndex 和 C-Tree。

3.Fast Graph Pattern Matching

Jiefeng Cheng, *Jeffrey Xu Yu*, Bolin Ding, Philip S. Yu, Haixun Wang

ICDE, pp.913-922, (2008)

论文链接：

https://www.aminer.cn/pub/53e99a20b7602d9702278b35/fast-graph-pattern-matching

论文解读：本文基于传统的图形模式匹配，提出了一种新的两步 R-join（可达性联接）算法，该算法基于带有图代码群集的联接索引，实施过滤和提取步骤。其中过滤步骤为 R-半连接，并通过将 R-连接与 R-半连接交错连接的方法进行优化。

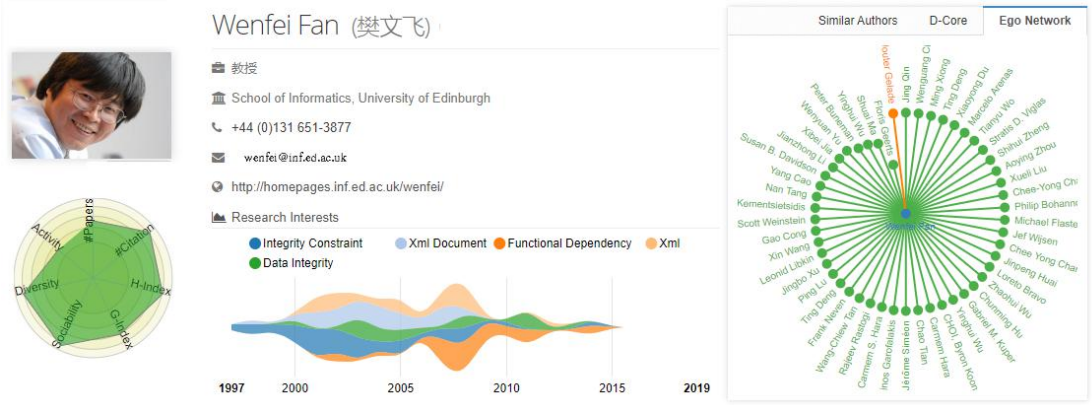
- **Wenfei Fan**

教授

爱丁堡大学

2020 AI 2000 Most Influential Scholar Award Honorable Mention in Database

2018 AMiner Most Influential Scholar Award in Database



樊文飞，英国爱丁堡大学信息学院首席教授。现为英国工程与物理科学研究等多项基金评委，曾为美国宾夕法尼亚大学、美国 Drexel 大学、北京航空航天大学客座教授；美国贝尔实验室研究员和科学家。

樊文飞主要研究领域为数据库理论与系统，包括大数据、数据质量、数据集成、分布式查询处理、查询语言、推荐系统、社会网络查询与分析、Web 服务等。

相关论文精选

1.Incremental Graph Pattern Matching

Wenfei Fan, Jianzhong Li, Jizhou Luo, Zijing Tan, Xin Wang, Yinghui Wu

ACM Trans. Database Syst., no. 3 (2013)

论文链接：

https://www.aminer.cn/pub/53e99a98b7602d970230f8f1/incremental-graph-pattern-matching

论文解读：本文研究了图形模式匹配的两个问题。首先，传统上的图形模式匹配是根据子图同构或图仿真定义图模式匹配的。但是，这些概念经常在图形上施加过强的拓扑约束，无法识别出有意义的匹配项。其次，在实践中，图形通常较大，即使只是较小的图形更新，也要通过批处理算法从头开始重新计算，造成过多消

耗。在此基础之上，作者提出了图形模式匹配的新方法（1）基于有界仿真的概念定义图模式匹配，该方法通过在预定义跳数内指定图中节点的连通性来扩展图仿真；（2）使用图仿真、有界仿真和子图同构定义的匹配模式；（3）通过实验验证了算法的有效性和效率。结果表明：（a）修正的图模式匹配概念使我们能够识别现实网络中常见的社区；（b）增量算法的性能明显优于他们的批次应对微小的变化。

2.Query Preserving Graph Compression

Wenfei Fan, Jianzhong Li, Xin Wang, Yinghui Wu

SIGMOD Conference, pp.157-168, (2012)

论文链接：

https://www.aminer.cn/pub/53e99afdb7602d970238b1fa/query-preserving-graph-compression

论文解读：在该论文中，笔者对图形的压缩方法进行了研究。（1）针对可达性和通过（有限）模拟的图形模式查询开发了压缩策略；（2）提供了根据原始图形 G 的变化ΔG 来获得压缩图形 Gr 的技术；（3）实验验证，该增量维护算法可靠有效。

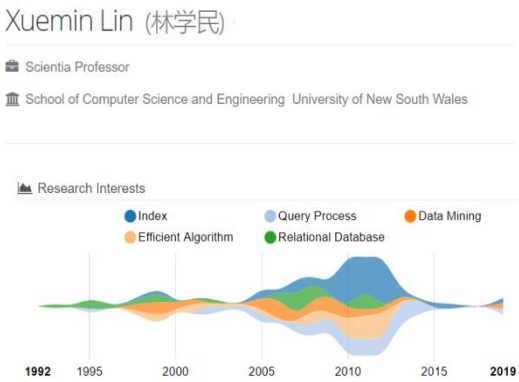
- **Xuemin Lin（林学民）**

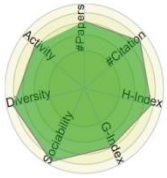
教授

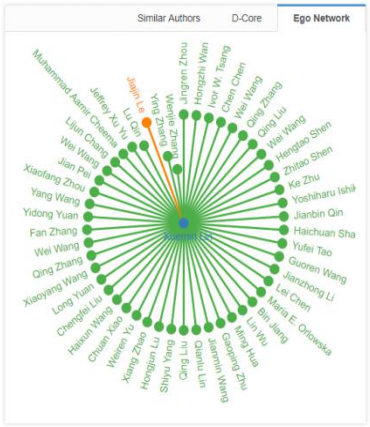
新南威尔士大学

2020 AI 2000 Most Influential Scholar Award Honorable Mention in Database

2018 AMiner Most Influential Scholar Award in Database







林学民，新南威尔士大学计算机科学及工程学院教授、数据库研究实验室主任、新南威尔士大学的首席教授。

林学民研究方向为数据库理论、算法与技术研究；时空数据和流数据的查询、图和文本的匹配查询、不确定数据的概化查询及图数据可视化等。林学民教授曾是顶级期刊 *ACM Transactions on Database Systems* 的编委（2008～2014 年）。目前是顶级期刊 *IEEE Transactions on Knowledge and Data Engineering* 的编委，并多次应邀担任 SIGMOD，VLDB，ICDE，KDD 等顶级会议的程序委员。

