动手学图机器学习

[英] 亚历山德罗·内格罗(Alessandro Negro) 著

郭 涛 译

清華大學出版社

北 京

北京市版权局著作权合同登记号 图字：01-2023-6159

Alessandro Negro
Graph-Powered Machine Learning
EISBN: 978-1-61729-564-5

图书在版编目(CIP)数据

动手学图机器学习 / (英) 亚历山德罗 • 内格罗 (Alessandro Negro) 著；郭涛译. —北京：清华大学出版社，2024.5

书名原文：Graph-Powered Machine Learning

ISBN 978-7-302-66042-2

I. ①动… II. ①亚…②郭… III. ①机器学习 IV. ①TP181

中国国家版本馆 CIP 数据核字(2024)第 070796 号

责任编辑：王 军
装帧设计：孔祥峰
责任校对：成凤进
责任印制：杨 艳

出版发行：清华大学出版社

网 址：https://www.tup.com.cn，https://www.wqxuetang.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-83470000 邮 购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：河北鹏润印刷有限公司
经 销：全国新华书店
开 本：170mm×240mm **印 张**：25.25 **字 数**：606 千字
版 次：2024 年 6 月第 1 版 **印 次**：2024 年 6 月第 1 次印刷
定 价：128.00 元

产品编号：094822-01

译 者 序

机器学习正从学术界走向工业界，从理论走向应用。越来越多的机器学习相关项目纷纷落地，并产生了巨大的商业价值。这不仅要求从业者精通理论，能够提出新算法、新模型，还需要其具备很强的动手能力和工程实践能力。同时具备深厚的理论基础和丰富的工程实践经验的工程师目前在国内寥寥无几，仅华为、阿里和小米等互联网大厂才有这样的人才。机器学习工程需要工程师熟悉整个机器学习项目的生命周期，涉及数据收集和准备、特征工程、监督模型训练、模型评估、模型服务、监测和维护等方面。此外，还需要团队成员之间相互协作、紧密配合，才可以让整个项目顺利开展。

混合型机器学习是近几年来机器学习的发展趋势。单凭某个机器学习算法很难解决问题，需要将多个机器学习算法结合进行优势互补，才能解决特定问题。图深度学习、概率图模型、贝叶斯深度学习、元深度学习和图机器学习等是这几年新兴的混合机器学习算法。本书将图论和机器学习相结合，从机器学习工程实践的角度对图机器学习进行理解。本书研究主题包括机器学习工程项目生命周期、图数据工程、图数据的存储与管理、图机器学习应用和工程实践等，这些主题是实现图机器学习工程的强有力工具，在人工智能领域越来越重要。近几年，图机器学习已广泛用于知识图谱、推荐系统、推理与学习等领域，成为人工智能相关研究不可或缺的技术。

本书并不是一本纯粹介绍图机器学习理论的著作，Alessandro Negro 博士作为科学家和 Reco4 公司的 CEO，长期维护图数据源的推荐系统。他结合机器学习工程和图机器学习方法，通过推荐引擎、欺诈检测和知识图谱等案例，讲述了图机器学习工程实战。他以源代码为示例，逐步讲述其实现过程，以及如何更有效地管理图数据、实施算法、存储预测模型和可视化结果。本书适合作为数据科学家和数据科学从业者以及企业工程师的参考书。

本书内容涉及图数据工程、图数据库存储、图机器学习技术、图机器学习结果可视化，涵盖了整个软件工程的生命周期。建议读者借鉴这种思维模式，将这种工程思维模式迁移到其他机器学习项目实战中。另外，本书很好地将图机器学习算法和应用案例相结合，以核心代码为例进行讲解，如果读者要思考机器学习理论如何解决实际项目问题，本书值得借鉴。在现实中，往往很难用前沿技术来解决实际问题，机器学习项目也很难落地，本书在这两个方面有很多值得借鉴的意义。此外，从本书中也可发现，单一的算法或模型很难解决实际问题，往往要使用混合模型或者将多个机器学习算法相结合形成混合机器学习算法，本书在这方面也值得借鉴，例如，近几年新兴的概率图模型、图深度学习、贝叶斯深度学习、深度强化学习等混合机器学习算法，可以将本书的经验迁移

到实际的应用场景中。

本书在翻译过程中得到了很多人的帮助。电子科技大学外国语学院的史佳艳、尹思敏，吉林大学外国语学院的吴禹林和吉林财经大学外国语学院的张煜琪等参与了全书校对工作，他们在校对过程中不断地查阅资料，进一步对书中的翻译细节进行了修正，以期达到“信、达、雅”。在此，感谢他们为本书所做的大量工作。最后，感谢清华大学出版社编辑的包容和极大耐心，他们为保证本书的质量和顺利出版做了大量的编校工作，在此深表谢意。

由于本书涉及一定的广度和深度，加上译者水平有限，翻译过程中难免有不足之处，欢迎各位读者批评指正。

郭涛

2024 年 5 月于蓉城

译 者 简 介

郭涛，主要从事人工智能、智能计算、概率与统计学、现代软件工程等前沿交叉研究。翻译并出版过多部译作，包括《深度强化学习图解》《机器学习图解》和《概率图模型原理与应用(第 2 版)》。

推 荐 序

机器学习在科技领域中被广泛讨论，每天都有大量关于其应用和进展的文章涌现。但在业内，一场以图作为机器学习核心的革命正悄然酝酿。

经过近十年的实践后，Alessandro 将图和机器学习相结合，汇编成此书。如果 Alessandro 在某个网络巨头公司工作，他汇集了一大批博士专门研究特殊的一次性系统，那么本书对于他们来说将大有裨益；而对大部分人来说，本书并不是一本实用指南，只能满足我们的一些好奇心。幸运的是，虽然 Alessandro 确有博士学位，但他从事商业，并对企业构建的各种系统有着深刻的理解。由本书可知：Alessandro 巧妙地解决了在超大规模网络巨头之外构建现代系统时，软件工程师和数据专业人员必须规避的各种有关实际设计和实施的挑战。

《动手学图机器学习》展示了图对未来机器学习的重要性。本书不仅表明，图为推动当代 ML 流程提供了一种高级手段，还说明了图是组织、分析和处理机器学习数据的主要方式。本书为读者提供了内容丰富、条理清晰的图机器学习相关知识，每个主题都以详细的示例为基础，这些例子体现了 Alessandro 具有丰富的经验，以及他作为一名长期从业者，身上所具备的坚定的信心。

本书内容并不复杂，提供了一个整体框架来推理机器学习，并将其集成到我们的数据系统中。本书还提供了一种实用的推荐方法，它涵盖多种方法，如协同过滤、基于内容和会话的推荐以及混合风格。Alessandro 指出了现有技术中缺乏可解释性的问题，并表明这不是图方法存在的问题。接着，他引入了邻近性和社交网络分析等概念，解决了欺诈检测问题，从中通过犯罪网络重温“物以类聚、人以群分”这一谚语。最后，本书讨论了知识图谱：图技术能够使用文档并从中提取相关知识、消除模糊项以及处理模糊查询项。本书主题跨度广阔，但内容质量始终上乘。

整本书中，Alessandro 循序渐进地引导读者，从基础知识到高级概念逐一学习。借助于示例和配套代码，即使“凡夫俗子”也能快速理解示例，并根据个人需要进行调整。通读本书后，你将学会使用各种实用工具，如果你愿意，还可以处理一些细节。现在，你应该已准备好提取图特征，从而优化现有模型的性能，进而熟练地使用图。我坚信这将是一段美妙的旅程。

Jim Webber 博士，Neo4j 首席科学家

作 者 简 介

首先，我对计算机科学和数据研究充满热情。我专攻 NLP、推荐引擎、欺诈检测和图辅助搜索等研究方向。

在攻读计算机工程专业并从事该领域相关的多种工作后，我跨学科攻读了科学与技术博士学位。随着我对图数据库的兴趣达到顶峰，我成立了一家名为 Reco4 的公司，旨在支持开源项目 reco4j——第一个基于图数据源的推荐框架。

现在我是 GraphAware 的首席科学家，我们立志成为图技术研究领域的先驱。我们与 LinkedIn、世界经济论坛、欧洲航天局和美国银行等客户合作，将他们的数据转化为可搜索、可理解和可操作的知识，专注于帮助客户获得竞争优势。在过去的几年里，我一直领导着 Hume(知识图谱平台)的开发工作，并在世界各地的各种会议上进行宣讲。

致　谢

本书历经三年多的时间才得以出版。期间做了大量的工作——绝对比提出这个疯狂想法时我所预想的要多。同时，这也是迄今为止，我职业生涯中最激动人心的经历(是的，我正在筹备撰写第二本书)。我很享受这本书的创作过程，但该过程的确非常漫长。所以我要感谢这一路上帮助过我的人。

首先，我要感谢我的家人。在我通宵达旦工作时，独留妻子 Aurora 一人。在无数个周末，我忙于撰写本书，极少陪伴我的孩子们。感谢我的家人给予的理解和无条件的爱。

接下来，我要感谢 Manning 的策划编辑 Dustin Archibald。感谢你的配合，教会我关于写作的知识。特别感谢你在我月复一月延期交稿时依旧保持耐心。为了保证本书质量，你付出了大量努力，每个读者都将获益匪浅。还要感谢参与了本书出版工作的每位 Manning 员工，他们分工明确、能力出众，很荣幸能与他们共事。

致所有审稿人：Alex Ott、Alex Lucas、Amlan Chatterjee、Angelo Simone Scotto、Arnaud Castelltort、Arno Bastenhof、Dave Bechberger、Erik Sapper、Helen Mary Labao Barrameda、Joel Neely、Jose San Leandro Armendáriz、Kalyan Reddy、Kelvin Rawls、Koushik Vikram、Lawrence Nderu、Manish Jain、Odysseas Pentakalos、Richard Vaughan、Robert Diana、Rohit Mishra、Tom Heiman、Tomaz Bratanic、Venkata Marrapu 和 Vishwesh Ravi Shrimali，你们的建议为本书的定稿做出了重要贡献。

最后，如果没有 GraphAware，尤其是 Michal，就不会有这本书——不仅是因为他聘用了我并让我在如此出色的公司得到了成长，还因为我酒后告诉他我在考虑写一本书时，他鼓励我：“我觉得你就该这样做！”Chris 和 Luanne 从一开始就是我最忠实的支持者；KK 总是在关键时刻恰到好处地激励我；Claudia 帮助我检查了图片；我优秀的同事们与我进行了最有趣且最具挑战性的讨论。你们所有人都是本书成功付梓的重要力量！

关于封面图片

《动手学图机器学习》封面上的图片摘自 *Roman Tarantella Dancer with Tambourine and a Mandolin Player* (1865)。Anton Romako (1832—1889)是一位 19 世纪的奥地利画家，在一次旅行中他爱上了意大利，并于 1857 年移居罗马，开始绘画他在意大利的日常生活。这幅画中描绘了一个塔兰泰拉舞者，他一边跳舞一边演奏手鼓(还有一个演奏曼陀林的男人)。选择这幅图像是为了向作者本人在意大利南部的家乡，特别是南阿普利亚致敬，在那里塔兰泰拉仍然是一种经久不衰的传统舞蹈，表演者模拟被当地常见的一种称为“tarantula”的狼蛛(不要与现在常说的狼蛛混淆)咬伤后的样子。人们普遍认为这种蜘蛛带有剧毒，被它咬伤会导致出现疯狂状态，这种症状被称为狼蛛症。这种舞蹈的原始形式是无序的、自由流动的，以模拟中毒的状态。还有人认为人们发明这种舞蹈是因为连续不断的舞蹈动作会导致人体大量出汗，有助于将狼蛛的毒液排出体外。

无论这种舞蹈的真正起源是什么，其音乐和节奏都令人愉悦，无法抗拒。时至今日，仍能经常看到当地人表演这种舞蹈。

序　言

在我记忆中，2012 年的夏天是意大利南部最热的夏天之一。当时，我和妻子正等候着我们的第一个儿子降生，由于临近生产，我们很少外出，也没能享受阿普利亚清新干净的溪水。在这段时间里，你可以沉迷于 DIY(我并没有)，或者做一些具有挑战性的事情。因为我对数独游戏不太感兴趣，于是开展了一个仅在晚间和周末进行的项目：尝试构建一个通用的推荐引擎，该引擎可以服务于多个范围和场景，从小而简单的用户-条目交互数据集到复杂但清晰的数据集，最终包含相关的上下文信息。

就在这时，图(Graph)强行进入了我的生活。这种灵活的数据模型使我可以用相同方式存储用户的购买行为，还可以存储所有的推论信息(后文正式定义为上下文信息)以及生成的推荐模型。当时，Neo4j 1.x 刚刚发布。虽然那时它还没有 Cypher 和现在所具备的其他高级查询机制，但它足够稳定，可以作为我项目的主要图数据库。我利用图来解决项目中出现的难题，4 个月后，我发布了 reco4j 的 alpha 版本，这是史上第一个基于图的推荐引擎！

此后，我开始了一段真挚热烈的工作经历。我独自尝试，四处推广 reco4j 的理念，坚持了 3 年(说实话，并不是很成功)，直到我与 GraphAware(一家小型咨询公司，帮助许多公司成功完成了有关图的项目)的首席执行官 Michal Bachman 通了电话。几天后，我飞往伦敦，与其签订合同，成为该公司的第六名员工。此后，图便成了我生命中最重要的一部分(但当然，排在我的两个孩子之后)。

之后，图的生态系统发生了很大变化。越来越多的大公司开始采用图作为其核心技术，为客户提供高级服务或解决内部问题。GraphAware 取得了显著进展，我成了首席科学家，有机会利用图来帮助公司构建新服务并改进现有服务。图不仅能够解决传统问题——从基本的搜索工具到推荐引擎，从欺诈检测到信息检索——还能够作为重要技术手段，改进并增强机器学习项目。为了对自然连接数据和非连接数据执行不同类型的分析，网络科学和图算法提供了一些新工具。

从事咨询工作多年以来，每当与数据科学家和数据工程师交谈时，我发现了许多常见问题可以利用图模型或图算法解决。通过向人们展示处理机器学习项目的不同方式，积累了丰富经验，得益于此，我写下本书。图无法解决所有问题，但可以作为解决问题的一把利剑。通过学习本书，你也可以开启自己的美妙科研之旅。

关于本书

《动手学图机器学习》是一本有关在机器学习应用程序中如何有效使用图的实用指南，展示了构建完整解决方案的所有流程，其中，图发挥了关键作用。本书侧重于介绍与图相关的方法、算法和设计模式。根据作者在构建复杂机器学习应用程序方面的经验，本书提出了许多方法，假设其为食谱，那么图就是客户所得美味中的主要原料。在机器学习项目的整个生命周期中，此类方法非常有用，表现在多个方面，例如，更有效地管理数据源、实施更好的算法、存储预测模型以便更快地访问它以及更有效地可视化结果从而进一步分析数据。

本书读者对象

本书适合你吗？如果你是数据科学家或数据工程师，本书可以帮助你完成或开始你的学习之旅。如果你是经理，要启动或推动一个新的机器学习项目，本书可以帮助你为团队提出不同的观点。如果你是一位高级开发人员且有兴趣探索图的功能，本书可以帮助你以新视角理解图的作用，不仅可以将图作为一种数据库，还可以作为一种 AI 推动技术。

本书不是有关机器学习技术的笼统纲要，它侧重于介绍与图相关的方法、算法和设计模式，这是本书的突出主题。具体而言，本书重点介绍图方法如何帮助你开发和交付更优秀的机器学习项目。本书详细介绍了图模型技术，并描述了多种基于图的算法。对于最复杂的概念，将用具体的场景来进行说明，并为其设计了具体的应用程序。

本书旨在成为一本实用指南，帮助你在生产环境中安装应用程序以供使用。因此，本书描述了优化技术和启发式方法，以帮助你处理真实数据、真实问题和真实用户。本书不仅讨论了小型示例，还讨论了来自实际用例的端到端应用程序，并提供了一些处理具体问题的建议。

如果你对这些场景感兴趣，那么本书绝对是你的最佳选择。

本书结构

本书内容共 12 章，分为 4 个部分。第I部分介绍了本书的主题，从通用机器学习和图概念开始，然后了解结合这些概念的优势：

- 第 1 章介绍机器学习和图，并涵盖理解后续章节所需的基本概念。

- 第 2 章列出将大数据作为机器学习输入的主要挑战，并讨论如何使用图模型和图数据库来应对这些挑战。还介绍了图数据库的主要特征。
- 第 3 章详细描述图在机器学习工作流中的作用，以及一个用于大规模图处理的系统。

第Ⅱ部分讨论了几个真实用例，其中图促进了机器学习项目的发展并改进了最终结果，特别关注以下推荐方法：

- 第 4 章介绍最常见的推荐技术，并描述如何为其中一种技术设计合适的图模型：基于内容的推荐引擎。详细展示如何将现有(非图)数据集导入图模型并实现基于内容的推荐引擎以供使用。
- 第 5 章描述如何为协同过滤方法设计合适的图模型，以及如何充分实现协同过滤推荐引擎以供使用。
- 第 6 章介绍基于会话的推荐算法，并描述一个能够捕获用户会话数据的图模型。说明如何将样本数据集导入设计的模型，以及如何在其基础上实现真正的推荐引擎。
- 第 7 章介绍如何实现一个考虑用户上下文的推荐引擎。描述为上下文感知推荐引擎构建的图模型，并展示如何将现有数据集导入图模型。此外，还说明如何在单个引擎中组合多种推荐方法。

第Ⅲ部分介绍了欺诈检测：

- 第 8 章介绍欺诈检测，并描述不同领域中存在的不同类型的欺诈行为。还明确图对于建模数据的作用，从而更快、更容易地揭示欺诈行为，同时也指明一些用于打击欺诈的简单图模型采用的技术和算法。
- 第 9 章转向更高级的基于异常检测的反欺诈算法。展示如何使用图来存储和分析交易的 *k*-NN 并识别异常交易。
- 第 10 章介绍如何使用社交网络分析(Social Network Analysis，SNA)对欺诈者和欺诈风险进行分类。列出用于基于 SNA 进行欺诈分析的不同图算法，并展示如何从数据中导出正确的图。

第Ⅳ部分涵盖自然语言处理(Natural Language Processing，NLP)：

- 第 11 章介绍与基于图的 NLP 相关的概念。特别是，描述了一种简单方法：通过 NLP 提取非结构化数据的隐藏结构，来分解文本并将其存储在图中。
- 第 12 章介绍知识图谱，详细描述了如何从文本中提取实体和关系并从中创建知识图谱。列出与知识图谱共同使用的后处理技术，如语义网络构建和自动主题提取。

即使从头至尾通读本书可以最大限度地提高学习效果，但你不必如此。当遇到新挑战时，你都可以将本书用作参考书。对于本领域的初学者，我建议从前 3 章开始阅读，首先了解关键概念，然后跳到特定研究主题的章节。如果你对特定主题或应用程序感兴趣，最好从你感兴趣的部分开始：第 4 章(推荐方法)、第 8 章(欺诈检测)、第 11 章(自然语言处理)。如果你是图和机器学习方面的专家，只是想寻求建议，那么可以自行阅读感兴趣的章节。

关于代码、参考文献和彩图的下载

本书包含许多源代码示例，包括带有编号的代码清单和内嵌的普通代码示例。在这两种示例中，源代码都被格式化成宽度固定的字体，从而与普通文本进行区分。

许多情况下，源代码已被重新格式化；我们添加了换行符和重新设计的缩进，以适应书中可用的页面空间。在某些情况下，即使这样也还不够，代码清单还包括续行标记(➥)。此外，当在正文中对代码清单中的源代码进行描述时，经常会删除代码清单中的源代码注释。有些代码清单带有代码注释，用于突出重要的概念。

本书示例的源代码可以通过扫描本书封底的二维码下载。另外，各章与各附录所引用的参考文献、书中各图的彩图也可通过扫描本书封底的二维码下载。

目　录

第 I 部分　导论

第 1 章　机器学习和图：介绍 ……3
1.1　机器学习项目生命周期……5
1.1.1　业务理解……6
1.1.2　数据理解……6
1.1.3　数据预处理……7
1.1.4　建模……7
1.1.5　评估……8
1.1.6　部署……8
1.2　机器学习挑战……8
1.2.1　事实来源……8
1.2.2　性能……11
1.2.3　存储模型……11
1.2.4　即时性……12
1.3　图……12
1.3.1　什么是图……12
1.3.2　图作为网络模型……15
1.4　图在机器学习中的作用……20
1.4.1　数据管理……21
1.4.2　数据分析……21
1.4.3　数据可视化……22
1.5　本书心智模型……22
1.6　本章小结……23

第 2 章　图数据工程……24
2.1　处理大数据……26
2.1.1　数量……27
2.1.2　速度……29
2.1.3　多样性……31
2.1.4　真实性……32
2.2　大数据平台中的图……33
2.2.1　图对于大数据很有价值……34
2.2.2　图对于主数据管理意义重大……40
2.3　图数据库……44
2.3.1　图数据库管理……45
2.3.2　分片……47
2.3.3　复制……49
2.3.4　原生与非原生图数据库……51
2.3.5　标签属性图……55
2.4　本章小结……56

第 3 章　图在机器学习应用中的作用……58
3.1　机器学习工作流中的图……59
3.2　管理数据源……61
3.2.1　监控目标……64
3.2.2　检测欺诈……67
3.2.3　识别供应链中的风险……69
3.2.4　推荐条目……70
3.3　算法……76
3.3.1　识别供应链中的风险……76
3.3.2　在文档中查找关键词……78
3.3.3　监控目标……80
3.4　存储并访问机器学习模型……81
3.4.1　推荐条目……82
3.4.2　监控目标……84
3.5　可视化……87
3.6　剩余部分：深度学习和图神经网络……89
3.7　本章小结……91

