O'REILLY®

第2版

图数据库

Graph Databases

[美] Ian Robinson Jim Webber Emil Eifrem 著
刘璐 梁越 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

图数据库

（第2版）

Ian Robinson
[美] *Jim Webber* 著
Emil Eifrem
刘璐　梁越　译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北京

图书在版编目（CIP）数据

图数据库 /（美）罗宾逊（Robinson, I.），（美）韦伯（Webber, J.），（美）艾弗雷姆（Eifrem, E.）著；刘璐，梁越译. -- 2版. -- 北京 : 人民邮电出版社，2016.7

书名原文: Graph Databases, Second Edition

ISBN 978-7-115-41856-2

Ⅰ. ①图… Ⅱ. ①罗… ②韦… ③艾… ④刘… ⑤梁… Ⅲ. ①图象数据库 Ⅳ. ①TP311.132

中国版本图书馆CIP数据核字(2016)第041182号

◆ 著 ［美］Ian Robinson Jim Webber Emil Eifrem

译 刘 璐 梁 越

责任编辑 杨海玲

责任印制 张佳莹 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号

邮编 100164 电子邮件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京鑫正大印刷有限公司印刷

◆ 开本：787×1000 1/16

印张：12

字数：245 千字 2016 年 7 月第 2 版

印数：3 001 – 5 500 册 2016 年 7 月北京第 1 次印刷

著作权合同登记号 图字：01-2015-7473 号

定价：49.00 元

读者服务热线：(010)81055410 印装质量热线：(010)81055316

反盗版热线：(010)81055315

内容提要

本书系统地介绍了图数据库的历史由来、建模方法、工作原理和一些真实的用户用例，详细地说明了图数据解决的是什么样的问题，并以 Neo4j 数据库和 Cypher 查询语言为例，阐述了图数据库的建模方法和领域用例，最后还介绍了图数据库的工作原理以及一些实用的图论算法。

本书适合开发人员和数据库管理人员了解和学习图数据库时阅读，作为一门新的知识和独特的数据库领域来拓宽视野，也适合提供解决方案的负责人了解行业动向和新的解决问题的方式。通过阅读本书，读者可以对图数据库这一领域有一个透彻的了解。

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了Make杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——Wired

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——CRN

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

序

图正在吞噬世界，其趋势已无法逆转

自从 3 年前开始写《图数据库》以来，我们的业界已经见证了审视数据资产方式的根本性变化。

数据，应该出现在创新的层面，然而在过去的几十年里却只表现出一小部分潜力，其主要原因是可供我们使用的技术迫使我们只把它当成一般意义上的孤岛，除此以外别无他用。然而，图和图数据库彻底颠覆了这一切。

随着对关联数据转化能力的不断发掘，这些行业中具有超前意识的领导者们正在从他们的对手那里赢得不可逆转的优势。图无处不在，它们正吞噬着世界，其趋势已无法逆转。

正如我在第 1 版的序中所写，视野的变化大概在 20 年前就开始了，即早期的 Web 搜索创业公司挑战主导市场的领导者们（AltaVista、Lycos、Excite）的时候，他们简单的算法应用程序就已经证明 Web 文档是互相关联的了。

如今，Google 统治着 Web 搜索领域。紧随其后，其他的行业领军者也开始自我检视："如果我们带着数据之间的联系和连接，并根据它们重新构想我们的业务会是怎样？"如今这些问题的答案在我们的线上生活中无处不在，如 Facebook、Twitter 及其他类似的公司。

从曾经的专业领域里发现关联数据并从中挖掘内在机会已经变成一项商品化的技术。过去 3 年里，世界领先的图数据库的功能、可用性和性能都已经成熟了很多，业界对图数据库的了解和采纳都比我们所预想的更加广泛、深入和快速，将图数据库引入曾经面向离散数据的领域带来的创造性的和不可逆转的影响，每每都鼓舞和挑战了市场。