相关论文精选

1.Fast Computing Reachability Labelings for Large Graphs with High Compression Rate

Jiefeng Cheng, Jeffrey Xu Yu, *Xuemin Lin*, Haixun Wang, Philip S. Yu

EDBT, pp.193-204, (2008)

论文链接：

https://www.aminer.cn/pub/53e9a6e6b7602d970301d566/fast-computing-reachability-labelings-for-large-graphs-with-high-compression-rate

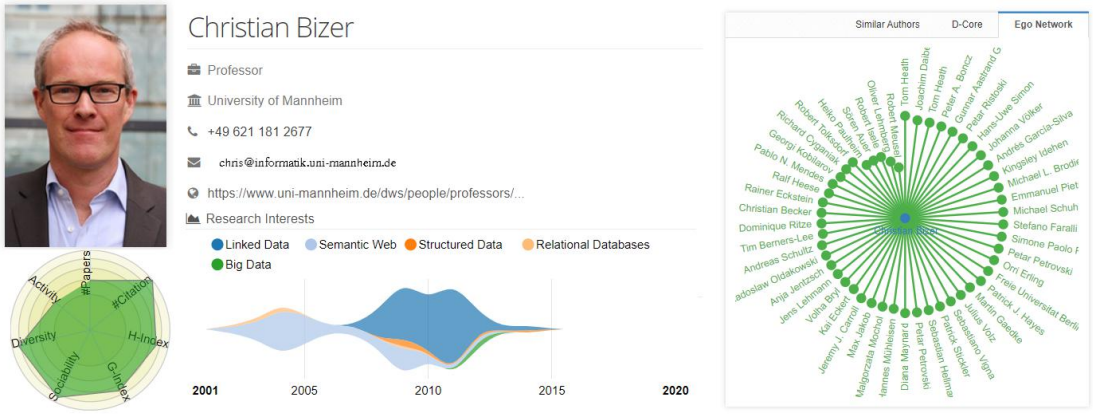
论文解读：目前，有许多应用程序需要处理大型图，并且需要查询图中节点之间的可达性。在本文中，作者提出了一种分层方法。实验表明，使用该方法可以在不到 30 分钟的时间内，以大约 40,000 的压缩率，获得具有 1,700,000 个节点和 1,690 亿个连接图的 2 跳覆盖。

- **Christian Bizer**

教授

曼海姆大学

2020 AI 2000 Most Influential Scholar Award in **Knowledge Engineering**



Christian Bizer 的研究方向是探索与关全球化、分散化信息环境发展的技术和经验问题。由他主持的项目包括 W3C 链接开放数据社区项目、DBpedia 项目（该项目是从 Wikipedia 派生出的一个全面的知识库）、WebData Commons 项目（该项目通过分析大型 Web 爬虫来监视 RDFa、Microdata 和 Microformat 数据在 Web 上的部署）。

Christian Bizer 的科研成果包括：（1）RDF 和 SPARQL 建议中采用的命名图数据模型；（2）在 Web 上发布关系数据库的 D2RQ 映射语言；（3）Silk-Identity Resolution Framework、Berlin SPARQL Benchmark 和 Mannheim Search Join Engine。

相关论文精选

1.DBpedia: A Nucleus for A Web of Open Data

Sören Auer, *Christian Bizer*, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives

ISWC/ASWC, pp.722-735, (2007)

论文链接：

https://www.aminer.cn/pub/53e9bbdbb7602d970482c288/dbpedia-a-nucleus-for-a-web-of-open-data

论文解读：DBpedia 是社区的一项工作，旨在从 Wikipedia 中提取结构化信息，并使该信息在 Web 上可用。DBpedia 允许客户对源自 Wikipedia 的数据集进行复杂的查询，并将 Web 上的其他数据集链接到 Wikipedia 数据。本文描述了：（1）DBpedia 数据集的提取，以及如何将结果信息发布在 Web 上以供人和机器使用；

（2）来自 DBpedia 社区的一些新兴应用程序，并展示了网站作者如何在他们的站点中促进 DBpedia 内容；（3）最后介绍了将 DBpedia 与 Web 上其他开放数据集互连的当前状态，并概述了 DBpedia 如何充当新兴开放数据 Web 的核心。

2.The Berlin SPARQL Benchmark

Christian Bizer, Andreas Schultz

Int. J. Semantic Web Inf. Syst., no. 2 (2009): 1-24

论文链接：

https://www.aminer.cn/pub/53e99ad7b7602d970235b073/the-berlin-sparql-benchmark

论文解读：随着 SPARQL 被社区所采用，越来越需要基准测试来比较通过 SPARQL 协议公开 SPARQL 端点的存储系统的性能。此类系统包括本机 RDF 存储以及针对非 RDF 关系数据库将 SPARQL 查询重写为 SQL 查询的系统。本文介绍的 Berlin SPARQL Benchmark（BSBM）同时拥有本机 RDF 存储的性能与跨体系结构的 SPARQL-to-SQL 重写器的性能。该基准测试建立在一个电子商务用例的基础上，在该用例中，不同供应商提供了一组产品，并且消费者已经发布了有关产品的评论。基准查询混合模拟了正在寻找产品的消费者的搜索和导航模式。本文还讨论了 BSBM 基准的设计，对四种流行的 RDF 存储（Sesame，Virtuoso，Jena TDB 和 Jena SDB）性能与两个 SPARQL-to-SQL 重写器（D2R）的性能进行了比较。