第 II 部分　推荐

第 4 章　基于内容的推荐……97
4.1　表示条目特征……99

4.2 对用户进行建模……112
4.3 提供推荐……118
4.4 图方法的优点……137
4.5 本章小结……137

第 5 章 协同过滤……138
5.1 协同过滤推荐……141
5.2 为 User-Item 数据集创建二部图……142
5.3 计算最近邻网络……147
5.4 提供推荐……156
5.5 处理冷启动问题……161
5.6 图方法的优点……164
5.7 本章小结……165

第 6 章 基于会话的推荐……166
6.1 基于会话的方法……166
6.2 事件链和会话图……169
6.3 提供推荐……174
6.3.1 基于条目的 *k*-NN……175
6.3.2 基于会话的 *k*-NN……180
6.4 图方法的优点……185
6.5 本章小结……185

第 7 章 上下文感知和混合推荐……186
7.1 基于上下文的方法……186
7.1.1 表示上下文信息……189
7.1.2 提供推荐……193
7.1.3 图方法的优点……208
7.2 混合推荐引擎……209
7.2.1 多模型，单图……210
7.2.2 提供推荐……212
7.2.3 图方法的优点……214
7.3 本章小结……214

第Ⅲ部分 打击欺诈

第 8 章 图欺诈检测的基本方法……217
8.1 欺诈预防和检测……218
8.2 图在打击欺诈行为中的作用……222
8.3 铺垫：基本方法……229
8.3.1 寻找信用卡诈骗的源头……229
8.3.2 识别欺诈环……236
8.3.3 图方法的优点……242
8.4 本章小结……242

第 9 章 基于邻近算法……243
9.1 基于邻近算法：介绍……244
9.2 基于距离的方法……245
9.2.1 将交易存储为图……247
9.2.2 创建 *k* 最近邻图……248
9.2.3 识别欺诈交易……255
9.2.4 图方法的优点……263
9.3 本章小结……263

第 10 章 社交网络分析反欺诈……264
10.1 社交网络分析概念……266
10.2 基于分数的方法……269
10.2.1 邻域度量……272
10.2.2 中心性指标……278
10.2.3 集体推理算法……285
10.3 基于聚类的方法……289
10.4 图的优点……293
10.5 本章小结……294

第Ⅳ部分 用图训练文本

第 11 章 基于图的自然语言处理……297
11.1 一个基本方法：存储和访问单词序列……300
11.2 NLP 和图……309
11.3 本章小结……322

第 12 章 知识图谱……323
12.1 知识图谱：介绍……323
12.2 知识图谱构建：实体……327
12.3 知识图谱构建：关系……334
12.4 语义网络……341

12.5　无监督关键字提取……346
12.5.1　关键字共现图……353
12.5.2　聚类关键字和主题识别……354
12.6　图方法的优点……357
12.7　本章小结……357
附录 A　机器学习算法分类……359
附录 B　Neo4j……362
附录 C　处理图模式和工作流……374
附录 D　表示图……381

第 I 部分

导　论

我们被各种各样的图包围着。Meta、LinkedIn 和 Twitter 是最著名的社交网络示例，即由人构成的图。也存在其他类型的图，例如，电网、管道等。

图具有强大的结构，不仅可表示相关信息，还可用于多种类型分析。它们的简单数据模型由两个基本概念组成(如节点和关系)，十分灵活，足以存储复杂信息。如果你把属性存储在节点和关系中，则实际上可以表示任意大小的信息。

此外，在一个图中，每个节点和每个关系都是用于分析的一个接入点，并且可以无限地从一个接入点连接到其他点，这为多种访问模式和分析提供了可能性。

另一方面，机器学习提供了一些工具和技术，用于表示现实并提供预测。推荐就是一个很好的例子，该算法纳入已与用户相交互的内容，并能够预测他们可能感兴趣的内容。另一个例子是欺诈检测，它分析先前的交易(不论是否合法)并创建一个模型，该模型可以以较高的准确性识别新交易是否为欺诈。

我们表示训练数据和存储预测模型的方式几乎可以直接影响机器学习算法在准确性和速度方面的性能。算法预测的效果与训练数据集的效果一样好。如果想要使预测达到合理的信任水平，则必须进行数据清洗和特征选择。系统做出预测的速度会影响整个产品的可用性。假如一个推荐算法为在线零售商做出推荐时需要三分钟，那这期间用户会打开另一个页面，或者更糟的是，打开其竞争对手的网站。

图通过做自己最擅长的事情，即以易于理解和访问的方式来表示数据，从而支持机器学习。图可以使所有必要过程变得更快、更准确、更有效。此外，对于机器学习从业者来说，图算法是一种强大的工具：图社区检测算法可以帮助识别人群；PageRank 算法可以显示文本中最相关的关键字，等等。

若你并未完全理解这个导论中提及的一些术语和概念，那么本书的第I部分将为你提供进一步学习本书所需的全部知识。第I部分将图和机器学习相关的基本概念分为独立的实体和强大的组合进行介绍。祝你阅读愉快！

第1章

机器学习和图：介绍

本章内容

- 机器学习简介
- 图简介
- 图在机器学习应用中的作用

机器学习是人工智能的一个核心分支：它是使计算机程序可以从数据中学习的一个计算机科学研究领域。该词创造于 1959 年，当时 IBM 计算机科学家 Arthur Samuel 编写了第一个下棋的计算机程序[Samuel, 1959]。他有一个清晰的想法：

对计算机进行编程，使其从经验中学习，这最终将大幅度减少详细编程的工作量。

Samuel 根据一个固定公式给每个棋盘位置打分，以此来编写最初的程序。这个程序运行得很好，但在第二种方法中，他使程序与自身进行了数千场比赛，并使用结果来完善棋盘得分。最终，该程序达到了人类棋手的熟练程度，而机器学习也迈出了第一步。

一个实体 ——如人、动物、算法或通用计算机智能体[1]——正在学习，观察世界后，它是否能够在未来任务中表现更佳。换句话说，学习是将经验转化为专业技能或知识的过程[Shalev-Shwartz 和 Ben-David, 2014]。学习算法输入代表经验的训练数据，并生成专业技能作为输出。该输出可以是计算机程序、复杂的预测模型或内部变量的调整。性能的定义取决于待实现的特定算法或目标。一般来说，我们认为性能是预测与特定需求相匹配的程度。

我们用一个例子来描述这种学习过程：思考如何为电子邮件安装一个垃圾邮件过滤器。较为简单直接的解决方案是编写一个程序，使其记住所有被人类用户标记为垃圾邮件的电子邮件。当接收到新邮件时，伪代理会在之前的垃圾邮件中搜索相似的匹配项，如果

1 根据 Russell 和 Norvig[2009]，智能体是做出动作的某物(Agent 来自拉丁语 agere，表示“做”)。所有计算机程序都会执行某些任务，但计算机智能体应该能执行更多任务：自主运行、感知环境、长期坚持、适应变化、创造并追求目标。

找到匹配项，该邮件将被发送至垃圾文件夹；若未找到匹配项，该邮件将顺利通过过滤器。

这种方法可能有效，并且在某些情况下很有用。然而，这不是一个学习过程，因为它缺乏学习的一个重要方面：归纳能力，即将单个示例转换为模型的能力。在这个特定用例中，其意味着标记未预见的电子邮件的能力，即使它们与之前标记的电子邮件不同。这个过程也称为*归纳推理*[1]。总的来说，算法应该扫描训练数据并提取一组词，这组词若在某封电子邮件中出现，则表示其为垃圾邮件。然后，智能体将检查一封新的电子邮件是否包含一个或多个可疑词并相应地预测其标签。

如果你是一位经验丰富的开发人员，你可能会想，“当我可以指示计算机执行当前任务时，为什么还要编写一个学习自主编程的程序？”以垃圾邮件过滤器为例，可以编写一个程序来检查某些词是否出现，并在出现这些单词时将该电子邮件归类为垃圾邮件。但是这种方法有三个主要缺点：

- 开发人员无法预测所有可能的情况。在垃圾邮件过滤器用例中，无法对垃圾邮件中可能使用的所有词进行预测。
- 随着时间的推移，开发人员无法预测所有变化。垃圾邮件可以使用生词，也可以采用一些技巧来避免被直接识别，例如在字符之间添加连字符或空格。
- 有时，开发人员无法编写出能完成任务的程序。例如，尽管对人类来说识别朋友的脸很容易，但如果不使用机器学习，就无法通过编写软件来完成这项任务。

因此，当你遇到新问题或新任务，想用计算机程序解决时，以下问题可以帮助你决定是否使用机器学习：

- 具体任务是否过于复杂而无法对其进行编程？
- 在其整个过程中，任务是否需要具有某种泛化能力？

任何机器学习任务的关键在于训练数据，并基于此建立知识。无论使用的学习算法的潜在性能或质量如何，一开始就错误的数据将导致出现错误的结果。

本书旨在帮助数据科学家和数据工程师从两个方面处理机器学习过程：*学习算法*和*数据*。在这两个方面，我们都将使用图(目前定义为一组节点和连接节点的关系)作为有价值的心智模型和技术模型。许多基于数据并以图表示的学习算法可以提供有效的预测模型，而其他算法可以通过在工作流中使用以图表示的数据或图算法来得到改进。使用图还有许多其他好处：图是一种有价值的存储模型，用于表示来自流程输入的知识、管理训练数据和存储预测模型的输出，提供多种快速访问其自身的方法。本书将带领读者了解整个机器学习项目的生命周期，逐步展示可证明图具备价值和可靠性的所有案例。

但图并非适用于所有机器学习项目。在流分析中，必须处理数据流以找出短期异常，在此情况下以图的形式存储数据可能毫无用处。此外，其他算法所需数据的格式无法用图表示，无论是在训练期间还是用于模型存储和访问时都如此。本书旨在帮助读者辨别在机器学习过程中使用图是利还是弊。

---

1 根据斯坦福哲学百科全书网站(https://plato.stanford.edu/entries/logic-inductive)，归纳的逻辑是，前提应该在一定程度上支持结论。相比之下，在演绎推理中，前提在逻辑上包含结论。因此(尽管存在反对意见)，归纳有时被定义为从具体观察中推导出一般原则的过程。

1.1　机器学习项目生命周期

机器学习项目是一个人工参与的过程，也是一个软件项目。它涉及大量人员、大量沟通、大量工作及各项技能，并且需要使用定义明确的方法才能行之有效。首先，我们将通过明确的步骤和组成部分来对工作流进行界定，这个理念将贯穿全书。这里提出的心智模式是许多可能模式中的一种，它将帮助你更好地理解在成功的机器学习项目开发和部署过程中，图所具有的作用。

提供机器学习解决方案是一个复杂的过程，需要的不仅仅是选择正确的算法。此类项目包括许多相关任务[Sculley, 2015]：

- 选择数据源
- 收集数据
- 理解数据
- 清洗和转换数据
- 处理数据以创建 ML 模型
- 评估结果
- 部署

部署后，需要监控应用程序并对其进行微调。整个过程涉及多种工具、大量数据和不同人员。

数据挖掘项目最常用的流程之一是跨行业数据挖掘标准流程，或 CRISP-DM[Wirth and Hipp, 2000]。尽管 CRISP-DM 模型是为数据挖掘而设计的，但它也可应用于通用机器学习项目。作为基本工作流模型的一部分，CRISP-DM 的主要特点如下：

- 具有非专有性。
- 与应用、行业和工具无关。
- 从应用程序和技术的角度明确地看待数据分析过程。

此方法可用于项目规划和管理、交流和文档编制。

CRISP-DM 参考模型概述了机器学习项目的生命周期。在采用算法观点之前，该模式或心智模型有助于从数据角度处理机器学习项目，并为清晰的工作流定义提供基线。图 1.1 显示了该过程的六个阶段。值得注意的是，数据是这个过程的核心。

查看图 1.1，我们可以看到各个阶段的顺序是流动的。箭头仅表示各个阶段之间最重要和最频繁的依赖关系；在特定项目中，每个阶段的结果决定了下一步必须执行的阶段或阶段的某个特定任务。

外圈象征着流程的循环性质，部署解决方案时，该循环过程不会结束。后续的机器学习过程不仅可以从先前过程的经验中受益([Linoff and Berry 2011]的良性循环)，还可以从先前过程的结果中受益。接下来将更详细地展示每个阶段。

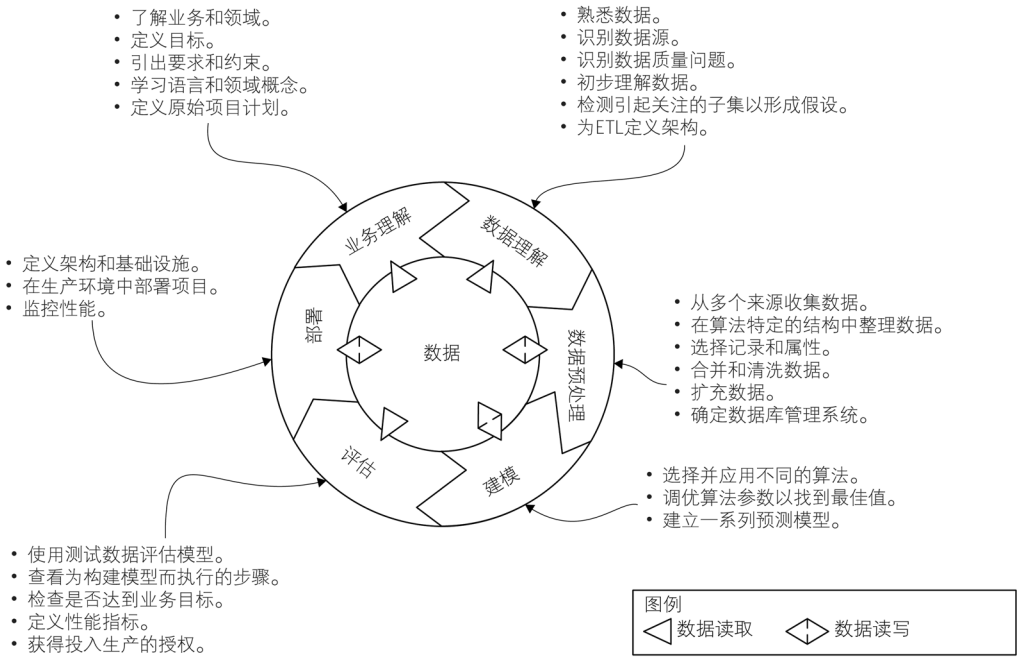

图 1.1 CRISP-DM 流程的六个阶段

1.1.1 业务理解

第一阶段需要定义机器学习项目的目标。这些目标通常都有特定表达：提高收入、改善用户体验、优化并定制搜索结果、增加产品销量等。要将这些高级问题定义转换为机器学习项目的具体要求和约束，就有必要了解业务及其所在领域。

机器学习项目是软件项目，在这个阶段，学习语言和领域概念也很重要。这些知识不仅有助于数据科学家和内部团队在后续阶段进行沟通，还可以提高文档的质量，优化结果的呈现效果。

这一阶段的结果如下：

- 清晰理解领域和业务。
- 定义目标、要求和约束。
- 将该知识转换为机器学习问题定义。
- 为实现目标，设计出初步、合理的项目计划。

第一次迭代的目标不应过于宽泛，因为这一轮需要进行大量工作，将机器学习过程运用到现有的基础设施中。同时，在设计第一次迭代时，牢记未来的扩展工作也很重要。

1.1.2 数据理解

在数据理解阶段，首先查询数据源，并从每个数据源收集一些数据，然后执行以下

步骤：

- 熟悉数据。
- 识别数据质量问题。
- 初步理解数据。
- 检测引起关注的子集以形成关于隐藏信息的假设。

理解数据需要对领域和业务有所了解。此外，查看数据有助于建立对领域和业务的理解，这就解释了为什么在这个阶段和前一个阶段之间存在反馈循环。

这一阶段的结果如下：

- 清楚了解可用的数据源。
- 清楚了解不同类型的数据及其内容(或至少清楚了解机器学习目标的所有重要部分)。
- 产生用于获取或提取此数据并将其提供给机器学习工作流中后续步骤的架构设计。

1.1.3　数据预处理

此阶段涵盖从多个来源收集数据以及以建模阶段算法要求的特定结构组织数据的所有活动。数据预处理任务包括记录和属性选择、特征工程、数据合并、数据清洗、新属性构建和现有数据扩充。如前所述，数据的质量对后续阶段的最终结果有着巨大的影响，因此这一阶段至关重要。

这一阶段的结果如下：

- 通过充分的设计技术得到一个或多个数据结构定义。
- 得到定义明确的数据流程，可为机器学习算法提供训练数据。
- 得到一组用于合并、清洗和扩充数据的过程。

这个阶段的另一个结果是确定了数据库管理系统，等待处理数据时，数据将被存储于该系统。

为完整起见，进一步处理数据之前并不一定需要一个显式的数据存储将数据持久化。可以在处理数据之前提取数据并进行转换。然而，实施这样一个中间步骤能够使数据在性能、质量以及进一步扩展方面具有更多优势。

1.1.4　建模

机器学习始于建模阶段。在此阶段，应选择和应用不同的算法，并将它们的参数校准到最佳值。这些算法用于构建一系列预测模型，在评估阶段完成后，从中选择最佳模型进行部署。有趣的是，一些算法会产生预测模型，而其他算法则不会[1]。

这一阶段的结果如下：

- 得到下一阶段要测试的算法集。
- 得到相关的预测模型(如果适用)。

1　附录 A(关于机器学习算法分类)包含一些用于创建预测模型的算法示例。

数据预处理和建模之间有着密切的联系，因为在建模过程中，你会经常发现数据问题，并产生构建新数据点的想法。此外，有些技术需要使用特定的数据格式。

1.1.5 评估

在机器学习项目的这个阶段，你已经构建了一个或多个高质量预测模型。在部署模型之前，必须对其进行全面评估，并审查建模所执行的所有步骤，以便可以确定它是否正确地实现了流程伊始定下的业务目标。

应以正式的方式进行评估，例如将可用数据划分为训练集(80%)和测试集(20%)。另一个主要目标是确定模型是否充分考虑到了重要业务问题。

这一阶段的结果如下：

- 得到可用于衡量性能的一组值(良好表现的具体衡量标准取决于算法类型和范围)。
- 全面评估业务目标是否实现。
- 得到在生产环境中使用解决方案的授权。

1.1.6 部署

因为构建机器学习模型是为了满足组织中的某些需求，所以模型的创建并不意味着项目的结束。根据不同需求，部署阶段可以像生成报告一样简单，也可以像发布一套为最终用户提供服务的完整基础设施一样复杂。在许多情况下，执行部署步骤的是客户(而不是数据科学家)。无论如何，重要的是预先了解需要执行哪些操作才能使用创建的模型。

此阶段的结果如下：

- 得到一份或多份包含预测模型结果的报告。
- 得到用于预测未来和支持决策的预测模型。
- 得到一个基础设施，为最终用户提供一组特定服务。

当项目投入生产时，有必要对其进行持续监控(例如，评估性能)。

1.2 机器学习挑战

机器学习项目存在一些内在挑战，这带来了工作难度。本节总结了你在构建新的机器学习项目时需要考虑的主要方面。

1.2.1 事实来源

CRISP-DM 模型从数据角度描述了整个机器学习的工作流，将数据置于机器学习过程的核心。训练数据就是事实的来源，可以从中获悉信息、进行预测。管理训练数据需要进行大量工作。引用华盛顿大学计算机科学教授 Jeffrey Heer 的话，“人们可以通过使用一种算法来处理原始数据并从中获取信息，这绝对是无稽之谈。”据估计，数据科学家将会

花费高达 80%的时间进行数据预处理[Lohr, 2014]。

在讨论算法细节之前，我经常使用下句将重点转移到数据上：

即使是用最佳学习算法，使用错误数据也会产生错误结果。

[Banko and Brill 2001]以及[Halevy, Norvig and Pereira 2009]的论文开创性地指出，对于复杂问题，数据往往比算法更重要。这两篇论文都考虑了自然语言处理，但这个概念可以泛化到一般的机器学习[Sculley, 2015]。

图 1.2 取自[Banko and Brill 2001]，显示了一组学习器的学习曲线，考虑了每个学习器在不同规模训练数据(多达 10 亿个词)上的平均性能。这里，使用何种算法并不重要；关键是，训练阶段的可用数据量增加后，学习器的性能也有所提升(如图所示)。这些结果表明，我们应该重新考虑将时间和金钱花在语料库开发上，而不是花在算法开发上。从另一个角度看，作为数据科学家，你可以专注于垂直维度——寻找更好的算法——但该图显示水平方向上还有更大的改进空间——收集更多数据。作为证明，图 1.2 中性能最差的算法在 1000 万个元素上的性能比性能最好的算法在 100 万个元素上的性能要好得多。