2011 年，我们曾认为采用图数据库的垂直领域主要会是软件、金融服务和通信行业；如今事实上也大致如此。但是令我们更加震惊的却是这 3 个领域以外对图数据的应用。

我们看到图渐渐渗透到一个又一个行业。在每个案例里，对图技术的采纳最终都产生了更好的产品和更出色的用户体验。诸如 Pitney Bowes、eBay 和 Cisco 这样的公司通过图技术来解决其最关键的问题，并促使其竞争者追赶他们或被逐出市场。如今最顶尖的 10 个全球零售商中有 4 个使用了 Neo4j。在他们身后，那些还没有适应形势的对手还在挣扎着去追赶，就是因为他们没有采用。

图数据库在开疆拓土和颠覆行业方面的能力，在新兴的物联网（IoT）领域表现得尤为明显。这个领域也许更适合被称为关联的物联网，因为若是没有关联，物联网就无从谈起。一旦有了很多关联的事物，你就有了一个基于图的问题。

近几年，某大型通信设备供应商用一个产品进入了物联网领域，该产品内嵌一个大规模通信网络，嗅探网络流量并构建网络中所有连接设备的模型。如果某一类设备同时闪红灯告警，通过该模型可以轻易地确定它们是真的同时发生了故障，还是它们全部连接到同一个防火墙然后停电了。当你从关联的视角来看物联网，图数据库可以帮你做这种实时预测分析。

图数据的解决方案能够如此迅速地研发并上线的主要原因在于它底层技术的重大变化。2013 年，Neo4j 的 2.0 版本问世，标志着产品在功能、性能和可用性方面的重大改变。除了一个全新的可视化工具，Neo4j 2.0 还带来了一个增强的数据模型，它的主要功能、标签、选择性约束和声明式索引，加上对 Cypher 查询语言大量改进，都使得设计和开发图数据库应用比以往更加便捷和直观。

随着图数据库技术的成熟，社区力量也有了惊人的成长。据 db-engines.com 的数据，自 2013 年起，图数据库已成为发展最快的数据库类型。大数据是技术领域里发展最快最热门的技术，图数据库和它的发展绝对相关。图正在吞噬世界，而且已无法逆转。

我希望新版的《图数据库》能给正在成长的图技术领域带来一次更新（或是一个起点），并希望这本书能启发你在下个项目中开始使用图数据库，要是你已经投入了图的怀抱，希望它可以带给你更多更惊艳的方式来应用这一技术。

——Emil Eifrem

Neo4j 的联合创始人兼 Neo Technology 公司 CEO

2015 年 5 月于英国伦敦

前言

图数据库应对的是当今一个宏观商业世界的大趋势：凭借高度关联的数据中复杂而动态的联系获得洞察力并赢得竞争优势。无论我们想了解的是客户之间的联系，电话或数据中心网络元素之间的联系，娱乐产品制作者和消费者之间的联系，还是基因和蛋白质之间的联系，都会涉及大量的高度关联的数据。这些数据又会构成庞大的图，而理解和分析这些图的能力将成为公司在未来 10 年的核心竞争力。

对于任何达到一定规模或价值的数据，图数据库都是呈现和查询这些关联数据的最好方式。关联数据是这样的一种数据：它需要我们首先理解它的组成元素之间的关联方式。为了理解这个，很多时候我们需要去给这些事物之间的关联加以命名和限定。

尽管在一段时间以前，一些大公司就已经意识到这个问题并着手开发他们自己的图处理技术，但我们正处在一个技术全民化的时代。现如今，通用的图数据库已经成为现实，主流用户不必去投资建设自己的图架构，就可以享受关联数据带来的好处。