3.D2R Server-Publishing Relational Databases on the Semantic Web

Christian Bizer, Richard Cyganiak, Freie Universit

International Symposium on Wearable Computers, (2004)

论文链接：

https://www.aminer.cn/pub/53e9a0d1b7602d97029b960e/d-r-server-publishing-relational-databases-on-the-semantic-web

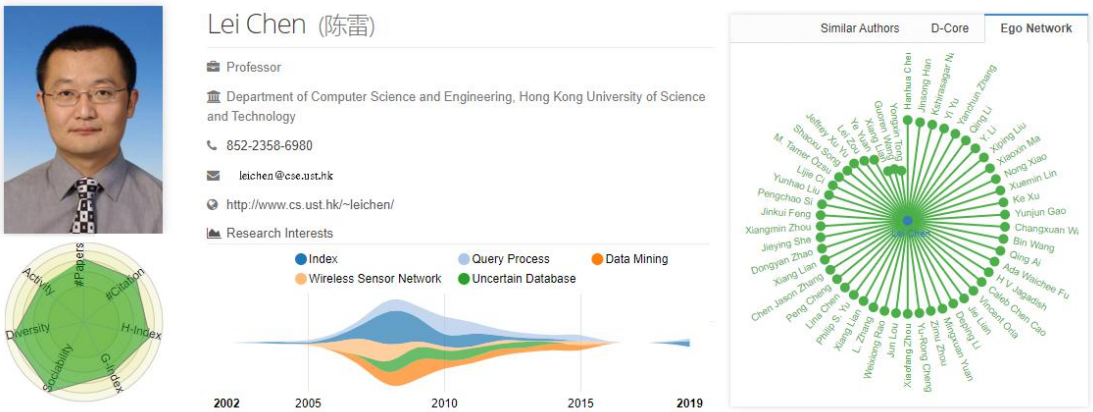
论文解读：本文中的 D2R Server 是用于在语义网上发布关系数据库内容的工具。通过声明性映射将数据库内容映射到 RDF，该声明性映射指定如何标识资源以及如何从数据库内容生成属性值。基于此映射，D2R 服务器允许 Web 代理使用 SPARQL 协议上的 SPARQL 查询语言来检索资源的 RDF 和 XHTML 表示并查询非 RDF 数据库。生成的表示在 RDF 和 XHTML 级别上进行了丰富的互连，以使浏览器和搜寻器能够导航数据库内容。

- **Lei Chen**

教授

香港科技大学

2018 AMiner Most Influential Scholar Award in Database



陈雷，香港科技大学计算机科学与工程系教授。

陈雷的研究领域为数据驱动的机器学习、基于众包的数据处理、不确定和概率数据库、Web 数据管理、多媒体和时间序列数据库、隐私保护数据的发布等。近年来，陈雷博士举办了一系列国际学术会议并担任 ACM SIGMM 2011、ACM CIKM 2012 和 IEEE ICDE 2012 会议的程序委员会副主席，担任 IEEE Transactions on Knowledge and Data Engineering（IEEE TKDE）与 Distributed and Parallel Databases（DAPD）等国际期刊编委（Associate Editors）。

相关论文精选

1.gStore: Answering SPARQL Queries via Subgraph Matching

Lei Zou, Jinghui Mo, *Lei Chen*, M. Tamer Özsu, Dongyan Zhao

PVLDB, no. 8 (2011)

论文链接：

https://www.aminer.cn/pub/53e9adbdb7602d97037c18e4/gstore-answering-sparql-queries-via-subgraph-matching

论文解读：由于 RDF 数据的使用越来越多，因此对 RDF 数据集进行 SPARQL 查询的有效处理已成为一个重要问题。但是，现有的解决方案有两个局限性：1）它们无法以可扩展的方式用通配符回答 SPARQL 查询；2）他们不能有效地处理 RDF 存储库中的频繁更新问题。

因此，大多数人必须从头开始重新处理数据集。本文提出了一种基于图的方法来存储和查询 RDF 数据。（1）该方法没有像大多数现有方法那样将 RDF 三元组映射到关系数据库中，而是将 RDF 数据存储为大图。然后将 SPARQL 查询转换为相应的子图匹配查询；（2）为了加快查询处理速度，本文还提出了一种新颖的索引以及一些有效的修剪规则和有效的搜索算法。该方法可以以统一的方式回答确切的 SPARQL 查询和带通配符的查询；（3）文本中提出的维护算法可以有效处理 RDF 存储库的在线更新。

2.DistanceJoin: Pattern Match Query in A Large Graph Database

Lei Zou, *Lei Chen*, M. Tamer Özsu

PVLDB, no. 1 (2009): 886-897

论文链接：

https://www.aminer.cn/pub/53e9ac28b7602d97035e85c7/distancejoin-pattern-match-query-in-a-large-graph-database

论文解读：在对图数据进行子图搜索、最短路径查询、可达性验证和模式匹配时，模式匹配查询比子图搜索更具灵活性，比最短路径或可达性查询具有更多信息。本文解决了大数据图 G 上的模式匹配问题。具体来说，给定一个模式图（即查询 Q），我们希望查找所有具有与 Q 中相似连接的匹配。（在 G 中）为了显著减少搜索空间，我们首先通过图嵌入技术将顶点转换为向量空间中的点，然后将模式匹配查询覆盖到转换向量空间上的基于距离的多方联接问题。本文还提出了几种修剪策略和联接顺序选择方法来有效地处理联接处理。大量实验结果表明，该方法比现有方法的性能要高出几个数量级。

5 趋势篇

领域技术分析系统（http://trend.aminer.cn）基于 AMiner 的近 3 亿篇论文和专利数据进行深入发掘，全面分析领域技术趋势、国际趋势、机构趋势及学者趋势等。本次研究以数据管理国际会议（The ACM Special Interest Group on Management of Data, SIGMOD）、超大型数据库国际会议（International Conference on Very Large Databases, VLDB）、IEEE 国际数据工程会议（IEEE International Conference on Data Engineering, ICDE）、图形数据管理经验与系统国际研讨会（International Workshop on Graph Data Management Experiences & Systems, GRADES）、扩展数据库技术国际会议（International Conference on Extending Database Technology, EDBT）会议上发表的图数据库相关论文、专利以及国家自然科学基金委员会（National Nature Science Foundation of China, NSFC）扶持的基金项目作为研究基础，对图数据库领域的热点趋势进行详尽分析[2]，