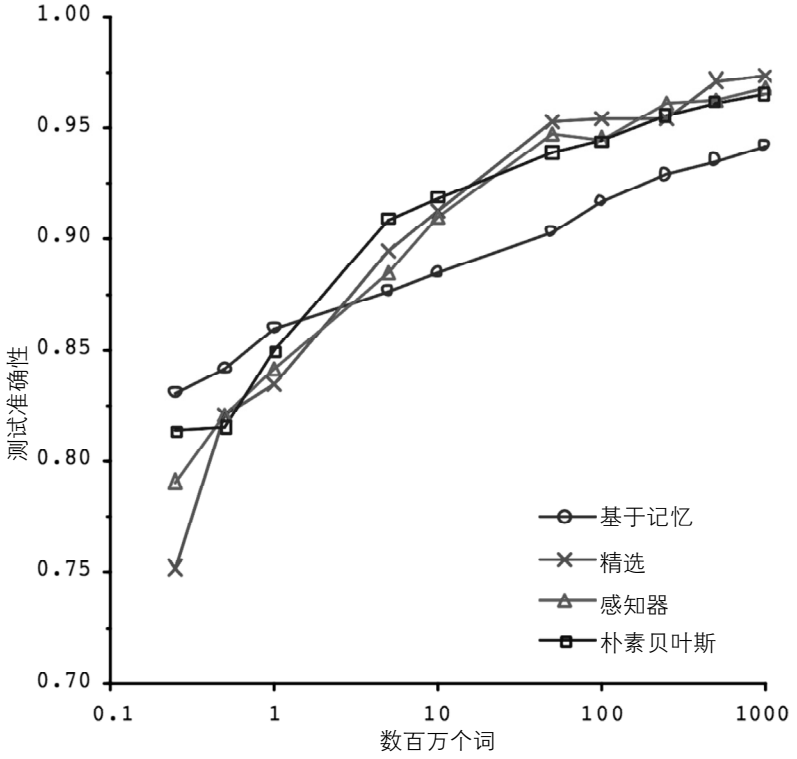

图 1.2　混淆集消歧的学习曲线

从多个来源收集数据不仅可以获得大量数据，还可以提高数据质量，解决诸如稀疏性、拼写错误、正确性等问题。从多种来源收集数据不成问题；我们生活在大数据时代，可以从网络、传感器、智能手机、企业数据库和开放数据源获得大量数字数据。但是，如果组

合不同的数据集切实可行，那么这种方法也会存在问题，因为不同来源的数据格式不同。在使用学习器对其进行分析之前，必须先对数据进行清洗、合并和归一化，形成统一的同构模式，这样算法才能理解数据。此外，对于许多问题来说，获得额外的训练数据需要付出代价，而对于监督学习而言，这种代价很大。

由于这些原因，数据成了机器学习过程中的第一大挑战。数据问题可以总结为以下四类：

- *数据量不足*——机器学习需要大量的训练数据才能正常工作。即使简单用例也需要数千个示例，而对于深度学习或非线性算法等复杂问题，你可能需要用到数百万个示例。
- *数据质量差*——数据源总是充满错误、异常值和噪声。较差的数据质量直接影响机器学习过程结果的质量，因为对许多算法来说，避开错误(不正确、不关联或不相关)的值，然后在混乱中检测潜在模式很难。
- *非代表性数据*——机器学习是一个归纳过程：模型根据它观察到的内容进行推断，并且可能排斥训练数据并未涉及的边缘情况。此外，如果训练数据充满噪声或仅与可能案例的一个子集相关，则学习器会产生偏差或过拟合训练数据，并且无法泛化到所有可能的案例中。对基于实例和基于模型的学习算法来说都是如此。
- *不相关特征*——如果数据包含一组较好的相关特征并且没有太多不相关的特征，算法将以正确的方式进行学习。尽管选择更多特征通常是一种有用策略，但为了提高模型的准确性，特征更多不一定更好。使用更多特征能让学习器得到从特征到目标更详细的映射，这增加了所计算的模型过拟合数据的风险。特征选择和特征提取是数据预处理过程中的两个重要任务。

为了克服这些问题，数据科学家必须从多个来源收集、合并数据，对其进行清洗，并使用外来数据进行扩充(此外，通常情况下，数据是为特定目的而准备的，但在此过程中，你会有新发现，而目的也会发生变化)。这些任务并不简单；它们不仅需要具备大量的专业技能，还需要一个数据管理平台，从而可以轻松地进行更改。

与训练示例质量相关的问题决定了机器学习项目基础设施的数据管理约束和要求。这些问题可以概括如下：

- *管理大数据*——从多个数据源收集数据并将其合并到一个统一的数据源中，以生成一个庞大的数据集，如前所述，增加(质量)数据的数量将提高机器学习过程的质量。第 2 章考虑了大数据平台的特征，并展示了在解决困难问题时，图所发挥的重要作用。
- *设计一个灵活模式*——尝试创建一个模式模型，该模型能够将多个异构模式合并到一个统一的同构数据结构中，以满足信息和导航需求。模式应该随机器学习项目目的变化而改变。第 4 章介绍了多个数据模型模式和最佳实践，以便为多个场景建模数据。
- *开发高效的访问模式*——快速的数据读取提高了训练过程在处理时间方面的性能。使用提供了多种灵活访问模式的数据平台，将有利于对训练数据进行特征提取、过滤、清洗、合并等预处理任务。

1.2.2　性能

性能是机器学习中的一个复杂主题，因为它可能与多种因素有关：

- 预测准确性：可通过使用不同的性能指标对其进行评估。回归问题的传统性能衡量标准是均方根误差(Root Mean Squared Error，RMSE)，它衡量系统在其预测中产生的误差的标准差[1]。换句话说，它关注测试数据集中所有样本的估计值和已知值之间的差异，并计算其平均值。在本书后面讨论不同算法时，我将介绍其他衡量性能的技术。准确性取决于多种因素，例如可用于训练模型的数据量、数据质量和所选算法。正如 1.2.1 节所讨论的，数据在保证预测具有适当的准确性水平方面发挥着主要作用。
- 训练性能：指计算模型所需的时间。要处理的数据量和使用的算法类型决定了计算预测模型所需的处理时间和存储空间。显然，这个问题对在训练阶段构建模型的算法的影响更大。对于基于实例的学习器[2]，性能问题会出现在后期，如预测阶段。在批量学习中，由于要处理的数据量较大，训练时间通常较长。相比之下，在线学习方法中，算法从较少量的数据中进行增量学习。虽然在线学习中要处理的数据量很小，但处理速度会影响系统与最新可用数据匹配的能力，从而直接影响预测的准确性。
- 预测性能：指提供预测所需的响应时间。机器学习项目的输出可能是帮助管理人员做出战略决策的静态一次性报告，或为最终用户提供的在线服务。在第一种情况下，完成预测和计算模型所需的时间并非重点，只要在合理的时间范围内(即，不是几年)完成即可。在第二种情况下，预测速度确实很重要，因为它会影响用户体验和预测的有效性。假设你正在开发一个推荐引擎，该引擎根据用户的兴趣来推荐与用户正在查看的产品相似的产品。此时用户导航速度相当快，这意味着需要在短时间间隔内进行大量预测；在用户浏览下一个条目之前，只有几毫秒用来提示有用内容。在这种情况下，预测速度是成功的关键。

这些因素也可以转换为机器学习项目的多项要求，例如在训练期间快速访问数据源、高数据质量和高效的访问模式以便模型加速预测等。在这种情况下，图可以为源数据和模型数据提供适当的存储机制，减少读取数据所需的访问时间，并提供多种算法技术来提高预测的准确性。

1.2.3　存储模型

在基于模型的学习器方法中，训练阶段的输出是一个将用于预测的模型。计算出该模型需要花费时间，并且必须将该模型存储在持久层中，以避免每次系统重启时对其重新计算。

---

1 标准差是表示组成员与组平均值之间的差的指标。

2 如果你不熟悉基于实例的算法和批量学习等概念，请参阅附录 A，其中涵盖了机器学习分类。

模型的结构与具体算法或所采用的算法类直接相关。例如：

- 使用最近邻法的推荐引擎的条目间相似度。
- 表示如何在簇中分组元素的项-簇映射。

两种模型的规模差异很大。假设有一个包含 100 个条目的系统。首先，条目间的相似度需要存储 100×100 个条目。利用优化过程，可以减少这个数字，只考虑前 k 个相似条目，在这种情况下，模型将需要存储 100×k 个条目。相比之下，项-簇映射只需要存储 100 个条目；因此，将模型存储在内存或磁盘中所需的空间可能很大也可能适中。此外，如前所述，模型访问/查询时间会影响预测阶段的全局性能。因此，模型存储管理成了机器学习中的一个重大挑战。

1.2.4 即时性

向用户提供实时服务时，会越来越多地使用到机器学习。从响应用户最后点击的简单推荐引擎，到受指示不伤害过马路行人的自动驾驶汽车，示例涵盖各个方面。在这两个示例中，尽管失败导致的结果大不相同，但其中学习算法对来自环境的新刺激做出快速(或适当及时)响应的能力很大程度上影响了最终结果的质量。

设想一个为匿名用户提供实时推荐的推荐引擎。这种匿名性(用户未注册或登录)意味着其并不存在之前的长期交互历史记录——只有 cookie 提供的基于会话的短期信息。这是一项复杂任务，其涉及多个方面并影响机器学习项目的多个阶段。所采用的方法因不同学习算法而异，但目标可描述如下：

- *快速学习*。在线学习器应能够在新数据可用时立即更新模型。该功能将减少事件或通用反馈之间的时间间隔，例如导航点击，或与搜索会话的交互以及模型的更新。模型与最新事件的匹配度越高，越能满足用户当前的需求。
- *快速预测*。模型更新后，预测会变得很快——最多几毫秒——因为用户可能会离开当前页面，甚至很快改变想法。

这两个目标都需要使用能快速匹配模型的算法，以及提供快速存储和高效访问模式的存储机制(在内存中、磁盘上或组合二者)。

1.3 图

正如本章导论中所述，图提供了可大力支持机器学习项目的模型和算法。即使图是一个简单的概念，了解如何表示它以及如何使用与之相关的主要概念也很重要。本节将介绍图的重要方面。如果你已经掌握了这些概念，可以跳过这一部分。

1.3.1 什么是图

图是一个简单且相当古老的数学概念：由一组顶点(或节点/点)和边(或关系/线)组成的数据结构，用于对一组对象之间的关系进行建模。据说，莱昂哈德·欧拉(Leonhard Euler)

于 1736 年首次提到了图。普鲁士的 Königsberg 坐落在普雷格尔河两岸，且包括两个大岛，它们彼此相连并通过七座桥与城市的两个大陆部分相连接。在游览 Königsberg 时，欧拉不想花费太多时间在城中散步。他将该问题定义为如何步行穿过这座城市的每座桥，且只穿过一次。他证明了这不可能实现，从而也促成了图和图论问世[Euler, 1736](因此他并没有去步行，只是待在家中)。图 1.3 显示了 Königsberg 的一张旧地图和欧拉用来证明其论点的图的一种表示。

正式来说，图是一个二元组 $G=(V, E)$，其中 V 是顶点的集合 $V=\{V_i, i=1, \cdots, n\}$，$E$ 是 V 上的边的集合，$E_{ij}=\{(V_i, V_j), V_i \in V, V_j \in V\}$。$E \subseteq [V]^2$；因此，$E$ 的元素是 V 的二元素子集[Diestel, 2017]。

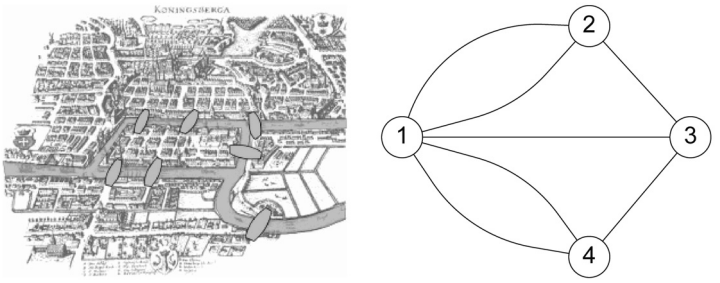

图 1.3　引发图论问世的 Königsberg 的桥

表示图最简单的方法是为每个顶点画一个点或一个小圆圈，若两个顶点形成一条边便用一条线进行连接，如图 1.4 所示。

图可以是有向的，也可以是无向的，这取决于是否在边上定义了遍历方向。在有向图中，可以将边 E_{ij} 从 V_i 遍历到 V_j，但相反方向不可行；V_i 被称为尾节点或起始节点，V_j 称为头节点或结束节点。在无向图中，边的遍历在两个方向上都有效。图 1.4 表示无向图，图 1.5 表示有向图。

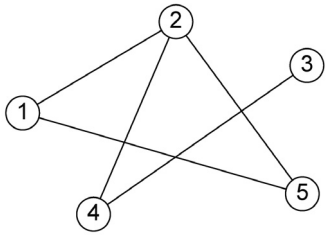

图 1.4　$V=\{1, 2, 3, 4, 5\}$ 上的无向图，边集 $E=\{(1,2),(1,5),(2,5),(2,4),(4,3)\}$

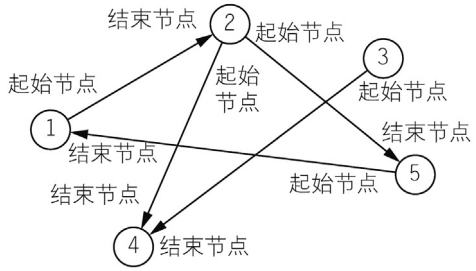

图 1.5 $V=\{1,...,5\}$上的有向图，边集 $E=\{(1,2),(2,5),(5,1),(2,4),(3,4)\}$

箭头指示关系方向。默认情况下，图中的边是未加权的；因此，相应的图也为未加权的。当一个权重(一个用来表达某种意义的数值)被分配到边上时，称该图为加权图。图 1.6 与图 1.4、图 1.5 相同，但给每条边都分配了权重。

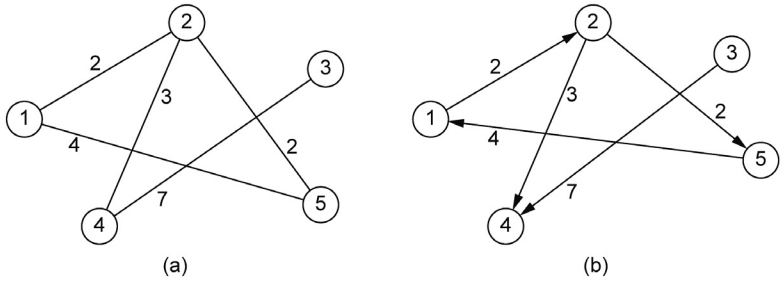

图 1.6 (a)一个无向加权图和(b)一个有向加权图

如果$\{x,y\}$是 G 的一条边，则将 G 的两个顶点 x 和 y 定义为相邻或互为邻点。连接它们的边 E_{ij} 被视为与两个顶点 V_i 和 V_j 相关。如果两个不同的边 e 和 f 具有一个共同顶点，则它们是相邻的。如果 G 的所有顶点都是成对相邻的，则 G 是完备的。图 1.7 显示了一个完备图，其中每个顶点都与所有其他顶点相连接。

图中顶点最重要的属性之一是它的度，度被定义为与该顶点相关的边的总数，也等于该顶点的邻点数。例如，在图 1.4 所示的无向图中，顶点 2 的度数为 3(顶点 1、4 和 5 为其邻点)；顶点 1(邻点是 2、5)、4(邻点为 2、3)和 5(邻点为 1、2)的度数为 2，而顶点 3 的度数为 1(仅与 4 连接)。

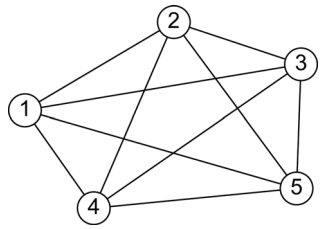

图 1.7 一个完备图，其中每个顶点都与所有其他顶点相连接

在有向图中，顶点 V_i 的度分为顶点的入度(以 V_i 作为其结束节点的边的数量，结束节点即箭头的头部)和顶点的出度(以 V_i 作为其起始节点的边的数量，起始节点即箭头的尾部)。在图 1.5 所示的有向图中，顶点 1 和 5 的入度和出度为 1(它们各有两个关系，一个传入和一个传出)，顶点 2 的入度为 1、出度为 2(一个来自 1 的传入关系，两个分别传向 4 和 5 的传出关系)，顶点 4 的入度为 2，出度为 0(两个分别来自 2 和 3 的传入关系)，顶点 3 的出度为 1，入度为 0(一个传向 4 的传出关系)。

图的平均度定义如下：

$$a = \frac{1}{N} \sum_{i=1,\cdots,N} degree(V_i)$$

其中 N 是图中的顶点数。

若某顶点序列中每一对连续的顶点都由一条边连接起来，则将这个序列称为路径。没有重复顶点的路径称为简单路径。若路径中第一个顶点和最后一个顶点重合，则称这个路径为环。图 1.4 中，[1, 2, 4]、[1, 2, 4, 3]、[1, 5, 2, 4, 3]等为路径；特别地，顶点[1, 2, 5]的路径代表一个环。

1.3.2　图作为网络模型

图可表示事物在简单或复杂结构中的任何物理或逻辑链接。我们在图中为边和顶点命名并赋予含义，这就构成了网络。在这些情况下，图是描述网络的数学模型，而网络是对象之间的一组关系，对象可以包括人、组织、国家、谷歌中的搜索条目、脑细胞或变压器。这种多样性说明了图的强大功能及其简单结构(这也意味着它们需要的磁盘存储容量较小)，可以使用这些结构对复杂系统进行建模[1]。

我们通过一个例子来说明这个概念。假设有图 1.8 所示的图。

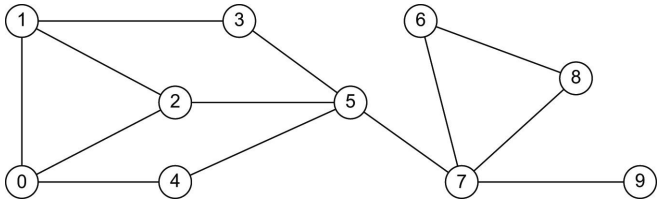

图 1.8　一个重要的通用图

在数学定义上，该图是纯粹的，可根据边和顶点的类型对多种类型的网络进行建模：

- 若顶点是人，且每条边代表人与人(朋友、家人、同事)之间的任何关系，则可以建模一个社交网络。
- 若顶点是信息结构，如网页、文档或论文，边代表逻辑连接，如超链接、引用或交叉引用，则可以建模一个信息网络。

1 在这种情况下，用动词建模以简化方式表示系统或现象。建模的目的在于用一种便于计算机系统处理的方式来表示数据。

- 若顶点是计算机或其他可以转发消息的设备，边代表可以传输消息的直接链接，则可以建模一个通信网络。
- 若顶点是城市，边代表使用航班、火车或公路的直接连接，则可以建模一个交通网络。

这组示例演示了同一个图如何通过为边和顶点分配不同的语义从而表示多个网络。图 1.9 说明了不同类型的网络。

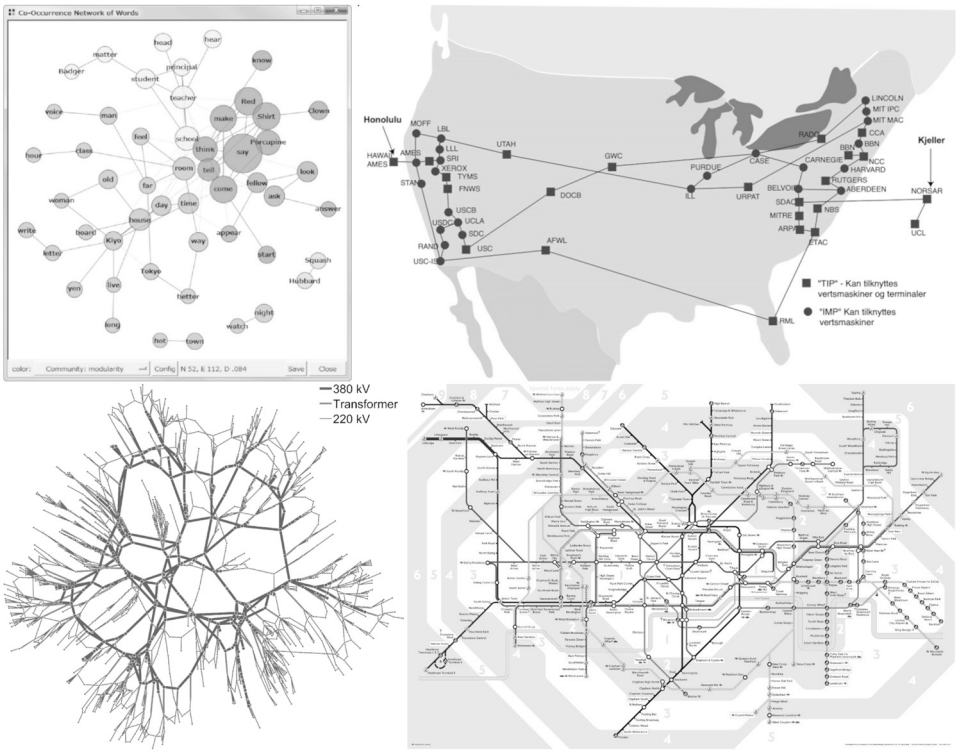

图 1.9 从左上角开始顺时针方向依次是共现网络[1]、ARPA 网络 1974[2]、伦敦地铁网络[3]和电网[4]

根据图 1.9，我们可以发现图的另一个有趣特征：它们具有高度的流动性。图能够清晰地表示信息，这就是人们常将其用作信息图的原因。将数据表示为网络并使用图算法，可以：