这次图数据和图思考复兴的伟大之处正在于图论本身并不是一个新事物。自 18 世纪欧拉创建了图论以来，数学家、社会学家、人类学家和其他领域工作者一直在研究和完善图论。然而，图论和图思考在信息管理中的应用却是最近几年的事情。那个时候，图数据库已经在社交网络、主数据管理（master data management）、地理空间、推荐系统以及其他领域帮我们解决了许多重要问题。有两股力量驱动我们对图数据库日益关注：一股力量是那些获得巨大商业成功的公司，如 Facebook、Google 和 Twitter，他们都将自己的商业模式紧紧地围绕在他们专有的图技术上；另一股力量就是通用的图数据库开始进入到技术领域里。

关于第 2 版

本书的第 1 版写于 Neo4j 2.0 还在开发的过程中，那时标签、索引和约束的最终形式还没有完全确定下来。现在 Neo4j 已经进入了 2.x 时代（本书写作的时候是 2.2，2.3 也要出来了），我们可以自信地将图属性模型里的新元素落在纸上了。

至于本书的第 2 版，我们修订了所有的 Cypher 示例，这样可以引入 Cypher 所有的最新语法。我们在查询和图中都加入了标签，并且提供了对于 Cypher 声明式索引和可选限制的解释。在其他地方我们还加入了额外的建模指导，引入了关于 Neo4j 最新的内部

架构变化的说明，并且更新了使用最新测试工具的测试案例。

关于本书

本书的目的是为技术实践者介绍图和图数据库，技术实践者包括开发人员、数据库专业人士和技术决策者。阅读本书将会让你对图数据库有一个贴近实际的理解。我们将演示图模型如何“塑造”数据，以及如何用图数据库查询、推断、理解和处理数据。我们讨论了多种适合用图数据库处理的问题，配以从实际应用中提取的用户案例，还将展示如何规划和实施一个图数据库解决方案。

本书的排版约定

本书使用下列排版约定。

中文楷体

用于新术语、文件名称以及文件扩展名。

等宽字体（constant width）

用于程序代码，包括文字段落中引用的程序元素，如变量、函数名、数据库、数据类型、环境变量、代码语句和关键字。

加粗等宽字体（**constant width bold**）

用于命令以及其他需要用户逐字输入的文字。

等宽斜体（*constant width italic*）

应该由用户提供的值或根据上下文确定的值。

此图标表示一个提示、建议或一般性注释。

此图标表示警告或慎用。

代码示例的使用

补充材料（代码示例、练习等）可以从 https://github.com/iansrobinson/graph-databases-use-cases 下载。

这本书的目的是帮助你完成工作。一般来说，如果代码示例是本书中提供的，你可以在你的程序和文档中使用。你不需要联系我们申请权限，除非你要直接复制相当大的一部分代码。例如，在编写程序的过程中使用了本书中的几段代码，这不需要授权。售卖或者分发 O'Reilly 的图书示例光盘显然是需要授权的。引用本书或引用示例代码来回答问题是不需要授权的，但将本书的大量示例代码纳入产品的文档是需要授权的。

我们对你在使用时声明引用信息表示感谢，但并不做强制要求。引用信息通常包括书名、作者、出版社和 ISBN，如“*Graph Databases* by Ian Robinson, Jim Webber, and Emil Eifrem (O'Reilly). Copyright 2015 Neo Technoogy, Inc., 978-1-491-93089-2”。

如果你认为对示例代码的使用需要授权，请通过这个邮箱联系我们：permissions@oreilly.com。

联系我们

如果你想就本书发表评论或有任何疑问，敬请联系出版社。

美国：

O'Reilly Media Inc.