5.1 国家趋势

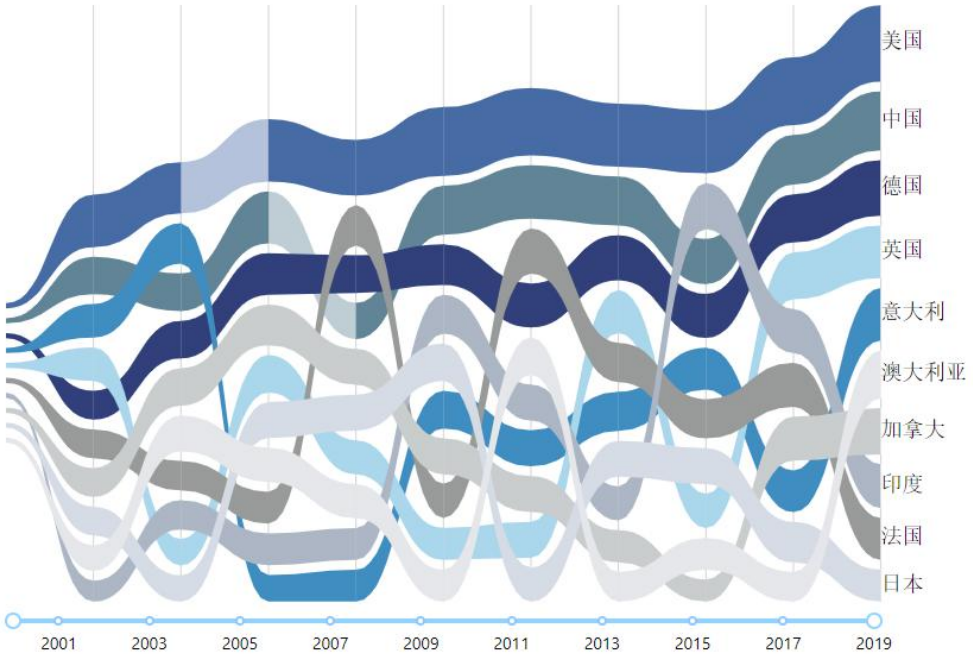


图 5-1 图数据库国家趋势

[2] 图数据库专利数据来自智慧芽专利分析系统 https://analytics.zhihuiya.com/

国家趋势分析如图 5-1 所示。图中每条色带表示一个国家，其宽度表示该国家在当年的研究热度，与当年该国论文数量呈正相关，每一年份中按照其热度由高到低进行排序。通过国家趋势分析可以发现当前图数据库领域研究热度 Top10 的国家分别是：美国、中国、德国、英国、意大利、澳大利亚、加拿大、印度、法国、日本。

根据国家趋势分析我们可以发现，图数据库领域当前研究热度最高的国家是美国，从全局热度来看，美国早期就有着领先优势并一直保持着较高的热度。同时可以看出，中国在图数据库领域的研究热度仅次于美国，尤其是在 2015 年以后。

5.2 论文技术趋势

我们根据图数据库的关键词，从 AMiner 数据库中查找出历年论文，其中包含论文所在领域的分支术语和年份，统计含有这些术语论文数量，给出论文数量排名前 10 的术语，再统计这些术语的起止年份，划分时间窗格，生成大数据智能的发展趋势图，如图 5-2 所示。

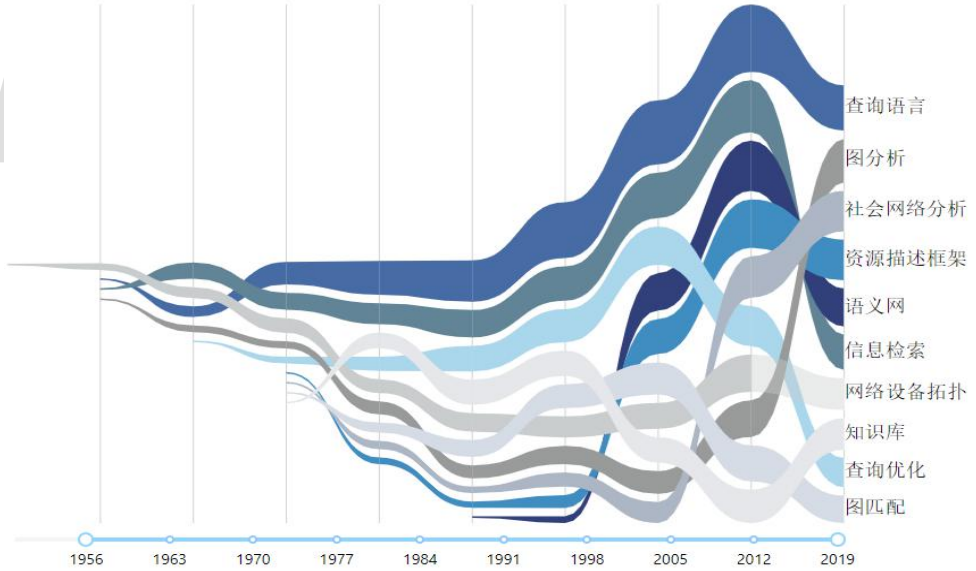


图 5-2 图数据库的热点趋势图

上图中的每个色带表示一个领域分支术语，其宽度表示该术语在当年的热度，与当年该分支领域的论文数量呈正相关；各分支在每一年份中按照其热度进行排序，越热的在越上方。对目前热度靠前的 10 个分支领域进行历史热度展示，从

趋势图中可以看出，查询语言的研究热度一直位居图数据库领域的首位；图分析、社会网络分析知识库在近五年来热度有所上升，尤其是图分析已经发展成为图数据库的重要分支领域；信息检索、语义网和查询优化的研究热度有下降趋势。

5.3 领域热点话题

为了帮助读者了解图数据库领域的热点研究话题，本报告针对 AMiner 平台上收录的专家推荐的 100 篇必读论文（https://www.aminer.cn/topic/5eec8ad092c7f9be2177bcc6），采用主题生成模型（Latent Dirichlet Allocation, LDA），分析了这些论文的研究主题分布情况。其中，查询语言、资源描述框架、图分析、社交网络和图数据库管理都是该领域的研究热点，热点话题的代表性论文如下所示。

- **查询语言**

1. 标题：Performance of Graph Query Languages: Comparison of Cypher, Gremlin and Native Access in Neo4j

出处：Conference: Joint EDBT/ICDT 2013 Workshop GraphQ, 2013

作者：Holzschuher F, Peinl R.