- 查找复杂模式。
- 使它们为人类所用，以供进一步调查和解释。

1 Higuchi Koichi——KH 编码器的共现屏幕截图(https://en.wikipedia.org/wiki/Co-occurrence_network)。

2 Yngvar——截至 1974 年 9 月的 ARPANET 的符号表示(https://en.wikipedia.org/wiki/ARPANET)。

3 由伦敦交通局提供 (http://mng.bz/G6wN)。

4 Paul Cuffe——高压传输系统的网络图，显示了不同电压等级之间的交互(https://en.wikipedia.org/wiki/Electrical_grid)。

当机器学习与人脑的力量相结合时，可以进行高效、先进和复杂的数据处理和模式识别。通过突出显示元素间的连接，网络可用于显示数据。报纸和新闻网站使用网络的频率越来越高，这不仅可以帮助人们导航数据，还可以为人们提供一种强大的调查工具。最近(在撰写本书时)，巴拿马报纸[1]展示了网络的惊人特征。国际调查记者联盟(International Consortium of Investigative Journalists，ICIJ)分析了泄露的财务文件，此举曝光了世界上最富有的精英人群使用的高度连接离岸税收结构网络。记者从文件中提取实体(人、组织和任何类型的中介)和关系(保护人、受益人、股东、董事等)，将它们存储在网络中，并使用可视化工具对其进行分析。结果如图 1.10 所示。这里，网络、图算法和图可视化使无法靠传统数据挖掘工具发现的信息变得显而易见。

Valdis Krebs[2]是一位专门研究社交网络应用程序的组织顾问，他博客中的文章也提供了许多相关的有趣示例。示例包含通过图可视化将图机器学习与人类思维相结合。这里，我们引用其中最著名的例子之一。

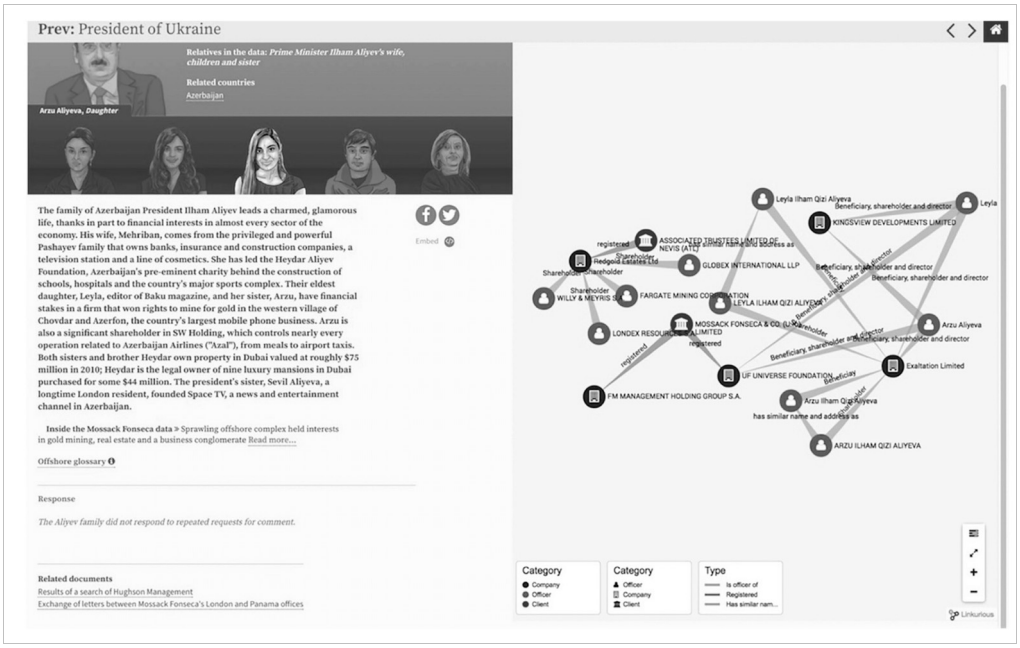

图 1.10　巴拿马报纸的图可视化示例

图 1.11 中的数据来自 Amazon.com，代表了 2008 年在美国购买最多的政治书籍[Krebs, 2012]。Krebs 对数据进行了网络分析，以创建与当年总统选举相关的书籍地图。如果某两本书经常被同一客户购买，则这两本书是相链接的。这些书被称为“同购对”(因为购买了这本书的顾客同时购买了那本书)。

1 https://panamapapers.icij.org。

2 http://www.thenetworkthinkers.com。

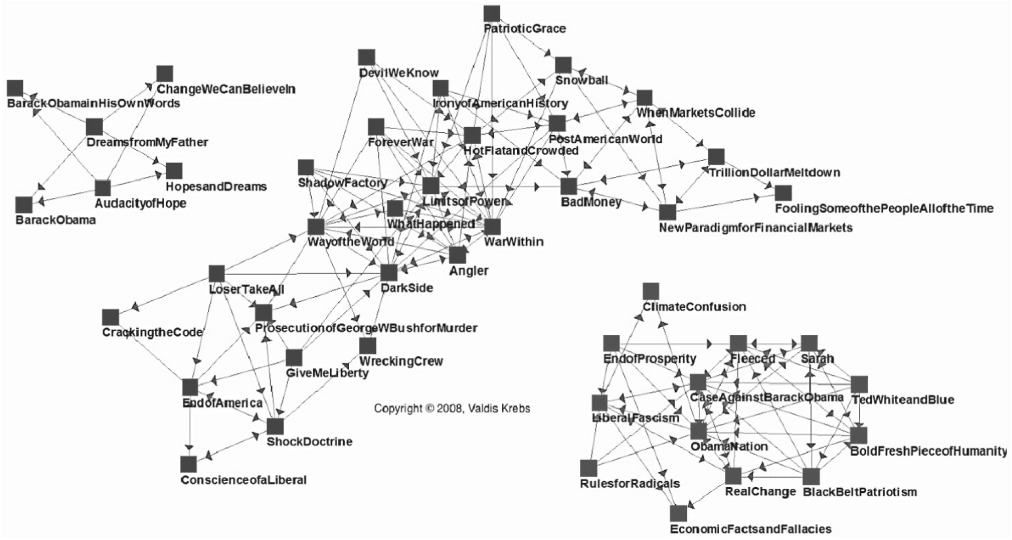

图 1.11 2008 年美国政治书籍网络图(Krebs，2012)

有三个不同且不重叠的簇：

- 左上角的奥巴马书籍簇。
- 中间的民主党(蓝色)簇。
- 右下角的共和党(红色)簇。

2008 年，美国政治环境两极分化严重。这一事实反映在亚马逊的政治书籍数据中，图 1.11 显示了保守派和自由派选民之间的巨大分歧。红皮书和蓝皮书之间没有联系或中介；每个簇都与其他簇完全不同。如前所述，某一群人在阅读总统候选人巴拉克 • 奥巴马的传记，但他们显然没有兴趣阅读或购买其他的政治书籍。

四年后，在 2012 年，同样的分析产生了一个看起来截然不同的网络(图 1.12)。该网络展示了许多作为簇间桥梁的书籍。此外，潜在选民似乎在阅读有关两位主要候选人的书籍。因此得出了一个更复杂的网络，其没有孤立的簇。

政治书籍网络的例子介绍了网络的一个重要方面。如果图是一个只存在于其柏拉图式世界中的纯数学概念[1]，那么网络作为某些具体系统或生态系统的抽象概念，会受到作用于它们的力的影响，从而导致其结构发生变化。我们将这些力量称为周围环境：这种因素存在于网络顶点和边之外，但影响着网络结构如何随时间演变。这种环境的性质和力的类型特定于网络的类型。例如，在社交网络中，每个人都有一组独特的个人特征，两个人的特征之间的相似度和兼容性会影响链接的创建或删除[Easley 和 Kleinberg, 2010]。

1 数学柏拉图主义(http://mng.bz/zG2Z)是一种形而上学的观点，其认为存在独立于我们和我们的语言、思想和实践的抽象数学对象。

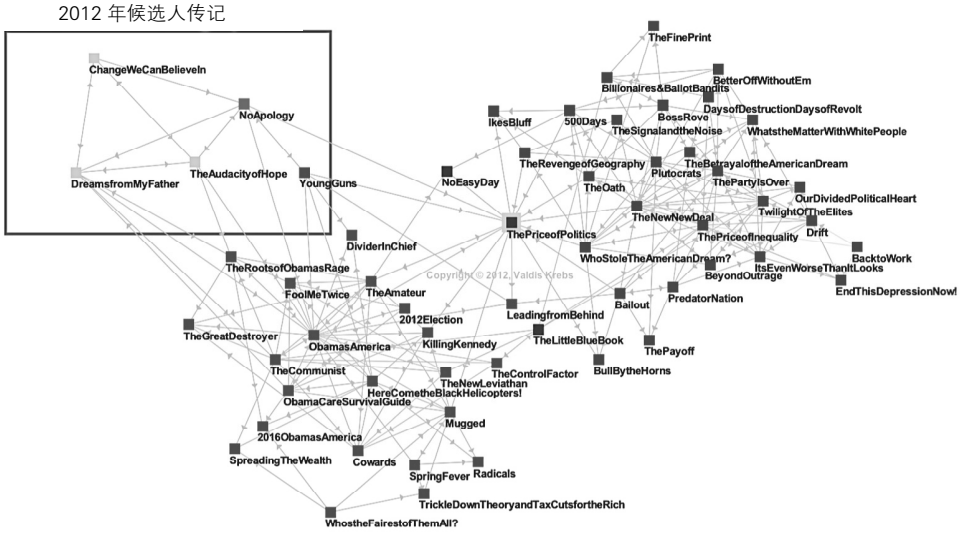

图 1.12　2012 年美国政治书籍网络图[Krebs, 2012]

决定社交网络结构的最基本概念之一是同质性(该词来自希腊语，意思是对同一事物的爱)：社交网络中的链接倾向于将相似的人联系起来。更正式地说，如果两个人的特征在他们所来自的人群或他们所在的网络中的匹配比例大于预期，则他们就更有可能相互联系[Verbrugge, 1977]。反之亦然：如果两个人相互联系，那么他们更有可能具有共同特征或属性。因此，(例如)我们在 Meta 上的朋友并不像是偶然遇到，相反，在民族、种族和地理维度，他们与我们大体相似；他们的年龄、职业、兴趣、信仰和观点往往与我们的相似。这种观察由来已久，早在 Mark Zuckerberg 写下他的第一行代码之前就开始了。在柏拉图(“相似度产生友谊”)和亚里士多德(人们“爱与自己相似的人”)的著作中以及民间谚语(例如“物以类聚、人以群分”)中都可以找到该种观察的基本思想。同质性原则也适用于群体、组织、国家或社会单元的任何方面。

了解周围环境和作用于网络上的相关力量有助于以下列多种方式完成机器学习任务：

- 网络是一把双刃剑。营销人员总是试图了解并说服人们。若能像滚雪球一样解决问题，那么人与人间接触是最有效的。这个概念就是所谓病毒式营销的基础。
- 了解这些力量后，则可以预测网络将如何随时间演变，并使数据科学家能够主动应对此类变化或将其用于特定业务目的。
- 社会学和心理学的研究发现表明，一个人的社交网络在一定程度上决定了其品位、偏好和活动。这一信息有助于构建推荐引擎。与推荐引擎相关的问题之一是冷启动问题：因为你没有新用户的历史记录，所以无法为他们预测任何事情。可以利用社交网络和同质性原则，根据相连接用户的品位进行推荐。

1.4 图在机器学习中的作用

图可用于描述感兴趣目标之间的交互，用于对简单和复杂的网络进行建模，或者更常用于表示现实世界的问题。因为图具有严格而直接的形式，所以图被用于从计算机科学到历史科学的许多科学领域。作为一种强大的工具，图被广泛用于机器学习中，可以实现人们的想法，并提供许多有用的特征，我们不必对此感到惊讶。随着时间的推移，图机器学习变得越来越普遍，超越了许多传统技术。

不论规模大小，许多公司都在使用这种方法，为其客户提供更高级的机器学习特征。一个著名例子是谷歌，它使用图机器学习作为其 Expander 平台的核心。这项技术为用户几乎每天都在使用的许多 Google 产品及特征提供支持，例如 Gmail 收件箱中的提醒或 Google 相册中最新的图像识别系统[1]。

构建一个图机器学习平台有很多好处，因为图不仅可以克服之前所述的挑战，还可以提供更高级的特征，而如果没有图的支持，这些特征就无法实现。

图 1.13 说明了机器学习和图之间的主要联系点，其中考虑到了不同任务的目标。

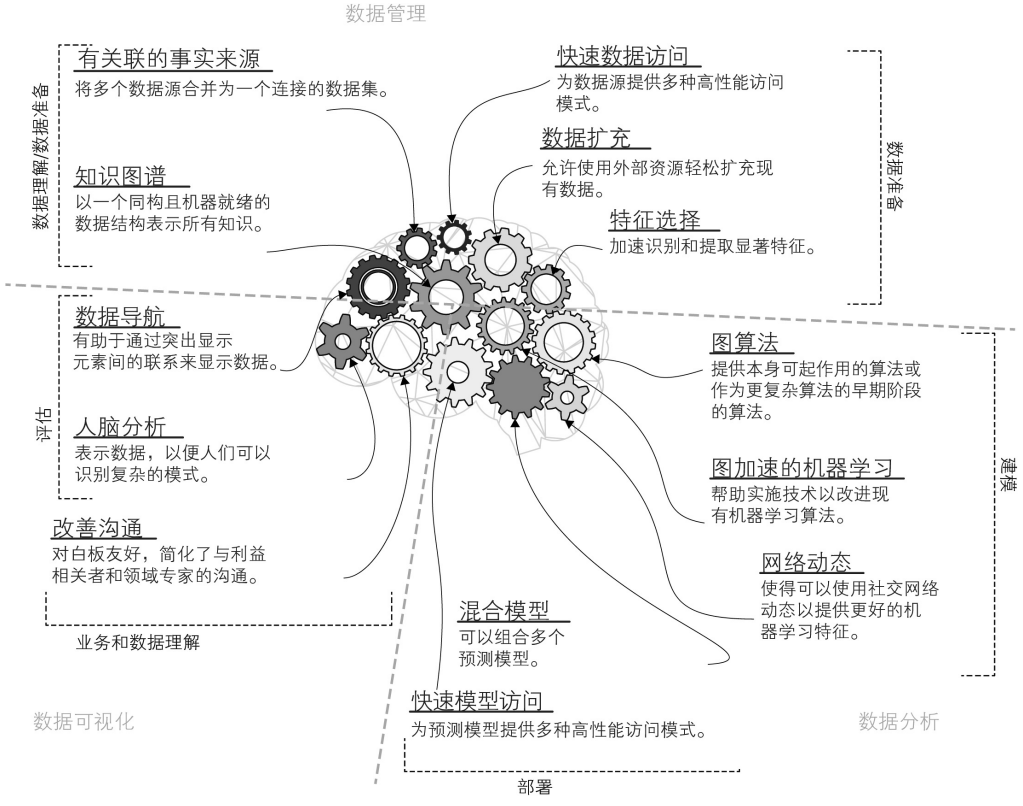

图 1.13 图机器学习思维导图

1 http://mng.bz/0rzz。

可以使用该思维导图快速地在概念上将图在机器学习中的作用可视化。在图 1.13 中，图特征分为三个主要区域：

- *数据管理*——该区域包含图提供的特征，其帮助机器学习项目处理数据。
- *数据分析*——该区域包含对学习和预测有用的图特征和算法。
- *数据可视化*——该区域强调了图作为一种可视化工具的实用性，可帮助人们通过人脑进行交流、与数据交互并产生观点。

该模式还显示了基于图的技术与 CRISP-DM 模型中的阶段之间的映射。

1.4.1　数据管理

图使学习系统可以探索更多的数据，更快地获取数据，并轻松地清洗数据和扩充数据。传统学习系统在研究人员准备的单个表上进行训练，而图原生系统可以访问的不仅仅是这个表。

图驱动的数据管理特征包括：

- *有关联的事实来源*——图使你能将多个数据源合并为一个统一的、连接的数据集，从而为训练阶段做好准备。此特征可减少数据稀疏性、增加可用数据量并简化数据管理，显示出了巨大优势。
- *知识图谱*——在前一点的基础上，知识图谱提供了一种同构的数据结构，不仅可以组合数据源，还可以组合预测模型、手动提供的数据和外部知识源。生成的数据是机器就绪的，可以用于训练、预测或可视化过程。
- *快速数据访问*——表提供与行和列过滤器相关的单一访问模式。另一方面，图为同一组数据提供了多个访问点。此特征将要访问的数据量减少到特定需求集的基线最小值，从而提高了性能。
- *数据扩充*——除了可以轻松地使用外部资源扩展现有数据，图的无模式特性和图数据库中提供的访问模式还有助于数据清洗和合并。
- *特征选择*——识别数据集中的相关特征是多项机器学习任务(如分类)的关键。通过快速访问数据和多种查询模式，图可以加速特征识别和提取。

在 CRISP-DM 模型的数据理解和数据预处理阶段，有关联的事实来源和知识图谱十分重要，而快速数据访问、数据扩充和特征选择在数据预处理阶段十分有用。

1.4.2　数据分析

图可用于建模，并分析实体之间的关系及其属性。这可以带来另一个信息维度，即图机器学习可以将其用于预测和分类。图提供的模式灵活性还允许不同的模型共存于同一数据集中。

图驱动的数据分析特征包括：

- *图算法*——有几种图算法有助于识别数据中的见解并进行分析，如聚类、PageRank 和链接分析算法。此外，它们可以用作更复杂分析过程中的第一个数据预处理步骤。

- *图加速的机器学习*——前面讨论的图驱动特征提取是说明图如何加速或提高学习系统质量的一个例子。图可以帮助在训练阶段前或训练期间过滤、清洗、扩充和合并数据。
- *网络动态*——了解周围环境和作用于网络的相关力量，将使你不仅可以了解网络动态，还可以将其用于提高预测质量。
- *混合模型*——利用灵活快速的访问模式，多个模型可以共存于同一个图中，前提是它们可以在预测阶段被合并。此特征提高了预测的最终准确性。此外，有时可以用不同方式使用相同模型。
- *快速模型访问*——实时使用需要快速预测，这意味着要尽可能快地访问模型。图为这些任务提供了正确的模式。

图算法、图加速的机器学习和网络动态主要参与建模阶段，因为与其他特征相比，它们与学习过程的联系更多。部署阶段使用混合模型和快速模型访问方法，因为它们在预测阶段运行。

1.4.3 数据可视化

图具有很强的交流能力，可以用易于人脑理解的方式同时显示多种类型的信息。此特征在机器学习项目中非常重要，可用于共享结果、分析结果或帮助人们导航数据。

图驱动的数据可视化特征包括：

- *数据导航*——网络可通过突出显示元素之间的连接来显示数据。它们既可以帮助人们正确导航数据，也可以用作强大的调查工具。
- *人脑分析*——通过将机器学习与人脑相结合，以图的形式表示数据，这将释放机器学习的效能，实现高效、先进、复杂的数据处理和模式识别。
- *改善交流*——图(尤其是属性图)是“白板友好型的”，这意味着当存储在数据库中时，在概念上它们表示在板上。此特征缩小了复杂模型的技术性与复杂模型被传达给领域专家或利益相关者的方式之间的差距。有效沟通可以提高最终结果的质量，因其减少了对领域、业务目标以及项目需求和约束的理解方面的问题。

在业务和数据理解阶段，改善沟通尤为重要，而数据导航和人脑分析主要与评估阶段相关。

1.5 本书心智模型

本章中提供的思维导图可帮助你轻松地将图在机器学习项目中的作用进行可视化。这并不意味着，在所有项目中，你都要对其中列出的所有内容使用图。在本书的几个例子中，图被用来解决问题或(在质量和数量方面)提高性能；通过查看一个简单的图，使用心智模型可以帮助你了解该图在特定情况下的作用。

下一个模式将图机器学习思维导图中的关键特征(接触点)分类归入机器学习工作流的四个主要任务中：

- *管理数据源*，指为学习阶段收集、合并、清洗和准备训练数据集的所有任务。
- *算法*，涉及将机器学习算法应用于训练数据集。
- *存储并访问模型*，包括存储预测模型的方法和用于提供预测的访问模式。
- *可视化*，指的是可以将数据可视化，从而辅助分析的方式。

图 1.14 所示的思维导图中总结了这些要点。

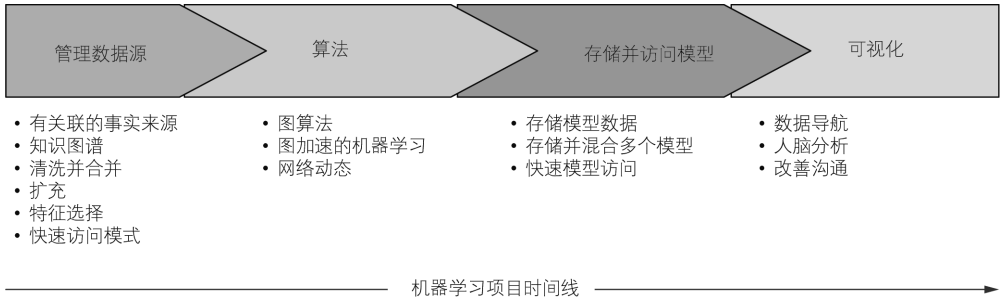

图 1.14　描述一个机器学习项目四个阶段的心智模型

该图将在本书中经常出现。该模式将帮助你快速定位当前讨论在项目工作流中的位置。

这种心智模型从整个流程的角度展示了机器学习项目，并且是确定你在机器学习生命周期中所处位置的最佳方式。但是，从更广泛的、面向任务的角度考虑项目也很有用。

1.6　本章小结

- 机器学习旨在开发能够自主从样本数据中获取经验的计算机程序，在不明确编程的情况下将其转化为专业知识。
- 机器学习项目不仅是一个软件项目，还是一个人工过程，涉及一群具有不同技能的人和大量工作。需要使用定义明确且系统的方法才能成功。CRISP-DM 提供正式的项目生命周期来推动这样的项目，帮助得出正确的结果。
- 任何机器学习项目都必须面对的难题主要与数据管理——无论是在训练数据集还是预测模型方面——以及学习算法的性能有关。
- 图是简单的数学概念，可用于对复杂网络进行建模和分析。网络外部的周围环境对其起作用，决定了其演变方式。
- 图和网络可以通过多种方式为机器学习项目提供支持，体现在三个维度：数据管理、数据分析和数据可视化。