1005 Gravenstein Highway North

Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）

奥莱利技术咨询（北京）有限公司

我们为本书提供了专门的网页，上面有勘误表、示例，以及其他额外的信息，可以通过 http://bit.ly/graph-databases-2e 访问该网页。

为本书提供建议和咨询技术问题，请发送邮件到 bookquestions@oreilly.com。

想了解更多关于我们的书籍、课程、会议，以及新闻等信息，请登录我们的网站：http://www.oreilly.com。

我们的其他联系方式如下。

Facebook: http://facebook.com/oreilly

Twitter: http://twitter.com/oreillymedia

YouTube:http://www.youtube.com/oreillymedia

致谢

我们要感谢本书的技术审稿人 Michael Hunger、Colin Jack、Mark Needham 和 Pramod Sadalage。

我们赞赏并感谢本书第 1 版的编辑 Nathan Jepson。

在本书成书的过程中，Neo Technology 的同事极大地贡献了他们的时间、经验和努力。特别感谢 Anders Nawroth 对本书的工具组提供了宝贵的帮助，感谢 Andrés Taylor 对 Cypher 部分提供的热情帮助，感谢 Philip Rathle 对行文提供的建议和所做的贡献。

感谢 Neo4j 社区的所有人，感谢你们这些年对图数据库做出的许多贡献。

还要特别感谢我们的家人 Lottie、Tiger、Elliot、Kath、Billy、Madelene 和 Noomi，谢谢他们付出的爱和支持。

感谢 Cristine Escalante 和 Michael Hunger 的辛勤工作让本书的第 2 版得以出版，谢谢你们巨大的帮助。

目录

第 1 章

简介

虽然本书大部分内容是讨论图数据模型的，但这并不是一本关于图论的书。[①] 使用图数据库并不需要太多的理论知识：只要知道什么是图就够了。记住这一点，下面来大体回顾一下我们对图的认识。

1.1 图是什么

形式上，图不过是顶点和边的集合，或者说更简单一点儿，图就是一些节点和关联这些节点的联系的集合。图将实体表现为节点，实体与其他实体连接的方式表现为联系。我们可以用这个通用而富有表现力的结构来为各种场景建模，从宇宙火箭的建造到道路系统，从食物的供应链及原产地追踪到人们的病历，甚至更多。

无处不在的图

在我们了解科学、政府和商业领域的数据集广泛多样性的过程中，图起到了极大的作用。现实世界完全不同于关系型数据库背后的基于表的模型，它是丰富的且相互之间充满关联：有些部分是统一而规则的，而其他部分是特殊的、不规则的。一旦理解了图，你就会发现图无处不在。比如，Gartner 定义了商业世界的 5 个图——社交、意向、消费、兴趣和移动，并指出运用这些图的能力是一个“可持续的竞争优势”。

① 关于图论的介绍，请参考 Richard J. Trudeau 的 *Introduction to Graph Theory*（Dover，1993）和 Gary Chartrand 的 *Introductory Graph Theory*（Dover，1985）。如果想要了解图是怎样给复杂的时间和行为提供洞察力的，请参考 David Easley 和 Jon Kleinberg 的 *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*（Cambridge University Press，2010）。

就拿 Twitter 来说，它的数据很容易表示为一张图。在图 1-1 中，我们可以看到由 Twitter 用户组成的一个小型社交网络。每个节点都被标为 User，表明了他在这个网络中的角色。然后这些节点又用联系连接起来，帮助更好地建立语义上下文。也就是说，Billy 关注了（FOLLOWS）Harry，相应地，Harry 也关注了 Billy。Ruth 和 Harry 也是互相关注的，不过，尽管 Ruth 关注了 Billy，但 Billy 还没有关注她。