2. 标题：Cypher: An Evolving Query Language for Property Graphs

出处：2018 ACM SIGMOD Conference, 2018

作者：Francis N, Green A, Guagliardo P, Libkin L, Lindaaker T, Marsault V, Plantikow S, Rydberg M, Selmer P, Taylor A

3. 标题：Foundations of Modern Query Languages for Graph Databases

出处：ACM Computing Surveys, 2017

作者：Angles R, Arenas M, Barceló P, Hogan A, Reutter J, Vrgoč D.

- **资源描述框架**

1. 标题：A Graph Model for RDF

出处：Diploma Thesis, Technische Universitat Darmstadt，Universidad de Chile, 2004

作者：Hayes J.

2. 标题：Reconciliation of RDF* and Property Graphs

出处：Computing Research Repository, 2014

作者：Hartig O.

3. 标题：Querying RDF Data from a Graph Database Perspective

出处：he Semantic Web: Research and Applications, 2005

作者：Angles R, Gutierrez C.

- **图分析**

1. 标题：Survey of Graph Database Performance on the HPC Scalable Graph Analysis Benchmark

出处：Web-Age Information Management 2010 Workshops, 2010

作者：Dominguez-Sal D, Urbón-Bayes P, Giménez-Vanó A, Gómez-Villamor S, Martínez-Bazan N, Larriba-Pey J-L

2. 标题：LDBC Graphalytics: A Benchmark for Large_Scale Graph Analysis on Parallel and Distributed Platforms

出处：Proceedings of the VLDB Endowment, 2010

作者：Iosup A, Hegeman T, Ngai WL, Heldens S, Prat-Pérez A, Manhardto T, Chafio H., Capotă M., Sundaram N., Anderson M

3. 标题：Management and Analysis of Big Graph Data: Current Systems and Open Challenges

出处：In book: Handbook of Big Data Technologies, 2017

作者：Junghanns M., Petermann A., Neumann M., Rahm E

- **社交网络**

1. 标题：BG: A Benchmark to Evaluate Interactive Social Networking Actions

出处：Future Generation Computer Systems, 2018

作者：Barahmand S, Ghandeharizadeh S.

2. 标题：The LDBC Social Network Benchmark: Interactive Workload

出处：2015 ACM SIGMOD Conference, 2015

作者：Erling O, Averbuch A, Larriba-Pey J, Chafi H, Gubichev A, Prat A, Pham M., Boncz, P.

3. 标题：LinkBench: a database benchmark based on the Facebook social graph

出处：Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 2013

作者：Armstrong TG, Ponnekanti V, Borthakur D, Callaghan M.

- **图数据库管理**

1. 标题：Modern Database Management

出版社: Pearson Higher Education

作者：Hoffer JA.

2. 标题：A Survey on Graph Database Management Techniques for Huge Unstructured Data

出处：International Journal of Electrical and Computer Engineering, 2018

作者：Patil N, Kiran P, Kiran N, KM NP.

3. 标题：A Survey of RDF Data Management Systems

出处：Frontiers of Computer Science, 2016

作者：Özsu MT.

5.4 国家自然科学基金支持情况

根据“图数据库”领域关键词，从 AMiner 数据库中查找出 2010 至 2020 年国家自然科学基金支持的图数据库相关项目（包含未结题的项目）。具体情况如下表所示，国家自然科学基金委共支持了 30 个与图数据库相关的项目，其中北京大学获得的支持项目数量最多（5 个），哈尔滨工业大学和中国人民大学各有 3 个支持项目。

图数据领域关键词包括：图数据库（Graph databases）、属性图（Property graphs）、资源描述框架（Resource Description Framework, RDF），图分析（Graph analysis）、ACID 事务属性（Atomicity, Consistency, Isolation, Durability, ACID transaction）、图匹配（Graph patterns）。

表 5-1 国家自然科学基金支持情况

项目类别（个数）	依托单位	项目个数
面上项目（13 个）	北京大学	3
	东北大学	1
	东南大学	1

项目类别（个数）	依托单位	项目个数
	哈尔滨工业大学	2
	华东师范大学	1
	南京大学	1
	天津大学	1
	中国人民大学	3
青年科学基金项目（15 个）	北京大学	1
	东南大学	1
	复旦大学	1
	哈尔滨工业大学	1
	深圳大学	1
	天津大学	1
	武汉大学	1
	西安理工大学	1
	香港浸会大学深圳研究院	1
	浙江大学	2
	中北大学	1
	中国科学院计算技术研究所	1
	中国科学院深圳先进技术研究院	1
	中国人民解放军国防科学技术大学	1
应急管理项目（1 个）	西安电子科技大学	1
优秀青年科学基金项目（1 个）	北京大学	1

5.5 专利趋势

根据“图数据库”领域关键词，从 AMiner 数据库中搜索 2000 年至 2019 年图数据库相关专利在全球范围内的申请情况。领域关键词包括：图数据库（Graph databases）、属性图（Property graphs）、资源描述框架（Resource Description Framework, RDF）、图分析（Graph analysis）、ACID 事务属性（Atomicity, Consistency, Isolation, Durability, ACID transaction）、图匹配（Graph patterns）。