(注：本章的参考文献，请扫描本书封底的二维码进行下载。)

第 2 章

图数据工程

本章内容

- 大数据作为机器学习输入的主要挑战
- 如何使用图模型和图数据库进行大数据分析
- 图数据库的形式和特征

第 1 章强调了数据在机器学习项目中的关键作用。正如我们所见，相比于微调或替换算法本身，在大量高质量数据上训练学习算法更能提高模型的准确性。Greg Linden 为亚马逊发明了如今广泛使用的条目间协同过滤算法，他在一篇关于大数据的综述中[Coyle, 2016]提到：

亚马逊的推荐如此有效，关键在于大数据。大数据可以调整搜索并帮助我们找到需要的东西。大数据使网络和手机变得智能化。

在过去几年中，信息技术、工业、医疗保健、物联网(Internet of Things，IoT)等多个领域产生的数据量呈指数级增长。数据来源数不胜数：用于收集气候信息的传感器、社交媒体网站上的帖子、数字图片和视频、购买交易记录和手机 GPS 信号，这些仅是其中几例。这些数据即大数据。图 2.1 显示了每分钟从知名应用程序或平台生成的当前数据量的一些统计数据[Domo, 2020]。

是什么导致过去十年间产生这种巨大变化？原因既不是互联网用户数量的爆炸式增长，也不是数据传感器等新系统的创建，而在于人们更加意识到数据作为知识来源的重要性。人们想要更多地了解各种用户、客户、企业和组织，这种强烈愿望对数据采集、收集和处理提出了新的需求和要求，且使数据科学家采用不同的方式收集数据并进行分析。多年前，他们还不得不在杂乱无章的数据中寻找格式奇怪的数据。现在，随着公司逐渐认识到其生成数据中隐藏的价值，数据科学家已经成为引领数据生成和收集工作的核心力量。

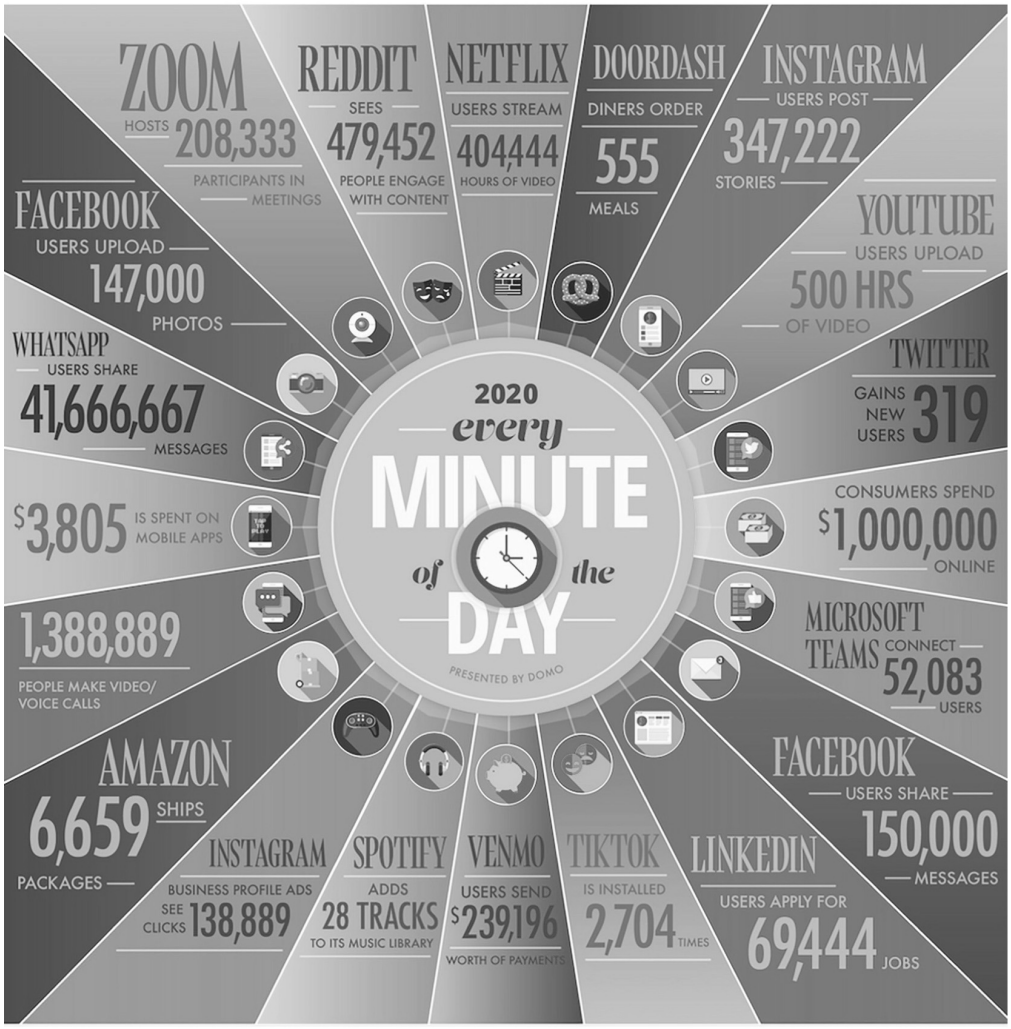

图 2.1　2020 年每分钟生成的数据(Domo, Inc. 提供)

例如，一个旅游网站曾经可能只收集星级评价，将结果应用于一个简单的推荐引擎，但现在该网站利用每个用户评论中的可用信息，将其作为更详细的知识来源。这种心智过程会产生一种良性循环(图 2.2)，其能够在每个循环中收集更多数据。

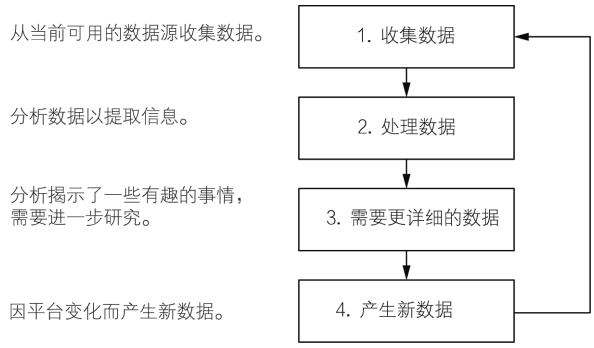

图 2.2 数据收集循环

这种数据源的可用性前所未有，使机器学习从业者能够访问多种形式的海量数据。查找和访问这些数据还相对容易，然而存储和管理数据则完全不同。具体来说，机器学习过程中，识别和提取所观察现象的相关特征(或是可测量特性或特征)很有必要。选择信息丰富、有辨别力且独立的特征是创建有效学习算法的关键步骤，因为这些特征定义了算法训练阶段的输入结构并决定了预测的准确性。若算法的类别不同，对特征列表的要求也会相应地变化，但一般而言，更准确的数据生成的模型更好。将 300 个因素纳入考虑得出的预测结果优于仅考虑 6 个因素得出的结果。然而，如果因素过多，就会面临过拟合的风险。

用于分析大数据系统的生命周期由一系列步骤组成，第一步就是从多个数据源收集数据。需要用专门的工具和框架将不同来源的数据提取到已定义的大数据存储中。数据存储在特定的可扩展解决方案中(如分布式文件系统或非关系数据库)。更正式地说，完成这些任务所需步骤如下：

- *收集*。汇集并收集来自多个数据源的数据。
- *存储*。以适当的方式将数据存储在单个(有时为多个)易于访问的数据存储中，为下一阶段做好准备。
- *清洗*。使用统一、同构的模式合并、清洗和(只要有可能)归一化数据。
- *访问*。数据处于可用状态。提供多个视角或访问模式以简化并加快访问将用于训练的数据集。

本章重点介绍这四个步骤中的后三步：存储、清洗和访问数据。本章描述了大数据的主要特征并讨论了处理数据的方法，且详细说明了基于图模型和图数据库的特定方法，并提供了最佳实践。

2.1 处理大数据

为了解决难题——定义大数据分析平台的要求——我们需要了解它。下面介绍使大数据“大”的基本特征。

2001 年(没错，20 多年前！)，META 集团的分析师 Doug Laney 发表了一篇题为“3D Data Management：Controlling Data Volume，Velocity，and Variety”的研究报告[Laney, 2001]。

尽管在 Laney 的报告中并没有出现“三 V”一词，但十年后，“三 V”——数量(volume)、速度(velocity)和多样性(variety)——已成为大众普遍接受的定义大数据的三个维度。

后来又增加了另一个维度：真实性(veracity)，指数据集或数据源的质量、准确性或真实性。随着新的、不受信任的且未经验证的数据源的出现(例如 Web 2.0 中用户生成的内容)，所收集信息的可靠性和质量成为一个大问题，这使得人们普遍接受将真实性作为大数据平台的重要维度。图 2.3 总结了“四 V”的主要方面，之后将对其进行更详细的描述。

多年来，随着“大数据”一词受到广泛关注并逐渐流行，分析师和科技记者在维度列表中添加了越来越多的 V。截止本书撰写时，已经提出了 42 个 V[Shafer, 2017]；毫无疑问，随着时间的推移，还会加入更多 V。

我们重点关注最初的“三 V”以及真实性，因为它们仍然是最为广泛接受的。这些维度贯穿全书，用于强调图模型和数据库对于管理大量数据的作用。

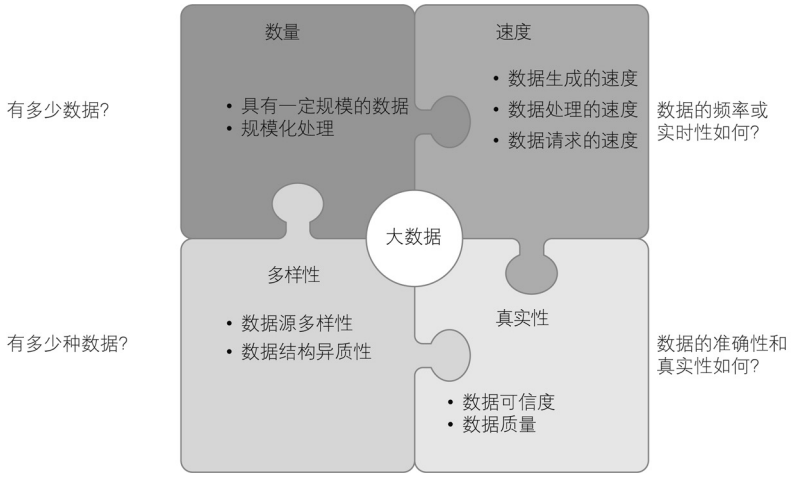

图 2.3　大数据的“四 V”

2.1.1　数量

数量带来的好处(图 2.4)——处理大量信息的能力——是大数据用于机器学习的主要优势。拥有更多且更好的数据胜过拥有更好的模型。若有大量数据，即使是简单的数学运算也可能十分高效。

假设你想要使用机器学习方法，利用多种电子健康记录(Electronic Health Record，EHR)数据，从而对不同症状所需的治疗类型进行实时预测。患者健康数据的数量呈指数级增长，这意味着 EHR 数据的数量也在猛增。根据 Health Data Archiver[2018]：

EMC 和研究公司 IDC 的一份报告通过新颖的方式将健康数据的剧增可视化，预计健康数据的总体年增长率为 48%。该报告显示 2013 年的医疗保健数据量为 153 EB。按照预计增长率，到 2020 年，数量将增长到 2,314 EB。

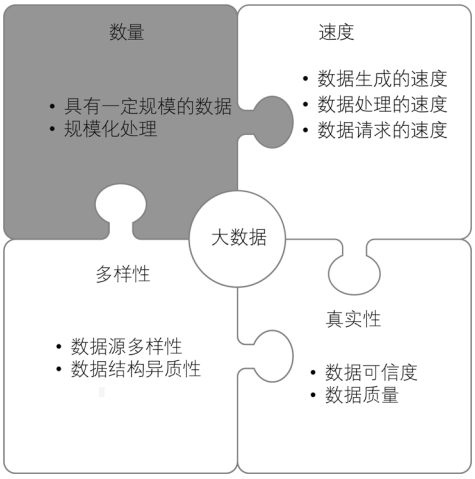

图 2.4 大数据的数量

如前所述，现代 IT、工业、医疗保健、IoT 等系统生成的数据量都呈指数级增长。这种增长一方面是由于数据存储和处理架构的成本降低，另一方面是因为能够从数据中提取有价值的信息(这创造了新需求)——可以改善业务流程、效率和为最终用户或客户提供的服务。尽管对于数据量来说不存在固定的阈值用以判断数据是否为“大”，但该术语通常表示一定规模的数据，“使用传统关系数据库系统和数据处理架构难以对数据进行存储、管理和处理”[Bahga and Madisetti, 2016]。

持续收集和分析这些大数据已成为 IT 所有领域的主要挑战之一。应对这一挑战的解决方案分为两大类：

- *可扩展存储*——扩展存储通常是指添加更多机器并在这些机器上分配负载(读取、写入或同时进行这两项工作)。此过程称为水平扩展。还可以通过查询或访问机制来实现可扩展性，这些机制为完整数据存储的一个子集提供多个访问点，而不需要使用过滤器或索引查找来遍历整个数据集。原生图数据库属于第二类情况，相关内容将在 2.3.4 节中讨论。
- *可扩展处理*——处理过程的水平扩展不仅意味着多台机器并行执行任务；它还需要使用一种分布式查询方法、一个通过网络进行有效通信的协议、编排、监控以及分布式处理的特定范式(如分治、迭代和流程)。

在 CRISP-DM 生命周期的数据理解和数据预处理阶段(见图 2.5)，需要确定数据源和每个阶段的大小和结构，用于设计模型并确定待使用的数据库管理系统(Database Management System，DBMS)。

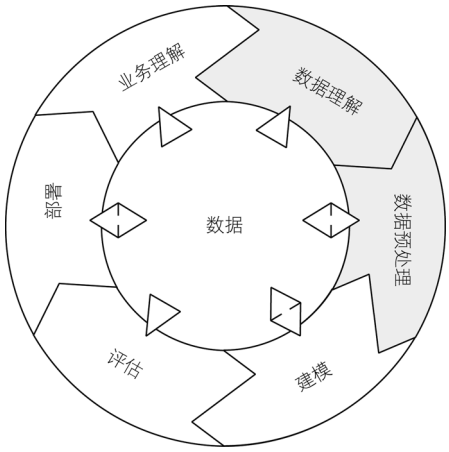

图 2.5　CRISP-DM 模型中的数据理解和数据预处理

在这些阶段，图可以提供有价值的支持，帮助解决数据数量方面存在的问题。基于图的模型使来自多个数据源的数据可以存储在高度连接的单一同质事实来源中，该来源可提供多种快速访问模式。具体来说，在大数据平台中，图通过以下两种方法帮助解决数量问题：

- *主数据源*——在这种情况下，图包含具有最低粒度的所有数据。学习算法直接访问图以执行分析。从这个意义上说，根据分析的类型，一个合适的大数据图数据库必须展示：
  - 一个索引结构(在其他 SQL 和 NoSQL 数据库中很常见)，以支持随机访问。
  - 一种仅访问小部分图的访问模式，不需要复杂的索引查找或数据库扫描。
- *物化视图*——在这种情况下，图代表主数据集的一个子集或其中数据的一个聚合版本，这对于分析、可视化或结果评估很有用。视图可以是分析过程的输入或输出，在这种情况下，图提供的全局和局部访问模式也很有帮助。

2.2 节通过展示两个示例场景及其相关实现来说明这些方法。

2.1.2　速度

速度(图 2.6)指生成、累积或处理数据有多快。例如，在一小时内接收并处理 1,000 个搜索请求不同于在不到一秒内接收并处理相同数量的请求。一些应用程序对数据分析有严格的时间约束，包括股票交易、在线欺诈检测和实时应用程序。数据速度的重要性遵循与数量相似的模式。以前局限于行业特定部门的问题现在出现在更广泛的场景中。

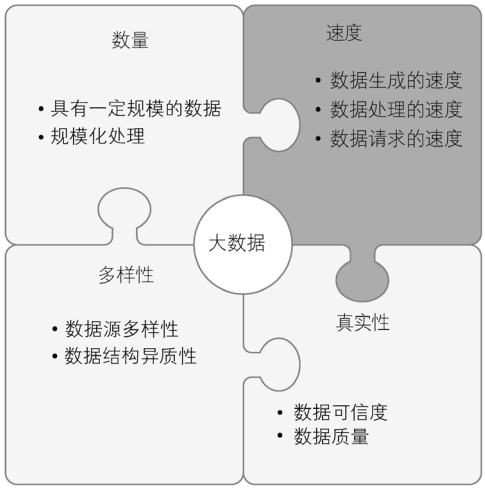

图 2.6 大数据的速度

假设你在研究自动驾驶汽车。每辆自动驾驶汽车都可以访问许多传感器，如摄像头、雷达、声呐、GPS 和激光雷达[1]。每个传感器每秒都会产生大量数据，如表 2.1[Nelson, 2016]所示。

表 2.1 某自动驾驶汽车传感器每秒产生的数据

传感器	每秒产生的数据量
摄像头	约 20-40 MB/s
雷达	约 10-100 KB/s
声呐	约 10-100 KB/s
GPS	约 50 KB/s
激光雷达	约 10-70 MB/s

在这种情况下，你设计的系统不仅应该能快速处理这些数据，还应该能尽快地生成预测，以避免撞到过马路的行人(一个示例)。

但是传入数据的速度并不是唯一需要考虑的问题，例如，可以将快速移动的数据传输到大容量存储器，以供后续进行批量处理。速度的重点在于反馈循环的整体速度(见图 2.7)，这涉及获取从输入到决策过程中所产生的数据：

通过反馈循环，系统监控预测的有效性并在需要时进行再训练，从而持续学习。机器学习的核心是监控并使用由此产生的反馈[2]。

1 “激光雷达和雷达的工作原理非常相似，但它不发射无线电波，而是发射红外光脉冲——也就是肉眼不可见的激光——并测量它们在击中附近目标后返回所需的时间。”来源：http://mng.bz/jBZ9。

2 Puget and Thomas [2016]。

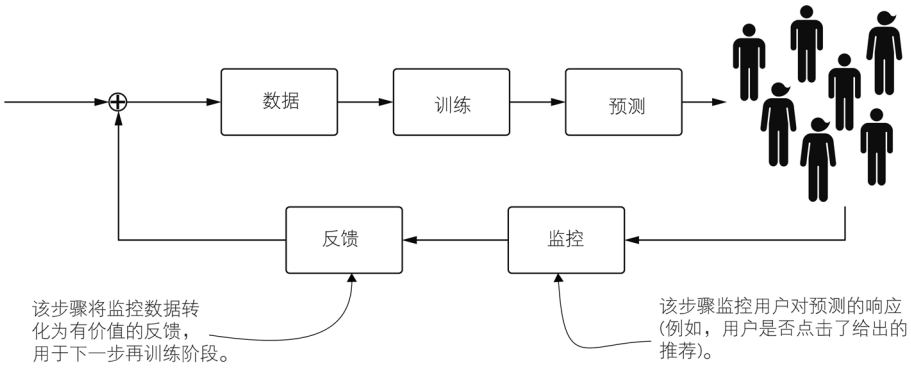

图 2.7　反馈循环示例

IBM 的一则广告表明，如果需要获取一张 5 分钟前的交通快照，那你根本就不会过马路。这个例子说明，有时人们无法等待报告运行或 Hadoop 作业完成。换句话说，“反馈循环越紧密，竞争优势就越大”[Wilder-James, 2012]。理想情况下，在数据生成时，实时机器学习平台应该能够立即对其进行分析。

随着时间的推移，出现了用于管理大数据的机器学习架构最佳实践。Lambda 架构[Marz 和 Warren, 2015]作为一种用于构建实时数据密集型系统的架构模式，就是其中一种实践；我们将在 2.2.1 节中呈现基于图的实现。

应对速度的数据基础设施必须能够快速访问必要数据，这些数据可能是整体数据的一部分。假设你要为出租度假屋实现一个实时推荐引擎。学习算法根据用户最近的点击和搜索记录来推荐房子。在这种情况下，引擎不必访问整个数据集；它检查最后 N 次点击或最后 X 时间范围内的点击记录并进行预测。原生图数据库(2.3.4 节中对其进行描述)维护每个节点的关系列表。从最后一次点击开始，并使用一个合适的图模型，引擎可以根据每个返回节点的关系返回到之前的点击。该示例展示了图在提供高性能访问数据集，以应对速度要求方面的显著优势。