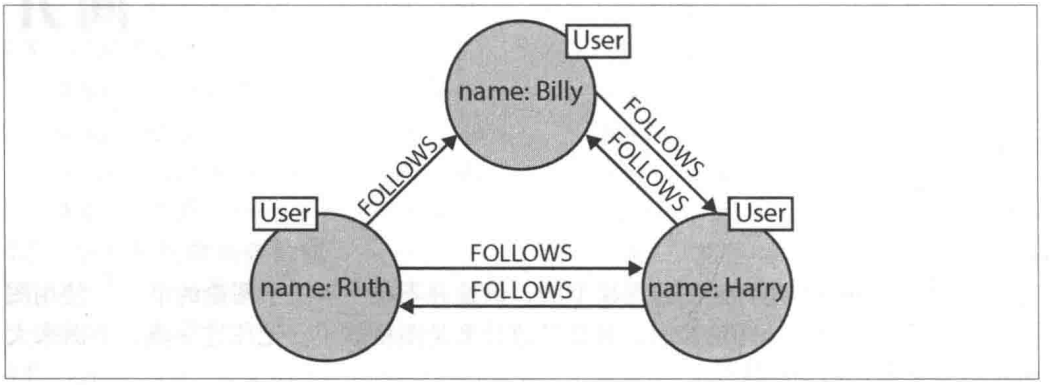

图 1-1　一个小型社交图

带标签的属性图模型

在讨论图 1-2 的过程中，我们也顺便提一下现在最流行的图模型形式——带标签的属性图（labeled property graph）（在附录 A 中，我们会更详细地讨论其他几种图数据模型）。带标签的属性图具有如下特征。

- 它包含节点和联系。
- 节点上有属性（键值对）。
- 节点可以有一个或多个标签。
- 联系有名字和方向，并总是有一个开始节点和一个结束节点。
- 联系也可以有属性。