从增长趋势来看，2000 年至 2019 年，全球图数据库的有效专利为 1,458 项，呈现显著地上升态势。其中，2015 年的申请数量最高，达 150 项。

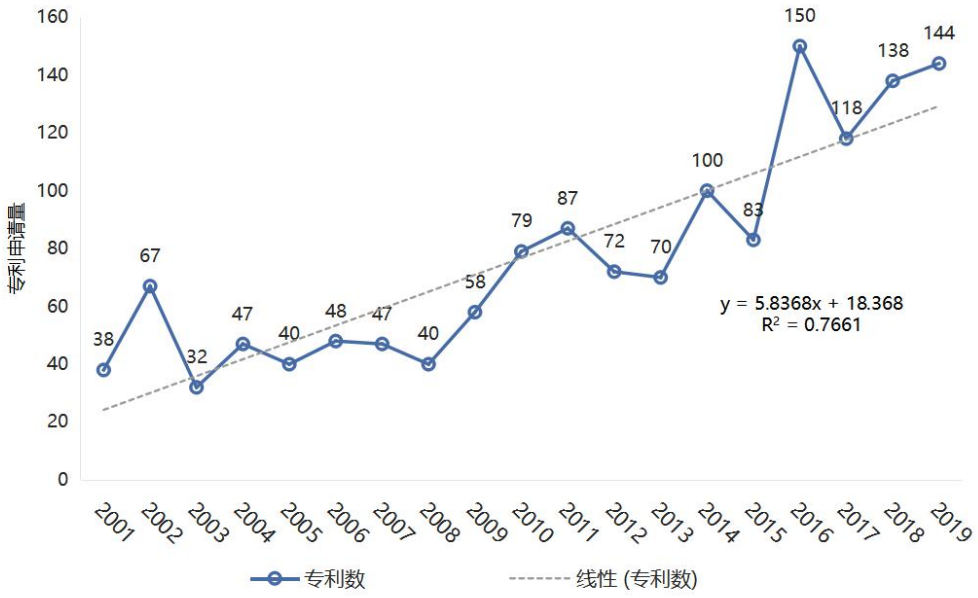


图 5-3 2000 年至 2019 年图数据库相关专利变化趋势

从国家层面来说，中国（579 项）、美国（405 项）和日本（118 项）是申请图数据库专利最多的三个国家。与美、日两国相比，中国的图数据库专利意识较晚，直至 2006 年以后申请量才开始明显上升，并且在 2013 年以后，一直处于领域专利申请量的领先地位。