2.1.3　多样性

多样性(图 2.8)与所分析数据的不同类型和性质有关。数据的格式、结构和大小各不相同；其形式通常杂乱无章，很少处于准备好进行处理的状态。由于大数据来源不同，因此其具有多种形式(结构化、非结构化或半结构化)和格式；它可以包括文本数据、图像、视频、传感器数据等。任何大数据平台都需要足够灵活，以应对这种多样性，当考虑到数据潜在不可预测的演变时更是如此。

假设你想对分布在多个数据源中的有关某公司的所有知识进行整理，从而为信息引擎创建知识库[1]。数据可以采用不同格式，涵盖从结构良好的关系数据库到对公司提供的产品或服务的非结构化用户评论，从 PDF 文档到社交网络数据。为了方便机器学习平台处理，

1 “信息引擎应用相关性方法来描述、发现、组织和分析数据。这使得数字工作者、客户或成员在办理业务期间，可以及时主动或交互式地提供现有或合成的信息。”(来源：http://mng.bz/WrwX)。

需要以同构方式将数据作为一个单元进行组织、存储和管理。

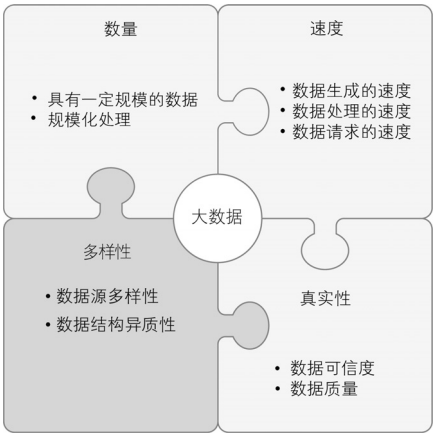

图 2.8 大数据的多样性

尽管关系数据库很流行且广为人知，但它们不再是大数据平台的首选。不同类型的数据适用于不同类型的数据库。例如，社交网络关系本质上是图，因此非常适合存储在诸如 Neo4j(附录 B 中介绍)之类的图数据库中，从而使得对连接数据的处理变得简单高效。此外，由于图在管理连接数据方面具有多功能性，因此其他类别的数据也非常适合使用图模型。正如 Edd Wilder-James[2012]所说：

> 即使数据类型基本匹配，关系数据库的一个缺点仍然在于其模式具有静态性。在灵活的探索环境中，计算结果将随着对更多信号的检测和提取而不断变化。而半结构化 NoSQL 数据库满足了对这种灵活性的需求：它们提供了足够灵活的结构来整理数据，但在存储数据之前不需要知道数据的确切模式。

作为 NoSQL 数据库的一种，图数据库也不例外。这种基于节点和边的简单模型在数据表示方面展现出极大的灵活性。此外，新类型的节点和边可以在设计过程的后期呈现，而不会影响先前定义的模型，由此赋予了图高度的可扩展性。

2.1.4 真实性

真实性(图 2.9)与所收集数据的质量和/或可信度有关。只有当数据正确、有意义且准确时，数据驱动的应用程序才能受益于大数据。

假设你想利用评论为旅游网站创建推荐引擎。这类引擎是推荐领域的新趋势，因为评论包含的信息比传统的星级评定包含的信息多得多。问题在于此类评论的真实性：

> 对于在线零售商来说，打击虚假评论是现在的主要业务。如今，当评论提交给 TripAdvisor 时，评论会通过一个跟踪系统。该系统检查数百个不同的属性，涵盖基本数据点(如评论者的 IP 地址)和更详细的信息(如用于提交评论的设备的屏幕分辨率)[1]。

1 Parkin [2018]。

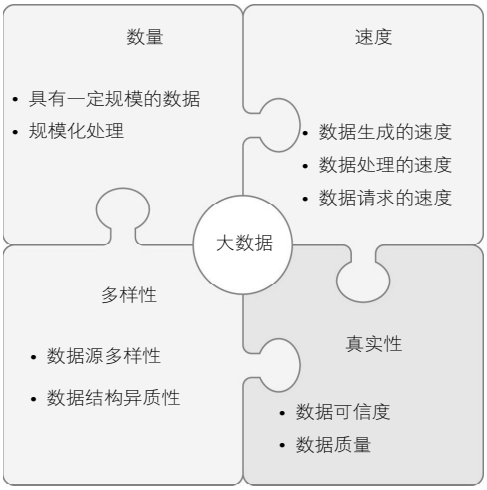

图 2.9　大数据的真实性

如第 1 章所述，训练数据集中数据的质量和数量会直接影响推断所得的模型质量。错误数据可能会影响整个处理流程。预测结果一定会受到影响。

数据很少会完全准确并完整，这就是必须对其进行清洗的原因。该任务可以用图方法来完成。具体来说，根据元素之间的关系，图访问模式使得可以轻松发现问题。此外，还可以通过在单一有关联的事实来源中组合多个来源，以此合并信息，从而减少数据稀疏性。

2.2　大数据平台中的图

处理大数据是一项复杂的任务。机器学习平台需要访问数据以提取信息并向终端用户提供预测服务。表 2.2 总结了与“四 V”相关的挑战，每个 V 与存储、管理、访问和分析数据的要求匹配。

表 2.2　大数据挑战

大数据的“四 V”	挑战
数量	数据库应该能够存储大量数据，或定义的模型应该能够通过聚合压缩数据，从而实现快速访问，但模型可以执行所有所需的分析
速度	通常数据传输速度很快，因此需要一个队列来解耦数据获取与存储机制。获取率和存储率必须充分平衡，以便可以对数据进行快速转换和存储，防止队列中元素的累积
多样性	数据库模式应具备较强的灵活性，从而同时存储多种信息，使用的模型可以存储当前所有类型数据以及项目后期可能会出现的任何类型数据
真实性	设计的模型和选择的数据库应该可以简单、快速地导航数据并识别不正确、无效或不需要的数据。还需要清洗数据，从而消除噪声。简化和合并数据(以对抗数据稀疏性)的任务也应受到支持

应对这些挑战的方法可以分为两类：方法论方法(或设计)和技术方法。为了获得最佳效果，你可将二者有机结合使用。

方法论方法包括所有涉及架构、算法、存储模式和清洗方法的设计决策，我们将在本节中结合一些特定的具体场景来研究这些方法。

技术方法包括与要使用的 DBMS、要采用的集群配置以及要提供的解决方案的可靠性相关的设计方面。这些内容在 2.3 节中介绍。

这里展示了两个场景，以说明图在管理大数据并从中提取信息方面的价值，其中大数据作为机器学习项目流程中的一部分。虽然这两个场景都与方法论方法相关，但它们也强调了技术方法的某些方面，相关内容将在本章后面讨论。

2.2.1 图对于大数据很有价值

为了了解图对于大数据的价值，我们将探索一个复杂用例，其需要处理大量数据才能起效。在本例中，只要有一组完整的特征来存储、处理和分析数据，那么图就可以处理问题复杂性。考虑以下场景：

你是一名警察。你如何用蜂窝塔的数据来跟踪嫌疑人？这些数据是通过连续监控每台手机向它能到达的所有蜂窝塔发送(或从其接收)的信号来收集的。

Eagle、Quinn 和 Clauset[2009]撰写的一篇有趣的文章解决了使用蜂窝塔数据的问题，这些数据是从手机与每个塔台交换的连续监测信号中收集的(图 2.10)。我们的示例场景旨在使用此类监控数据创建一个预测模型，该模型识别与目标生活相关的位置集群，并根据目标的当前位置来预测后续移动。

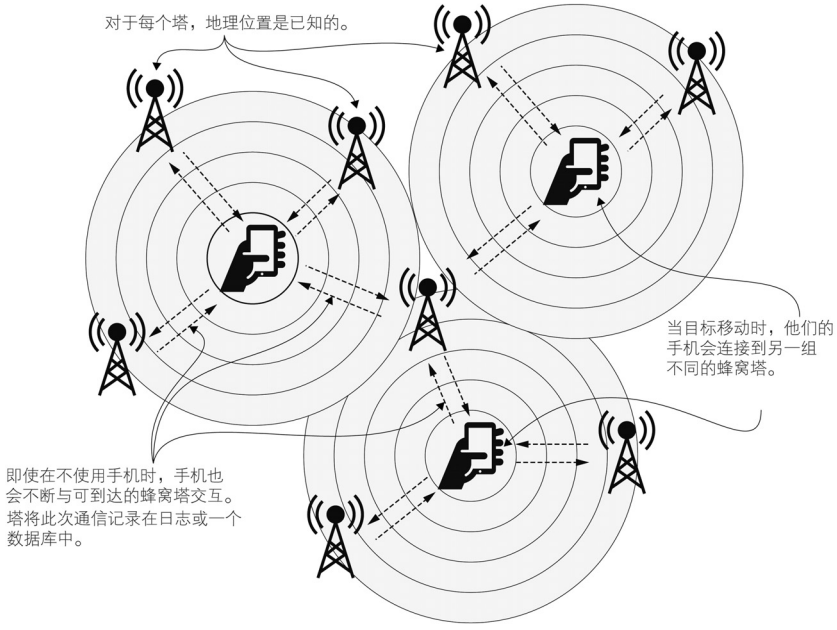

图 2.10 手机与蜂窝塔通信

这种方法的核心思想是，它使用图来折叠并整理来自蜂窝塔的可用数据，并针对目标的移动创建一个基于图的物化视图。使用可识别位置集群的图算法来分析生成的图。这些位置可用于分析流程中的下一个算法，从而构建位置预测模型。这里我不赘述算法的细节，因为该场景的目的是展示如何使用图模型来预处理和组织复杂问题中的数据。该图将信息压缩，使其适用于后续分析。

对于本场景中的心智模型，使用图模型来存储和管理数据源，处理流程中使用图算法，并使用图来存储中间模型(图 2.11)。

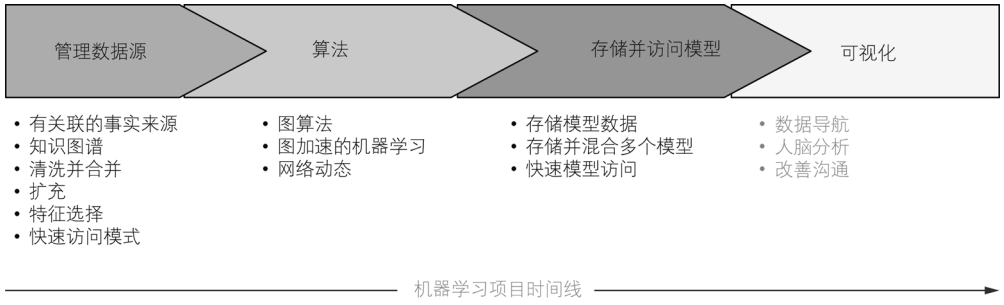

图 2.11　思维导图中与此场景相关的思维导图区域

今天，每部手机都可以持续访问有关附近蜂窝塔的信息。研究这些数据流可以很大程度上帮助了解用户的移动和行为。获取连续蜂窝塔数据的方法有：

- 在手机上安装日志应用程序以捕获连续的数据流。
- 使用来自蜂窝塔的原始连续数据(如果可用的话)。

本示例使用连续数据聚合过程来合并特定手机的数据，因此我们通过此场景中的第一个用例来简化叙述。

每个目标的手机每隔 30 秒就记录下四个最近的塔——信号最强的塔。数据被收集后，其可以表示为蜂窝塔网络(Cellular Tower Network，CTN)，其中节点是唯一的蜂窝塔。在同一记录中同时出现的每对节点之间有一条边，并且根据每对节点在所有记录中同时出现的总时间，对每条边进行加权。为每个目标生成一个 CTN，包括在监控期间由目标的手机记录的每个塔[Eagle、Quinn 和 Clauset, 2009]。图 2.12 显示了单个目标得到的图的示例。

对节点所有边的权重求和，从而确定节点的强度。总边权重最高的节点识别出最常靠近目标手机的塔。因此，高权重节点组应与目标停留了大量时间的位置相对应。基于这个想法，可以通过使用不同的聚类算法将图分成多个簇(从第 II 部分开始，我们将讨论这些算法)。这种聚类过程的结果将类似于图 2.13。

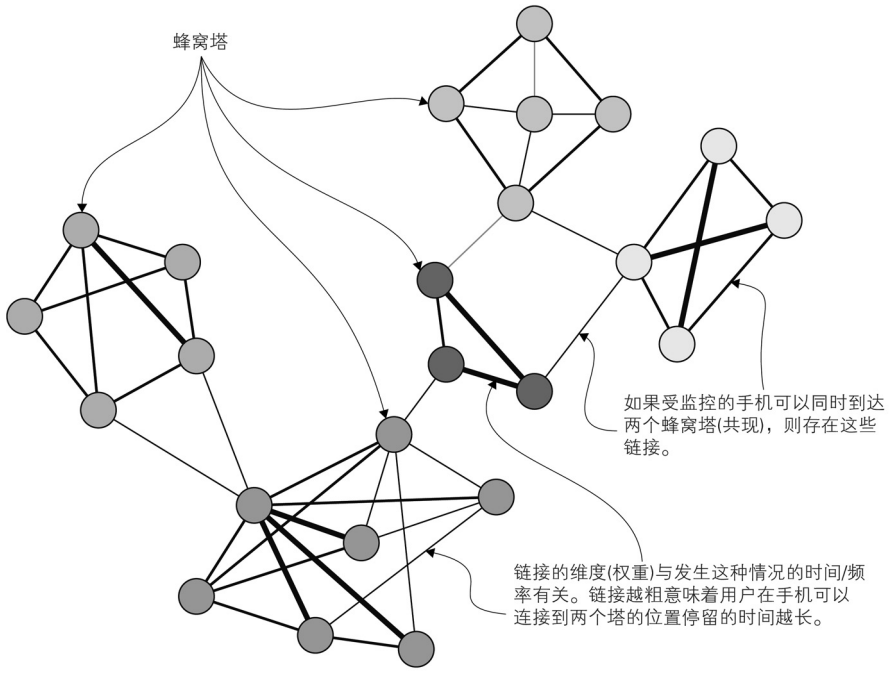

图 2.12 单个目标的 CTN 图表示

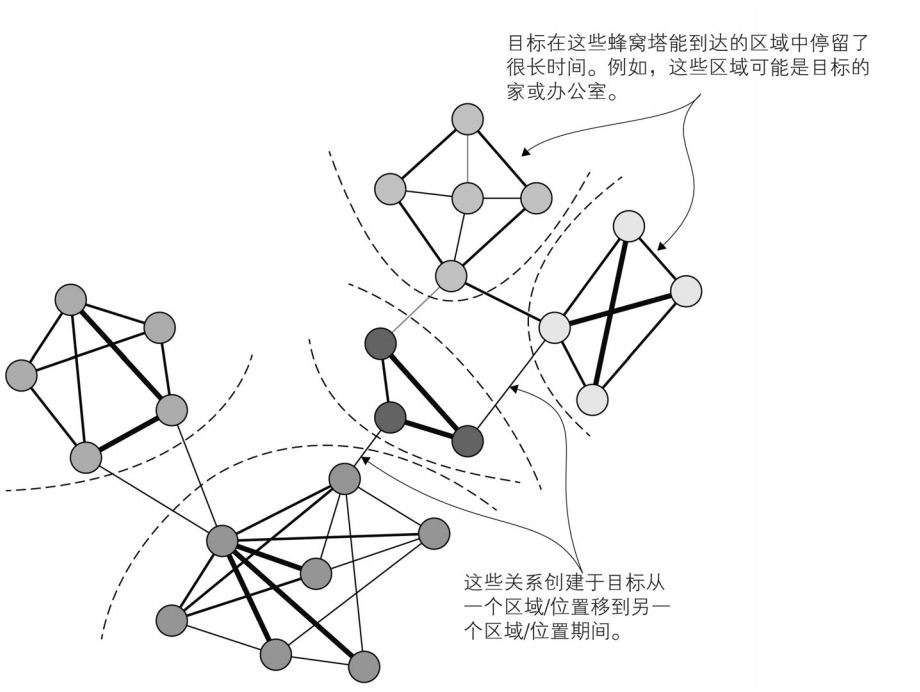

图 2.13 CTN 簇视图

簇代表目标停留时间较长的区域——可能是在家里、办公室、健身房或商店。簇间的关系(连接两个不同簇中的节点)代表两个不同区域之间的转换，例如早上从家移到办公室，下午则进行反向移动。

可识别蜂窝塔的簇可以转换为动态模型的状态。给定一个目标曾去过的一系列位置，就可以学习它们的行为模式并计算将来其移到不同位置的概率(为此可以使用不同的技术，但这些技术超出了本书的讨论范围)。在计算预测模型时，考虑到从同一数据源显示的最后位置，预测模型可以用于预测目标的未来移动。

此场景展示了一种方法论方法，其中图模型作为强大的数据视图，可用作某些机器学习算法的输入或分析师的可视化工具。此外，由于图模型无模式，所以历史数据的聚合版本(通常是大量数据，包含目标在每个位置停留的时间)和实时值(目标的手机当前能够访问的蜂窝塔，其可以表示他们当前所处的位置)可以共存于同一模型中。

从特定模型中，可以提炼出更通用的方法。问题和过程总结如下：

- 有大量事件形式的数据(可从蜂窝塔或手机获得的监控数据)。
- 数据分布在多个数据源(每个蜂窝塔或手机)中。
- 应以简化进一步流程和分析的形式聚合并整理数据(CTN)。
- 根据聚合格式，创建了一些视图(聚类算法和位置预测模型)。
- 同时，需要存储最近事件的实时视图，以便对这些事件做出快速反应。

此数据流的一些重要相关方面会影响机器学习项目架构：

- 由手机或蜂窝塔记录的事件是原始的、不变的且真实的。分析并不会改变已发生的事件。当手机连接到蜂窝塔时，该事件不会因为分析目的不同而发生改变。有必要将这些事件以原始格式一次性存储起来。
- 在此数据上创建多个视图作为函数(聚合就是一个示例)，视图随用于分析的算法而改变。
- 通常是在整个数据集上构建视图，这个过程较为耗时，尤其在处理大量数据时更是如此，正如我们这一特定用例所示。处理这些数据所耗费的时间导致当前事件的视图和先前事件的视图之间产生差距。
- 为获得数据的实时视图，就必须弥补这一差距。实时视图需要一种流过程来读取事件并向视图中添加信息。

Nathan Marz 和 James Warren 所著 *Big Data: the Lambda Architecture* 一书中介绍的一种特定类型的架构可以解决上述架构问题，该架构“提供了一种在任意数据集上运用任意函数的通用方法，使函数以低延迟返回其结果”[Marz and Warren, 2015]。Lambda 架构的主要概念是将大数据系统构建为三层：批处理、服务和速度。架构模式如图 2.14 所示。

每一层都满足一部分需求，并以其他层提供的功能为基础。一切都从方程 query = function(all data)开始。在我们的示例场景中，查询为“获取用户长时间停留的位置”。

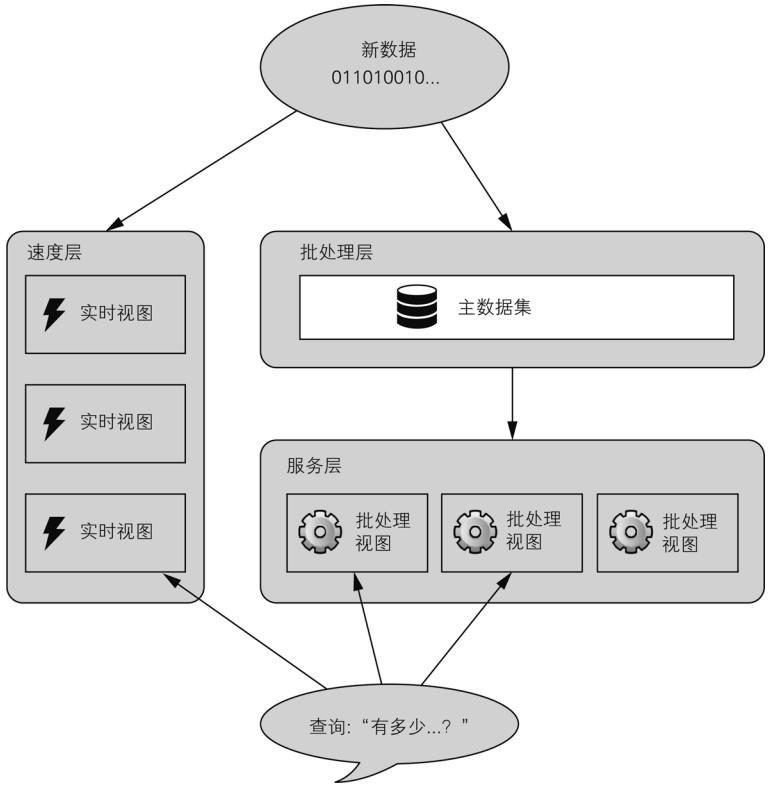

图 2.14 Lambda 架构[Marz 和 Warren, 2015]

理想情况下，我们可以即时运行查询以获取结果。但是由于要处理的数据量很大和要访问的数据源具有分布式特性，这种方法通常不可行。而且，这会占用大量资源，成本十分高昂，而且耗时较长。它也不适合任何实时监控和预测系统。

首选的替代方法是预先计算查询函数的结果或能加速最终查询结果的中间值。我们将预先计算的(最终或中间)查询结果称为批处理视图。我们并不动态计算查询，而是从这个视图中计算结果，这使得需要以一种提供快速访问的方式进行存储。那么前面的方程分解如下：

batch view = function(all data)

query = function(batch view)