对于大部分人来说，属性图模型是直观且容易理解的。不过简单归简单，它却可以描述绝大部分图的使用场景，并对我们的数据产生有价值的见解。

当然，实际的 Twitter 图比图 1-1 要大数亿倍，但它们的工作原理是一样的。在图 1-2 中，我们把 Ruth 发布的消息也包含到图里面来。

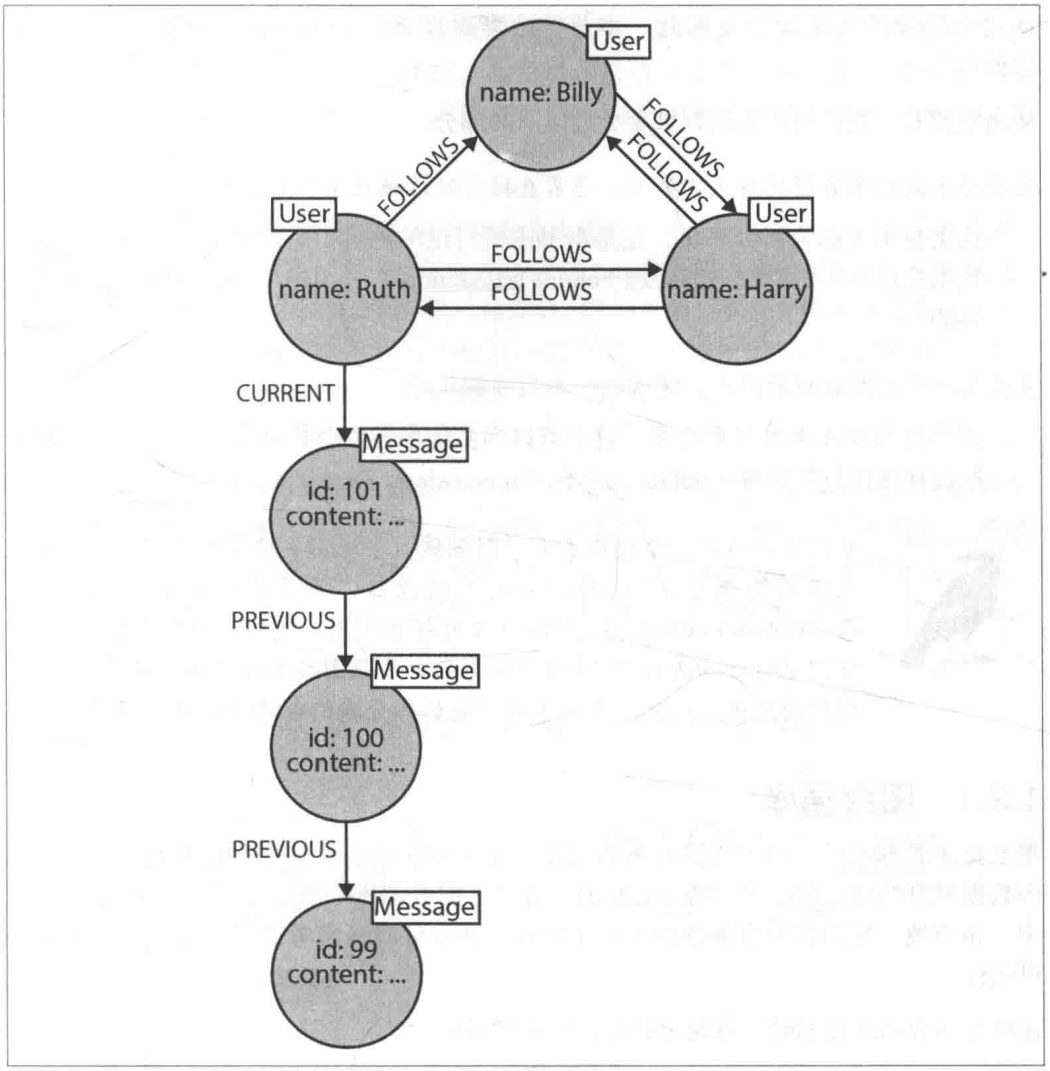

图 1-2　发布消息

尽管图很简单，但图 1-2 还是展示出了图模型的表现力。我们可以很容易从图中看出 Ruth 发布了一连串的消息。通过标记为 CURRENT 的联系可以找到她最新发布的消息，PREVIOUS 联系建立了 Ruth 的消息时间线。

1.2　图领域概览

近年来，无数用于管理、处理和分析图的项目和产品纷纷涌入市场。技术选择的陡增使我们难以跟进这些工具并摸清它们之间的区别，即便对我们这些一直活跃

在这个领域的人来说也是如此。本节的内容对理解新兴的图领域提供了一个“高空俯览”。

从高空俯看，我们可以将图领域划分成以下两部分。

主要用于联机事务图的持久化技术，通常直接实时地被应用程序访问

这类技术被称为图数据库，正是本书主要讨论的内容。它们和“常见的”关系型数据库世界中的联机事务处理（online transactional processing，OLTP）数据库是一样的。

主要用于离线图分析的技术，通常按一系列步骤执行

这类技术被称为图计算引擎。它们可以和其他大数据分析技术看做一类，如数据挖掘和联机分析处理（online analytical processing，OLAP）。

我们可以从另一个视角去划分图领域，去观察各种技术使用的图模型。主流的图模型有 3 种，分别是属性图、资源描述框架（Resource Description Framework，RDF）三元组和超图。我们将会在附录 A 中对它们进行详细的说明。市场上常见的大多数图数据库使用的都是属性图模型的变体，因此，在本书接下来的部分我们也将使用这一模型。

1.2.1 图数据库

图数据库管理系统（以下将简称图数据库）是一种在线的数据库管理系统，它支持对图数据模型的增、删、改、查（CRUD）方法。图数据库一般用于事务（OLTP）系统中。相应地，它们也对事务性能进行了优化，在设计时通常考虑了事务完整性和操作可用性。

在研究图数据库技术时，有两个特性需要多加考虑。

底层存储

一些图数据库使用原生图存储，这类存储是优化过的，并且是专门为了存储和管理图而设计的。然而并不是所有的图数据库使用的都是原生图存储，也有一些图数据库将图数据序列化，保存到关系型数据库或面向对象数据库，或是其他通用数据存储中。

处理引擎

一些定义要求图数据库使用免索引邻接，这意味着，关联节点在数据库里是物理意义上的“指向”彼此[①]。这里如果我们看的更宽泛些：站在用户的角度，任

① 参考 Rodriguez, Marko A.和 Peter Neubauer 的“The Graph Traversal Pattern”（2011）。*Graph Data Management: Techniques and Applications*, ed. Sherif Sakr and Eric Pardede, 29-46. Hershey, PA: IGI Global.