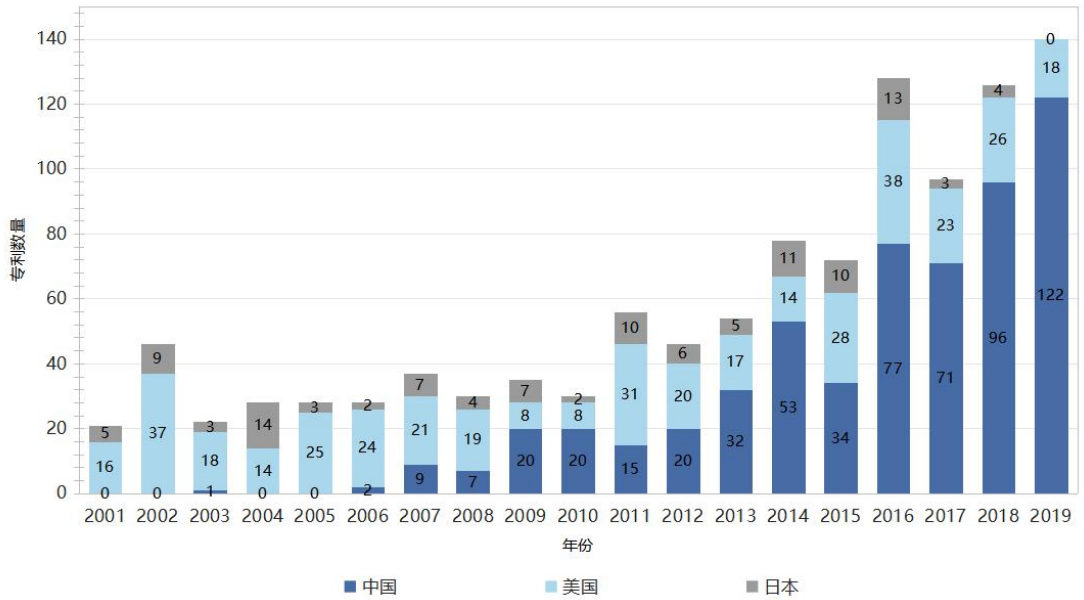


图 5-4 全球图数据库相关专利 TOP3 国家

从各省排名来看，当前申请人（专利权人）主要分布于北京、广东、江苏、上海等具有一定经济基础，科技投入度高的发达省市。

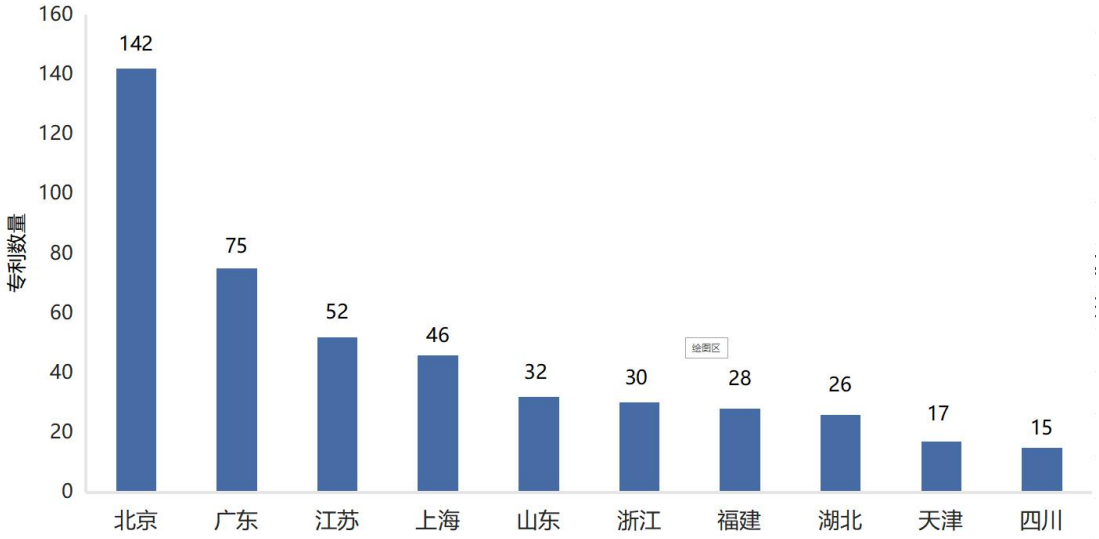


图 5- 5 中国图数据库相关专利各省排名

6 结语

图数据库作为存储和分析之间关系的系统，因其高性能、轻量级等优势，越来越受到业界关注，特别是在大数据处理方面，更是优势明显，被广泛应用于反欺诈、推荐引擎和知识图谱等多个场景。

目前，欧美国家的 Neo4j 和 ArangoDB 等数据库系统仍然是市场的主流。与欧美国家相比，中国人口众多，数据量巨大，相应的图数据库需求也更大。但是，中国的图数据库基础研究相对薄弱，缺少专业的数据处理人才，更是制约了图数据库的应用与发展。

因此，在发展层面，中国需要更专注产品的技术创新层面，打造更完全自主的图数据库；在服务层面，积极探索云计算模式的图数据库服务，打造基于云计算的图数据库；在应用方面，继续在图的可视化工具方面进行创新，通过拖拽的方式，自动生成查询语言、实时反馈结果，提升图数据库的应用便捷性。通过图数据库发现不同事物、数据之间的深度关联，获得更深层次的洞察。

参考文献

[1] 腾讯新闻. 一线 | 蚂蚁金服井贤栋：支付宝全球用户数超 12 亿[EB/OL]. 腾讯科技, 2019: 1. https://tech.qq.com/a/20190924/008583.htm.

[2] 张俊玲,王秀英,籍淑丽,等. 数据库原理与应用[M]. 北京: 清华大学出版社, 2007: 345-350.

[3] Gartner. An Introduction to Graph Data Stores and Applicable Use Cases[EB/OL], 2019. https://www.gartner.com/en/documents/3899263.

[4] Neo4j[EB/OL]. https://neo4j.com/.

[5] Db-engines. Popularity changes per category, April 2020[EB/OL]. https://db-engines.com/en/ranking_categories.

[6] Maciej Besta, Emanuel Peter, Robert Gerstenberger, Marc Fischer, Michał Podstawski, Claude Barthels, Gustavo Alonso, Torsten Hoefler. Demystifying Graph Databases: Analysis and Taxonomy of Data Organization, System Designs, and Graph Queries[EB/OL]. Arxiv Preprint Arxiv:1910.09017.

[7] 资源描述框架[EB/OL]. https://zh.wikipedia.org/wiki/資源描述框架.

[8] Luo Lannan,Sun Gang,Yu Wei. A Distributed Storage Access System for Mass Data using 3-tier Architecture[C]//Proceedings of 2011 International Conference on Computer Science and Information Technology(ICCSIT 2011): IACSIT Press, 2011: 292-300.

[9] George. L. HBase: the Definitive Guide: Random Access to Your Planet-Size Data[J]. O'Reilly Media, Inc., 2011.

[10] Hartig O. Reconciliation of Rdf and Property Graphs[J]. Arxiv Preprint Arxiv:1409.3288, 2014, 0(0): 1-18.

[11] 王鑫,邹磊,王朝坤,等. 知识图谱数据管理研究综述[J]. 软件学报, 2019, 30(7): 2139-2174.

[12] RDF 1.1 Concepts and Abstract Syntax[EB/OL]. https://www.w3.org/TR/rdf11-concepts/.

[13] Brickley d. Guha-RV.. RDF Schema 1.1. W3C Recommendation[EB/OL]. United States: W3C, 2018. https://www.w3.org/TR/rdf-schema/.

[14] W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview. 2nd ed.[EB/OL]. https://www.w3.org/TR/owl2-overview/.

[15] Harris S,Seaborne A,Prud'hommeaux E. Sparql 1.1 Query Language[J]. W 3c Recommendation, 2013, 21(10): 778.

[16] Francis N,Green A,Guagliardo P, et al. Cypher: an Evolving Query Langu age for Property Graphs[C]//Proceedings of the 2018 International Conferen ce on Management of Data, 2018: 1433-1445.

[17] Apache TinkerPop. TinkerPop3 Documentation v.3.3.3.[EB/OL](2020-1-1) [2020-4-1]. http://tinkerpop.apache.org/docs/3.3.3/reference/.

[18] Van Rest O,Hong S,Kim J, et al. Pgql: a Property Graph Query Languag e[C]//Proceedings of the Fourth International Workshop on Graph Data Ma nagement Experiences and Systems, 2016: 1-6.

[19] Angles R,Arenas M,Barceló P, et al. G-core: a Core for Future Graph Qu ery Languages[C]//Proceedings of the 2018 International Conference on Ma nagement of Data, 2018: 1421-1432.

[20] Neumann. T and Weikum. G. The RDF-3X engine for scalable manageme nt of RDF data[J]. VLDB J., 19(1):91–113,2010.

[21] Virtuoso[EB/OL]. https://virtuoso.openlinksw.com/.

[22] Martínez-Bazan. N, Muntés-Mulero. V, Gómez-Villamor. S, ÁguilaLorente. M, Dominguez-Sal. D, and Larriba-Pey. J.-L. Efficient Graph Management Based O n Bitmap Indices[J]. In IDEAS, pages 110–119, 2012.

[23] ArangoDB: Index Free Adjacency or Hybrid Indexes for Graph Databases [EB/OL]. https://www.arangodb.com/2016/04/index-free-adjacency-hybrid-inde xes-graph-databases/.

[24] Davoudian A,Liu C,Liu M. A Survey on Nosql Stores[J]. Acm Computing Surveys, 2018, 51(2): 1-43.