所有原始格式的数据都存储在批处理层中。该层还用于原始数据访问以及批处理视图计算和提取。产生的视图存储在服务层中，其中以适当的方式对它进行索引并在查询期间进行访问(图 2.15)。

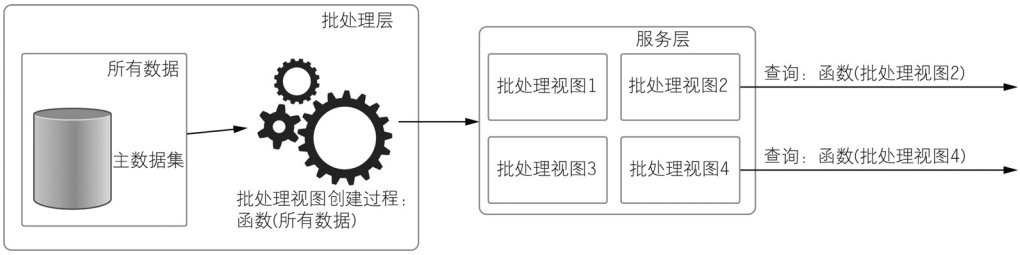

图 2.15　批处理层流程

在蜂窝塔场景中，第一个批处理视图是 CTN，这是一个图。处理这种视图的函数是图聚类算法，它可产生另一个视图：聚类网络。图 2.16 显示了以批处理视图的形式生成目标图的批处理层。为图 2.12 所示的每个受监控目标计算这些视图。

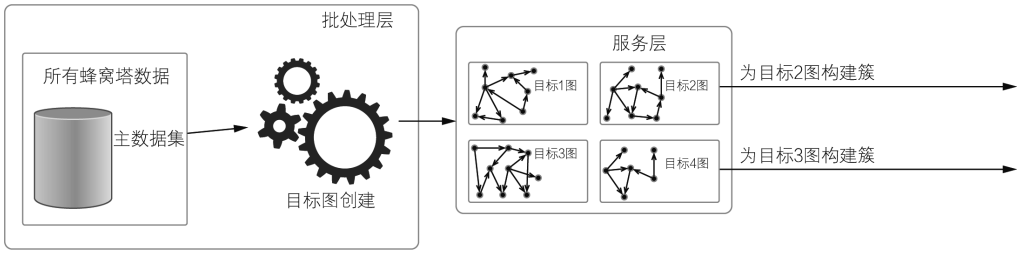

图 2.16　生成图视图的批处理层流程

图 2.16 中的主要区别在于批处理视图和后来的实时视图都被建模为图。这种建模决策在多个场景中具有优势，在这些场景中，图表示不仅使我们可以更快地回复查询，而且有助于分析。在所考虑的特定场景中，CTN 被创建为收集到的所有原始蜂窝数据的函数，并支持聚类算法。

每当批处理层完成对批处理视图的预计算时，服务层就会被更新。由于预计算需要时间，到达架构入口点的最新数据在批处理视图中并未体现。速度层通过提供最近数据的一些视图来解决这个问题，从而填补批处理层和最新传入数据之间的空白。这些视图可以具有与服务层相同的结构，也可以具有不同的结构并用于不同的目的。

两个层之间的一大区别是速度层只查看最近的数据，而批处理层查看所有数据。在蜂窝塔场景中，实时视图提供有关目标的最后位置和位置之间转换的信息。

Lambda 架构与技术无关；可以使用不同的方法来实现它。具体使用哪种技术可能会根据需求而变化，但 Lambda 架构定义了一种一致的方法来选择这些技术并将它们连接在一起以满足你的要求。

本节中介绍的场景展示了如何使用图模型作为服务层和速度层的一部分。由此产生的最终架构表示为图 2.17。

这种新的 Lambda 架构子类型可以定义为基于图的 Lambda 架构。在本例中，速度层仅根据每个目标的手机到达的最后一个蜂窝塔来跟踪每个目标的最后一个已知位置。然后将这些信息与聚类和预测模型结合使用，以预测目标将到达的位置。

图 2.17 架构中的主数据集可以存储在任何数据库管理系统或一个可以容纳大量数据的

简单数据存储中。最常用的数据存储是 HDFS、Cassandra 和类似的 NoSQL DBMS。

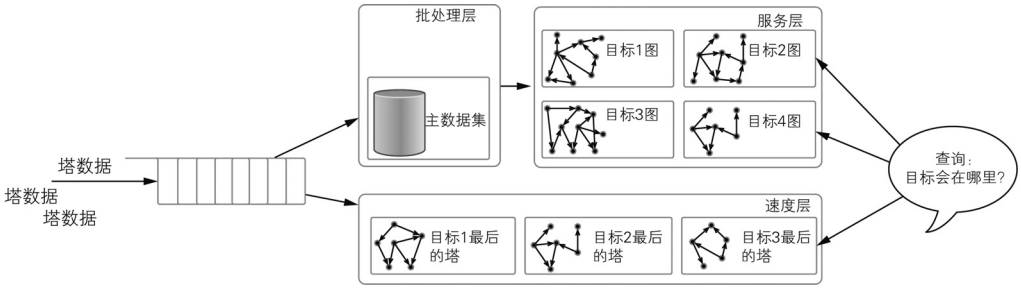

图 2.17 基于图的 Lambda 架构

在所考虑的特定场景中，需要先合并数据，然后再以图的形式提取数据。HDFS 基于文件系统存储提供基本访问机制，而 Cassandra 提供的访问模式更灵活。存储图视图需要图数据库。2.3 节介绍了此类数据库和特定工具的一般特征。

基于图的 Lambda 架构和这里描述的场景便是一个示例，可以说明图在机器学习领域中对于大数据分析所起的重要作用。可以在多个场景中使用相同的架构，包括：

- 分析银行交易以检测欺诈。
- 分析网络农场中的服务器日志以识别网络攻击。
- 分析通话数据以识别人群。

2.2.2 图对于主数据管理意义重大

在 2.2.1 节中，我们了解了如何使用图在主数据集(批处理层)或部分可用数据的实时(速度层)表示之上创建视图。在这种方法中，交易数据存储在主数据集中。当你能查询并分析存储在服务层中的聚合数据时，该方法非常有用。

但是，无法在数据的聚合版本中执行其他类型的分析。此类算法需要更详细的信息才能生效；它们需要访问数据的细粒度版本来完成任务。这种类型的分析还可以使用图模型来表示连接并从数据中提取信息。理解数据之间的联系并从这些链接中得出意义提供了不基于图的传统分析方法无法提供的功能。在这种情况下，图是知识的主要来源，它对一个单一有关联的事实来源进行建模。这个概念使我们思考第二个示例场景：

你要为银行创建一个简单但有效的欺诈检测平台。

银行和信用卡公司每年因欺诈而损失数十亿美元。传统的欺诈检测方法(例如基于规则的方法)在减少损失方面发挥着重要作用。但是欺诈者不断开发越来越复杂的方法来逃避检测，这使得基于规则的欺诈检测方法变得脆弱不堪，很快就被淘汰。在此，我们将关注一种特定类型的欺诈：信用卡盗窃。犯罪分子可以通过多种方法窃取信用卡数据，包括使用安装在自动提款机中的蓝牙数据扫描设备、大规模的黑客入侵，或者供收银员或餐厅工作人员刷卡的小型设备。任何可以合法使用你的卡的人甚至可以将卡的信息记录下来[Villedieu, n.d.]。为了揭露这种欺诈行为，有必要确定“入侵”的来源：信用卡窃贼以及他们操作的地点。将信用卡交易表示为图，我们可以寻找共性并追踪欺诈的源点位置。与大

多数其他查看数据的方式不同，图旨在表达相关性。图数据库可以发现使用传统表示(如表格)难以检测的模式。

假设表 2.3 中的交易数据库包含对某些交易提出异议的用户子集的数据。

表 2.3　用户交易[a]

用户标识符	时间戳	金额	商家	有效性
用户 A	01/02/2018	250 美元	Hilton Barcelona	无争议
用户 A	02/02/2018	220 美元	AT&T	无争议
用户 A	12/03/2018	15 美元	Burger King New York	无争议
用户 A	14/03/2018	100 美元	Whole Foods	有争议
用户 B	12/04/2018	20 美元	AT&T	无争议
用户 B	13/04/2018	20 美元	Hard Rock	无争议
用户 B	14/04/2018	8 美元	Burger King New York	无争议
用户 B	20/04/2018	8 美元	Starbucks	有争议
用户 C	03/05/2018	15 美元	Whole Foods	无争议
用户 C	05/05/2018	15 美元	Burger King New York	无争议
用户 C	12/05/2018	15 美元	Starbucks	有争议

a 此处以商家名称为例，以使用例更加具体。

我们可根据该交易数据集定义一个图模型。每笔交易都涉及两个节点：一个人(客户或用户)和一个商家。节点由交易本身链接。每笔交易都有一个日期和一个状态：合法交易无争议；欺诈交易有争议。图 2.18 将数据显示为图。

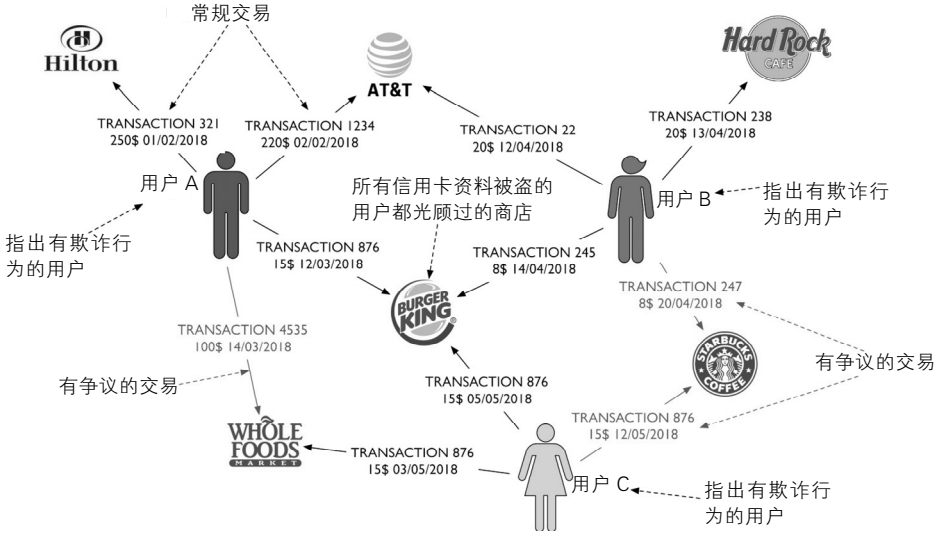

图 2.18　信用卡欺诈检测示例图模型

在这个图中，红色连接是有争议的交易；其他连接是常规(无争议的)交易。得到的图很大，但图的维度不会影响必须执行的分析类型。从这种数据集中可得，找出欺诈来源的分析步骤如下：

(1) 过滤欺诈交易。确定被攻击的人和卡的信息。

(2) 找出欺诈的源点位置。搜索欺诈开始前的所有交易记录。

(3) 隔离窃贼。确定一些常见模式，例如常见的商家，这可能是欺诈的源点位置。

根据图 2.18 中的样本图，欺诈交易和受影响的人列于表 2.4 中。

表 2.4 有争议的交易

用户标识符	时间戳	金额	商家	有效性
用户 A	14/03/2018	100 美元	Whole Foods	有争议
用户 B	20/04/2018	8 美元	Starbucks	有争议
用户 C	12/05/2018	15 美元	Starbucks	有争议

所有这些交易都发生在不同的月份。现在，对于每个用户，找出在有争议交易发生日期之前其进行的所有交易以及相关商家。结果如表 2.5 所示。

表 2.5 有争议的交易发生前的所有交易

用户标识符	时间戳	金额	商家	有效性
用户 A	01/02/2018	250 美元	Hilton Barcelona	无争议
用户 A	02/02/2018	220 美元	AT&T	无争议
用户 A	12/03/2018	15 美元	Burger King New York	无争议
用户 B	12/04/2018	20 美元	AT&T	无争议
用户 B	13/04/2018	20 美元	Hard Rock	无争议
用户 B	14/04/2018	8 美元	Burger King New York	无争议
用户 C	03/05/2018	15 美元	Whole Foods	无争议
用户 C	05/05/2018	15 美元	Burger King New York	无争议

我们按商店名称对交易进行分组。结果列于表 2.6 中。

表 2.6 聚合交易

商家	数量	用户
Burger King New York	3	[用户 A、用户 B、用户 C]
AT&T	2	[用户 A，用户 B]
Whole Foods	1	[用户 A]
Hard Rock	1	[用户 B]
Hilton Barcelona	1	[用户 A]

从这张表中可以清楚地看出，窃贼在 Burger King 餐厅进行了欺诈行为，因为这是所有用户都光顾过的唯一商家，并且在每个案例中，欺诈都发生在那里进行的交易之后。

根据这个结果，可进行深入分析，使用图搜索其他类型的模式，并将搜索结果转化为

一个防范行动，阻止来自已识别源点的任何进一步交易，查到进行更深入的调查。

在这种情况下，图用于存储单一事实来源，通过使用图查询对其进行分析。此外，可以以图的形式将数据可视化，以供进一步分析和调查。相关的心智模型如图 2.19 所示。

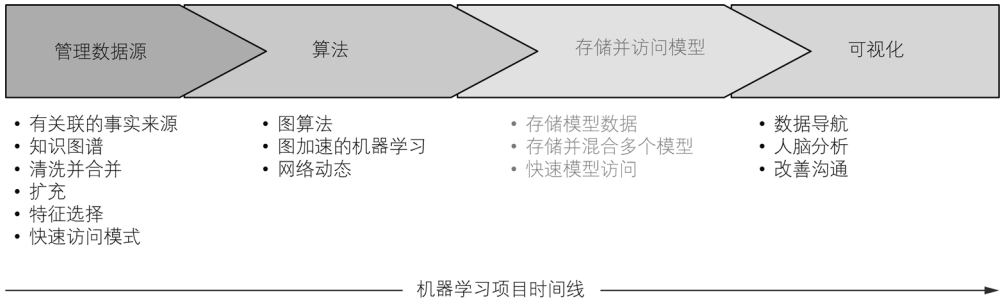

图 2.19　思维导图中与此场景相关的区域

此示例简化了揭示欺诈的方法，但它展示了使用图数据库进行此类分析的一些优势。这些优势可以总结如下：

- 可以将多个数据源(例如地理位置或 GPS 信息、社交网络数据、用户个人资料、家庭数据等)合并到单一有关联的事实来源中。
- 可以使用外部知识源(商店位置、人员地址等)或上下文信息(新商店、其他投诉等)扩展现有数据，并且这些信息可用于改进分析。
- 同一数据模型可以支持多种分析技术(例如用于发现诈骗团伙[1])。
- 可以将数据可视化为图以加速手动分析。
- 考虑到多个跃点，分析可以扩展到多个级别的交互。
- 由于图模型提供了灵活的访问模式，该结构简化了合并和清洗操作。

在欺诈分析场景中，图代表了已合并、清洗和扩展数据的主要知识来源，在此基础上执行分析，并基于此做出决策。与 2.2.1 节中描述的基于图的 Lambda 架构不同，这里的图作为主数据集，并且还是主数据管理(Master Data Management，MDM)的基础，MDM 用于识别、清洗、存储和管理数据[Robinson et al., 2015]。MDM 的主要关注点包括：

- 管理由组织结构改变、业务合并和业务规则变化所导致的变化。
- 合并新的数据源。
- 用外部数据源补充现有数据。
- 满足报告、合规性和商业智能消费者的需求。
- 当数据的值和模式更改时，更新数据版本。

尽管 MDM 和 DW(Data Warehousing，DW)这两种实践有很多共同之处，但 MDM 不是数据仓库的替代或现代版本。DW 存储历史数据，而 MDM 处理当前数据。MDM 解决方案包含某公司内所有业务实体的现有完整信息。DW 只包含历史数据，并将其用于某种静态分析。如果操作正确，MDM 具有许多优点，可总结如下：

1 根据法律词典(https://thelawdictionary.org/fraud-ring)，诈骗团伙是“一个专注于诈骗他人的组织。伪造、虚假索赔、盗用身份、伪造支票和货币都是欺诈行为。”

- 简化人员和部门之间的数据共享。
- 促进多系统架构、平台和应用程序中的计算。
- 消除数据的不一致和重复。
- 减少搜索信息时的不必要失败。
- 简化业务流程。
- 改善整个组织间的沟通。

此外，如果有适当的 MDM 解决方案，系统提供的数据分析就更加可靠，从而基于该数据所做的决策也更加可信。在这种情况下，图数据库“不提供完整的 MDM 解决方案，但它们非常适合于建模、存储和查询层次结构、主数据元数据和主数据模型。这些模型包括类型定义、约束、实体之间的关系以及模型与潜在源系统之间的映射”[Robinson et al., 2015]。

基于图的 MDM 具有以下优点：

- 灵活性——可以轻松更改捕获的数据，从而使其包含其他属性和对象。
- 可扩展性——该模型中，模型使得主数据模型可以根据业务需求的变化快速演变。
- 搜索功能——每个节点、每个关系以及它们的所有相关属性都是搜索入口点。
- 索引功能——图数据库由关系和节点自然索引，与关系数据相比其访问速度更快。

基于图的 MDM 解决方案处理不同类型的功能。在欺诈检测场景以及本书的其余部分中，它被认为是分析/机器学习平台的一部分，作为主要数据源运行并由结果模型进行扩展，代表了基于图的 Lambda 架构的一种替代方法。

另一个有趣的场景是在推荐引擎中，其中图可以表示存储数据用于训练的 MDM 系统。在这种情况下，图可以存储用户到条目(user-to-item)矩阵，其包含用户和条目之间的交互历史。我们将在第 3 章中更详细地介绍这种情况。

2.3 图数据库

2.2 节介绍了一些使用图的机器学习方法论方法，并提供了具体示例，说明了如何使用图作为数据的存储和访问模型，以此来增强预测分析能力。要以最佳方式使用此类模型，存储、操作和访问图的方式必须能与在数据流或算法中处理它们的方式相同。要完成此任务，你需要一个图数据库来作为存储引擎。这个图数据库可供你存储和操作实体(也称节点)以及这些实体之间的连接(也称关系)。

本节描述与图管理相关的技术方面。这种观点与机器学习项目的整个生命周期相关，在此期间你所操作、存储和访问的必须是真实数据。此外，在大多数情况下，你将处理大数据，因此必须考虑可扩展性问题。本节介绍了分片(跨多个服务器水平划分数据)和复制(跨多个服务器复制数据，以实现高可用性和可扩展性)。

许多图数据库都可用，但并非都是原生的(即一开始就是为图构建)；相反，它们在非图存储模型之上提供了一个图“视图”。这种非原生方法会导致存储和查询时出现性能问题。另一方面，适当的原生图数据库使用图模型来存储和处理数据，使图操作简单、直观且高效。2.3.4 节中重点介绍这两种图数据库的主要区别。

在许多情况下，也可以说几乎所有情况下，你需要使用至少一个节点标识，所以向节点和关系添加一些属性是很有帮助的。换句话说，必须对同一类中的节点进行分组。这些“特征”极大地提高了图数据库的表达能力和建模能力。2.3.5 节介绍了满足这些需求的标签属性图。

尽管本书中介绍的所有理论、示例和用例都与技术完全无关，但我们将使用 Neo4j(在附录 B 中介绍)作为参考数据库平台。Neo4j 不仅是为数不多的可提供高性能的可用图数据库之一，还拥有强大且直观的查询语言，其被称为 Cypher[1]。

2.3.1　图数据库管理

要想在机器学习项目中使用图，你需要存储、访问、查询并管理每个图。所有这些任务都属于图数据库管理的一般类别，如图 2.20 所示。

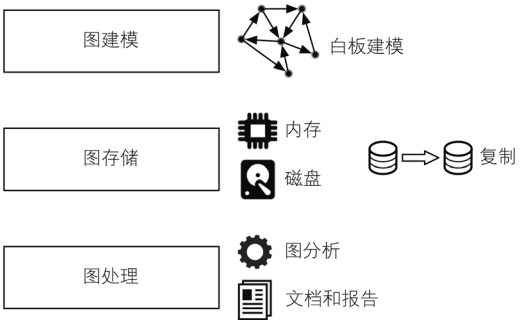

图 2.20　图数据库管理任务

该图显示了可以将图数据管理任务分组为三个主要区域：

- 图建模——一般来说，图模型是数据库系统蕴含的基本抽象概念——是一种用于建模现实世界实体及实体间关系的概念工具。建模非结构化数据所具有的简单性是图结构的主要特征之一。图模型中模式和数据之间的分离程度低于传统关系模型中的分离程度[Angles 和 Gutierrez, 2017]。同时，图模型具有灵活性和可扩展性。在图模型中可以用多种方式对现实的相同方面或相同问题进行映射。不同的模型可以从不同的角度解决不同的问题，因此定义正确的模型需要努力尝试和经验积累。幸运的是，图的“无模式”性质意味着，即使是对于在项目早期阶段定义的模型，花费较少的精力也可以对其进行更改。另一方面，当你使用其他类型的 NoSQL 数据库或关系数据库时，更改模型可能需要完全重新导入数据。但这种对于大数据的操作极其耗费时间、人力和财力。此外，模型设计会影响在图上执行的所有查询和分析的性能。因此，建模是数据管理的一个重要方面。本章前面介绍了一些模型示例，在接下来的章节中，我们将查看更多用例并讨论相关模型的优缺点。