[25] Iordanov B. Hypergraphdb: a Generalized Graph Database[C]//Web Age In formation Management, 2010: 25-36.

[26] Boag. S, Chamberlin. D, Fernández. M. F, Florescu. D, Robie. J, Siméon. J, and Stefanescu. M. Xquery 1.0: An xml query language[J]. 2002.

[27] MongoDB[EB/OL]. https://www.mongodb.com/.

[28] OrientDB[EB/OL]. https://orientdb.com.

[29] Titan Data Model[EB/OL]. http://s3.thinkaurelius.com/docs/titan/1.0.0/data-m odel.html.

[30] JanusGraph[EB/OL]. http://janusgraph.org/.

[31] Armstrong TG,Ponnekanti V,Borthakur D, et al. Linkbench: a Database Benchmark Based on the Facebook Social Graph[C]//Proceedings of the 2013 Acm Sigmod International Conference on Management of Data, 2013: 1185-1196.

[32] Barahmand S,Ghandeharizadeh S. Bg: a Benchmark to Evaluate Interactive Social Networking Actions.[C]//Cidr: Citeseer, 2013: 1-28.

[33] Jouili S,Vansteenberghe V. An Empirical Comparison of Graph Databases [C]//2013 International Conference on Social Computing: IEEE, 2013: 708-715.

[34] Ciglan M,Averbuch A,Hluchy L. Benchmarking Traversal Operations Over Graph Databases[C]//2012 IEEE 28th International Conference on Data Engineering Workshops: IEEE, 2012: 186-189.

[35] Dominguez-sal D,Urbón-bayes P,Giménez-vanó A, et al. Survey of Graph Database Performance on the Hpc Scalable Graph Analysis Benchmark[C]// International Conference on Web-age Information Management: Springer, 2010: 37-48.

[36] Eberle W,Graves J,Holder L. Insider Threat Detection Using a Graph-based Approach[J]. Journal of Applied Security Research, 2010, 6(1): 32-81.

[37] Capotă M,Hegeman T,Iosup A, et al. Proceedings of the Grades'15[M], 2015: 1-6.

[38] Mccoll RC,Ediger D,Poovey J, et al. A Performance Evaluation of Open Source Graph Databases[C]//Proceedings of the First Workshop on Parallel Programming for Analytics Applications, 2014: 11-18.

[39] Malewicz G,Austern MH,Bik AJ, et al. Pregel: a System for Large-scale Graph Processing[C]//Proceedings of the 2010 Acm Sigmod International Conference on Management of Data, 2010: 135-146.

[40] Gonzalez J E, Xin R S, Dave A, et al. Graphx: Graph processing in a distributed dataflow framework[C]//11th Symposium on Operating Systems Design and Implementation ({OSDI} 14). 2014: 599-613.

[41] Zhu X, Chen W, Zheng W, et al. Gemini: A computation-centric distributed graph processing system[C]//12th Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016: 301-316.

[42] Cheng J,Yu JX,Ding B, et al. Fast Graph Pattern Matching[C]//2008 IEEE 24th International Conference on Data Engineering: IEEE, 2008: 913-922.

[43] Singh K,Singh V. Graph Pattern Matching: a Brief Survey of Challenges and Research Directions[C]//2016 3rd International Conference on Computing for Sustainable Global Development (indiacom): IEEE, 2016: 199-204.

[44] Dijkstra EW. A Note on Two Problems in Connexion with Graphs[J]. Numerische Mathematik, 1959, 1(1): 269-271.

[45] Dantzig G,Fulkerson DR. On the Max Flow Min Cut Theorem of Networks[J]. Linear Inequalities and Related Systems, 2003, 38(1): 225-231.

[46] Kruskal JB. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem[J]. Proceedings of the American Mathematical Society, 1956, 7(1): 48-50.

[47] Page L,Brin S,Motwani R, et al. The Pagerank Citation Ranking: Bringing Order to the Web.[R]: Stanford Infolab, 1999.

[48] Green. A, Junghanns. M, Kießling. M, Lindaaker. T, Plantikow. S, and Selmer. P. opencypher: New directions in property graph querying[J]. In EDBT, 2018: 520–523.

[49] RedisGraph[EB/OL]. https://oss.redislabs.com/redisgraph/.

[50] AgensGraph[EB/OL]. https://bitnine.net/agensgraph-2/.

[51] TuGraph[EB/OL]. https://fma-ai.cn/product.

[52] Groovy[EB/OL]. https://groovy-lang.org/.

[53] Infinitegraph[EB/OL]. https://www.objectivity.com/products/infinitegraph/.

[54] Azure Cosmos DB[EB/OL]. https://azure.microsoft.com/en-us/services/cosmos-db/.

[55] DSE Graph (DataStax) [EB/OL]. https://www.datastax.com/

[56] Amazon Neptune[EB/OL]. https://aws.amazon.com/neptune/.

[57] TigerGraph[EB/OL]. https://www.tigergraph.com/.

版权声明

AMiner 研究报告版权为 AMiner 团队独家所有，拥有唯一著作权。AMiner 咨询产品是 AMiner 团队的研究与统计成果，其性质是供用户内部参考的资料。

AMiner研究报告提供给订阅用户使用，仅限于用户内部使用。未获得AMiner团队授权，任何人和单位不得以任何方式在任何媒体上（包括互联网）公开发布、复制，且不得以任何方式将研究报告的内容提供给其他单位或个人使用。如引用、刊发，需注明出处为“AMiner.org”，且不得对本报告进行有悖原意的删节与修改。

AMiner 研究报告是基于 AMiner 团队及其研究员认可的研究资料，所有资料源自 AMiner 后台程序对大数据的自动分析得到，本研究报告仅作为参考，AMiner 团队不保证所分析得到的准确性和完整性，也不承担任何投资者因使用本产品与服务而产生的任何责任。

顾 问：陈文光、李涓子
编 辑：叶静芸、林 恒、刘强强、朱晓伟
数 据：赵慧军

关注“学术头条”并回复“图数据库”下载报告