1 https://www.opencypher.org。

- 图存储——定义模型时，必须将数据存储在持久层中。图 DBMS 专门管理类似图的数据，遵循数据库系统的通用原则：持久数据存储、内存使用、缓存、物理/逻辑数据独立性、查询语言、数据完整性和一致性等。此外，图数据库供应商必须处理与可扩展性、可靠性和性能相关的所有方面，如备份、恢复、水平和垂直可扩展性以及数据冗余。在本节中，我们将讨论图 DBMS 的关键概念——重点讨论在处理过程中影响模型以及数据访问方式的概念。
- 图处理——这些任务涉及用于处理和分析图的框架(例如工具、查询语言和算法)。有时，处理图需要使用多台机器来提高性能。一些处理特征(如查询语言和一些图访问模式)在图 DBMS 上可用；其他特征可用作算法或外部平台，必须在图和图 DBMS 之上实现。

图处理是一个涵盖内容广泛的主题，相关任务大致可以分为几类。Özsu[2015]提出了一种对图处理进行分类的有趣方法，根据三个维度进行分类：图动态、算法类型和工作负载类型(见图 2.21)。

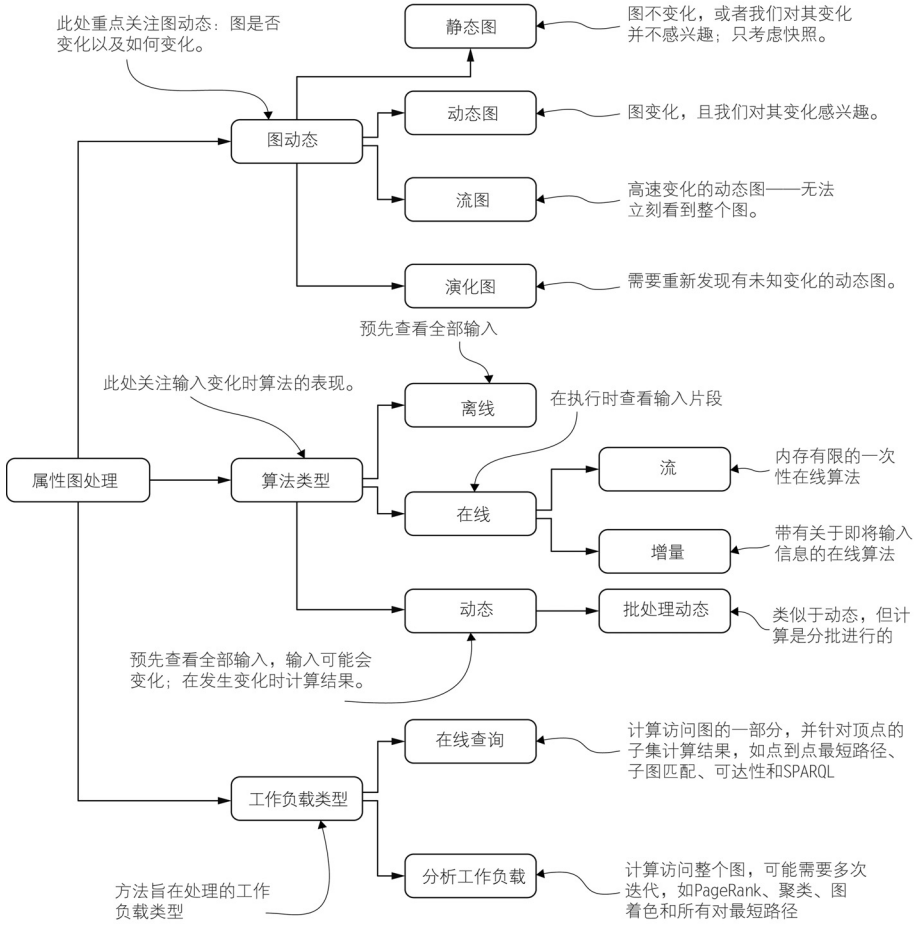

图 2.21 属性图处理分类[Özsu, 2015]

图处理这一复杂主题将贯穿全书。讨论过程中将介绍不同的算法，并将它们与特定的实际用例或应用程序示例进行映射，在这些例子中，算法有助于提取信息。

2.3.2　分片

只从数据存储的角度来看大数据应用，主要的“四 V”挑战如下：

- 数量——涉及的数据量很大，导致很难将数据存储在单一机器上。
- 速度——一台机器只能为有限数量的并发用户提供服务。

尽管垂直扩展(例如添加增加的计算、存储和内存资源)可以作为临时方案，用于处理大量数据并缩短多个并发用户的响应时间，但最终数据会变得太大而无法存储在单个节点上，且用户数量太多，导致单台机器不足以对其进行处理。

在 NoSQL 数据库中，一种常见的扩展技术是分片，这种技术将一个大型数据集划分为多个子集，子集分布在不同服务器上的多个分片中。这些分片或子集通常跨多个服务器复制而来，从而提高了可靠性和性能。分片策略决定将哪些数据分区发送到哪些分片。可以通过各种策略来实现分片，其可以由应用程序管理，也可以由数据存储系统本身管理。

对于面向聚合的数据模型，如键/值、列族和文档数据库[Fowler 和 Sadalage, 2012]，表达概念之间关系的唯一方法就是使用值或文档将它们聚合在单个数据条目中，这是较为明智的解决方案。在这些类型的存储中，用于检索任何条目的键是已知且稳定的，并且查找机制速度很快且可预测，因此很轻松便可以将想要存储或获取数据的客户指引到适当的分片[Webber, 2011]。

另一方面，图数据模型是高度以关系为导向的。每个节点都可以与任何其他节点相关，因此图不能进行可预测的查找。它的结构还高度可变：即使新的链接和节点很少，链接结构也可能发生重大变化。在这些情况下，对图数据库进行分片难度较大[Webber, 2011]。一种可能的解决方案是共同定位相关节点，从而共同定位相关边。这种解决方案将提高图的遍历性能，但在同一个数据库分片上，连接节点太多会使其负载过重，因为大量数据将出现在同一个分片上，从而导致分布不平衡。图 2.22 和图 2.23 说明了这些概念。

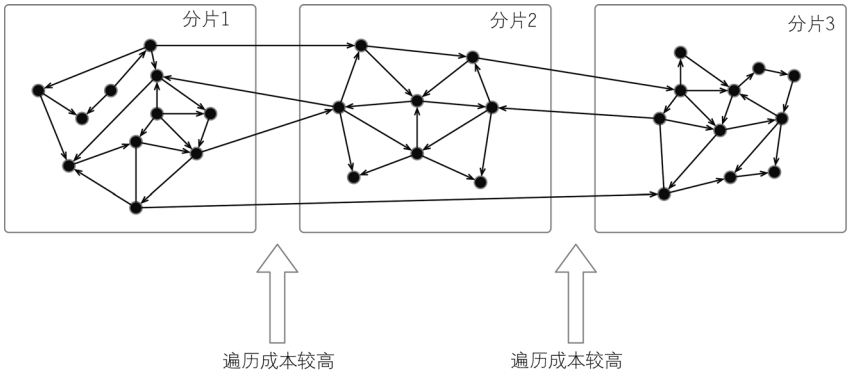

图 2.22　属于不同分片的遍历关系

图 2.22 显示了导航一个图可能涉及多次跨越分片边界。这种跨分片遍历成本较高，因为它需要很多网络跃点，导致查询时间大大延长。在这种情况下，与所有操作都发生在同一个分片上的情况相比，其性能会迅速下降。

在图 2.23 中，为了克服这个问题，将相关节点存储在同一个分片上。图遍历速度更快，但分片之间的负载高度不平衡。此外，由于图具有动态特性，运行时，图及其访问模式可以快速且不可预测地变化，使得该解决方案在实践中难以实现。

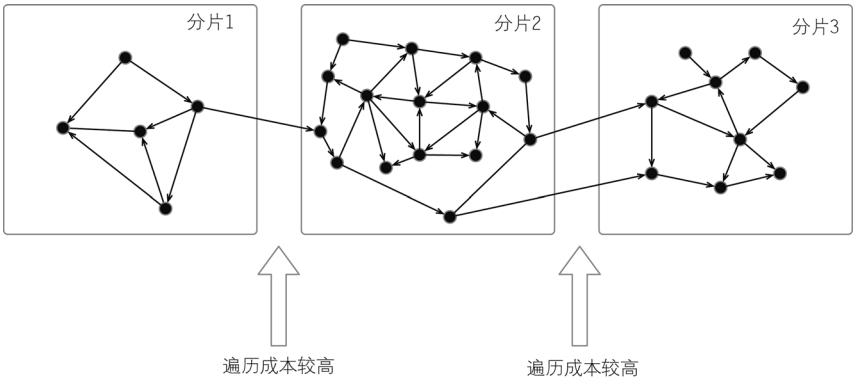

图 2.23 单个分片(分片 2)的过载

考虑到这些挑战，一般来说，有三种扩展图数据库的技术：

- *应用级分片*——在这种情况下，数据分片是在应用端通过使用特定领域的知识来完成的。对于全球业务，可以在一台服务器上创建与北美相关的节点，而在另一台服务器上创建与亚洲相关的节点。这种应用级分片需要了解：节点存储在物理位置不同的数据库中。分片还可以基于必须对数据执行的不同类型的分析或图处理。在这种情况下，每个分片都包含执行算法所需的所有数据，并且可以跨分片复制一些节点。图 2.24 描述了应用程序级分片。
- *增加 RAM 或使用缓存分片* —— 可以垂直扩展服务器，添加更多 RAM 以使整个数据库适用于内存。这种解决方案使得图遍历速度极快，但对于大型数据库来说既不合理也不可行。在这种情况下，可以采用缓存分片技术，从而在容量远远超过主存空间的数据集上保持高性能。因为我们希望每个数据库实例上都存在完整的数据集，所以这里的缓存分片并不是传统意义上的分片。为了实现缓存共享，我们对每个数据库实例的工作负载进行分区，以增加针对给定请求命中热缓存的可能性(像 Neo4j 这样的图数据库中的热缓存是高性能的)。
- *复制*——通过添加更多(相同的)数据库副本作为具有只读访问权限的从数据库实例，可以实现数据库的扩展。当你将数量相对较多的只读从数据库实例与少量主数据库实例配对时，可以实现高级别的可扩展性。该技术在 2.3.3 节中描述。其他技术也有各自的优点和缺点，这里不一一讨论。

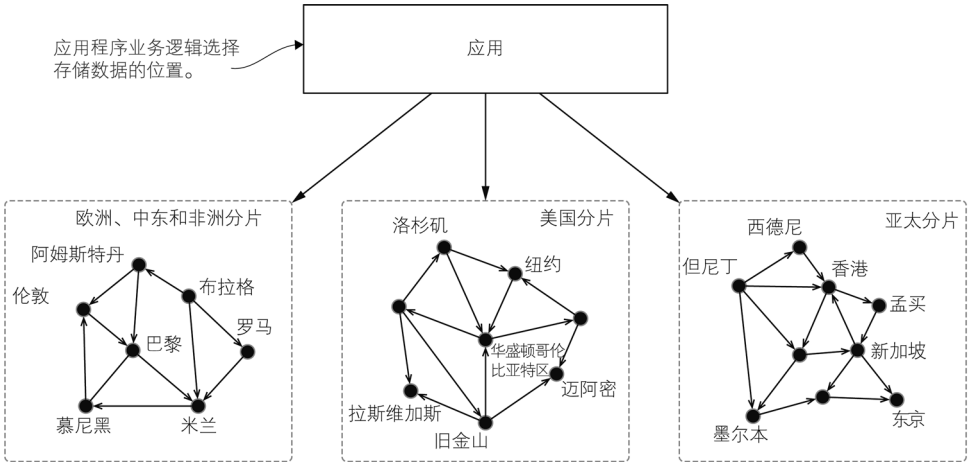

图 2.24　分片的应用级隔离

分片在某些情况下更有效。考虑本章前面讨论的两种情况：

- 在蜂窝塔监控示例中，为每个监控目标创建了一个图，因此机器学习模型会生成多个独立的、将被单独访问的图。在这种情况下，应用程序级分片非常简单，因为所有图都是独立存在的。总结来说，在基于图的 Lambda 架构场景中，通过在同一数据集上创建多个图视图，我们可以将这些视图存储在多个数据库实例中，因为可以独立对其进行访问。
- 在第二个用例(欺诈检测)中，分片会很困难，因为理论上所有节点都可以连接。可以应用一些启发式方法来减少跨分片遍历或将经常访问的节点保持在同一个分片上，但不能像前面的用例那样将图划分为多个孤立的图。在这种情况下，还有一种方法，即使用复制来扩展读取性能并加快分析时间。

2.3.3　复制

如 2.3.2 节所述，分片是图数据库中的一项艰巨任务。处理速度和可用性的一个有效替代方案是复制。数据复制包括在不同的计算机上维护多个称为副本的数据副本。复制有以下几个目的[Özsu 和 Valduriez, 2011]：

- *系统可用性*——复制使数据可以从多个站点访问，从而消除分布式 DBMS 存在的单点故障。即使某些集群节点关闭，数据也应该仍旧可用。
- *性能*——复制通过将数据定位在更靠近其访问点的位置，从而减少延迟。
- *可扩展性*——复制允许系统在地理空间和访问请求数量方面增长，同时保持可接受的响应时间。
- *应用程序要求*——作为其操作规范的一部分，应用程序可能需要维护多个数据副本。

数据复制的好处显而易见，但让不同副本保持同步并非易事。定义复制协议的基本设计决策是在何处首先执行数据库更新。这些技术的特点如下：

- 如果首先在主副本上执行更新，则为集中式。当系统中的所有数据项只有一个主数据库副本时，集中式技术可进一步被识别为单一主副本；当每个数据项(集)都可以有一个主副本时，集中式技术可以被识别为主副本。
- 如果允许对任何副本进行更新，则为分布式。

由于图具有高度连接性，实现集中式主副本协议或分布式协议较为困难，会严重影响系统性能和数据一致性，这一点是最关键的(在图中，数据项可以是一个节点或一个关系；根据定义，一个关系连接到另外两个数据项——节点——并且一个节点可能通过多个关系连接到其他节点)。因此，我们将重点关注单一主副本的集中式方法，也称主/从复制。

在这种方法中，将一个节点指定为数据的权威来源，称为主节点、领导节点或主要节点。该节点通常负责更新该数据。即使从节点设备接受写入，这些操作也必须通过要执行的主节点(见图 2.25)。如果你的大部分数据访问是读取，则主/从复制最有用。通过添加更多从节点并将所有读取请求路由到从节点，你可以将其水平扩展。主/从复制还提供弹性读取：如果主节点失败，从节点仍然可以处理请求[Fowler 和 Sadalage, 2012]。

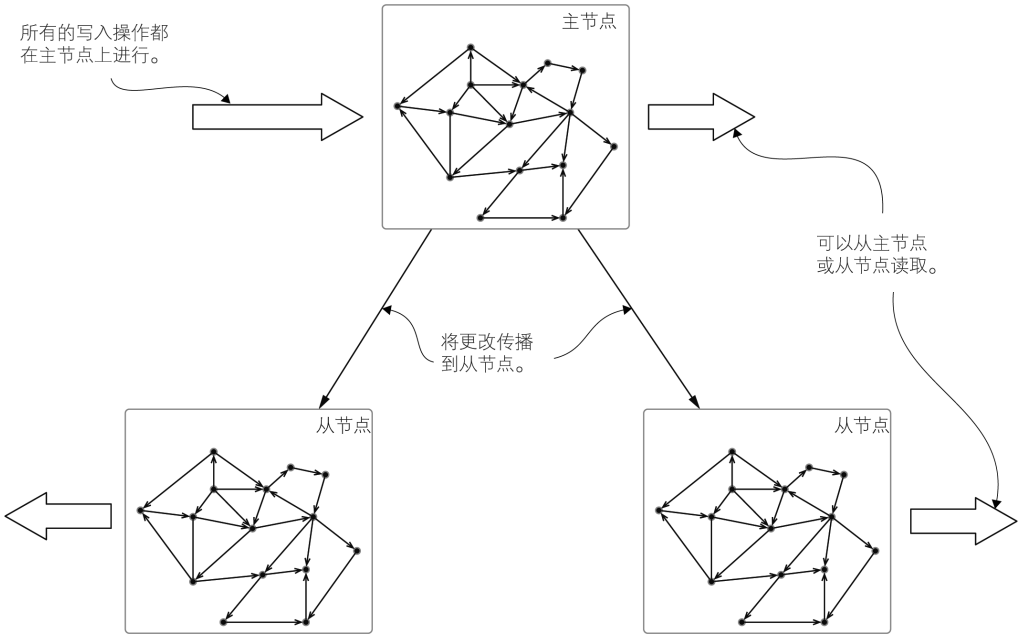

图 2.25 基于主/从协议的复制

大多主/从协议的实现都允许从节点在当前主节点不可用时连接到不同的主节点。这种方法提高了架构栈的可用性和可靠性。具体来说，在机器学习项目中，复制使得可以在训练或预测阶段将读取负载分散到所有节点。

2.3.4　原生与非原生图数据库

本书描述了多种方法，通过这些方法，图可以为机器学习项目提供支持。为了最大化利用图模型，需要使用一个合适的图 DBMS 来存储、访问和处理图。尽管在多个图数据库实现中模型本身相当一致，但在不同的数据库引擎中有许多方法可以对图进行编码和表示。从查询语言到数据库管理引擎和文件系统，从集群到备份和监控，为处理整个计算栈中的图工作负载而构建的 DBMS 被称为原生图数据库[Webber, 2018]。原生图数据库旨在以一种不仅理解图而且能支持图的方式使用文件系统，这意味着对于图工作负载而言，它们既高效又安全。更详细地说，原生图 DBMS 表现出一种名为无索引邻接的特性，这意味着每个节点都维护对其相邻节点的直接引用。邻接表是表示稀疏图的最常见方式之一。

形式上，图 $G=(V,E)$的这种表示由列表数组 *Adj* 组成，每个 *Adj* 对应 V 中的每个顶点。对于 V 中的每个顶点 u，邻接列表 *Adj*[u] 包含所有顶点 v，在 E 中的 u 和 v 之间存在一条边 E_{uv}。换句话说，*Adj*[u]由 G 中与 u 相邻的所有顶点组成[Cormen et al., 2009]。

图 2.26(b)是图 2.26(a)中无向图的邻接表表示。例如，顶点 1 有两个邻节点 2 和 5，所以 *Adj*[1]是列表[2, 5]。顶点 2 有三个邻节点 1、4 和 5，所以 *Adj*[2]是[1, 4, 5]。其他列表的创建方式相同。这表示发并不重要，因为关系中没有顺序，列表中也没有特定的顺序；因此，*Adj*[1]可以是[2,5]，也可以是[5, 2]。

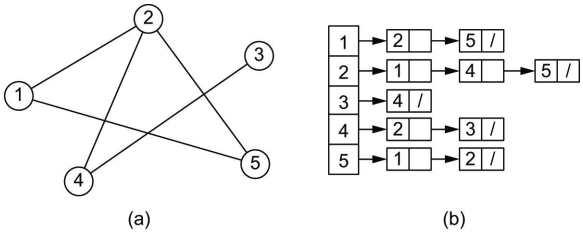

图 2.26　(a)无向图和(b)作为邻接表的相关表示

同理，图 2.27(b)是图 2.27(a)中有向图的邻接表表示。将这样的列表可视化为一个链接列表，其中每个条目都包含对下一个条目的引用。在节点 1 的邻接表中，第一个元素是节点 2，对下一个元素的引用是节点 5 的元素。这是存储邻接表最常用的方法之一，因为它可以提高添加和删除元素的效率。在这种情况下，我们只考虑传出关系，但对于传入关系，可以进行同样的操作；重要的是在创建邻接表的过程中选择一个方向并保持一致。这里，顶点 1 只有一个与顶点 2 的传出关系，所以 *Adj*[1]将是[2]。顶点 2 有两个传出关系，分别为 4 和 5，因此 *Adj*[2]为[4, 5]。顶点 4 没有传出关系，因此 *Adj*[4]为空([])。

如果 G 是有向图，则所有邻接表的长度之和为$|E|$。因为每条边都可以在一个方向上遍历，所以 E_{uv} 只会出现在 *Adj*[u]中。如果 G 是无向图，则所有邻接表的长度之和为 $2\times|E|$，这是因为如果 E_{uv} 是无向边，则 E_{uv} 会出现在 *Adj*[u]和 *Adj*[v]中。有向图或无向图的邻接表表示所需的内存与$|V|+|E|$成正比。

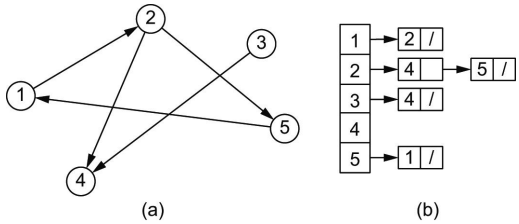

图 2.27 (a)有向图和(b)作为邻接表的相关表示

通过在 $Adj[u]$中存储边 E_{uv} 的权重 w，可以很容易地用邻接表表示加权图。对邻接列表表示也可以进行类似的修改，以支持其他变体图。在这样的表示中，每个节点都充当其他附近节点的微索引，这比使用全局索引的成本低得多。在这种数据库中，遍历关系的成本是恒定的，与图的大小无关。此外，查询次数与图的总大小无关；相反，它们与搜索的图数量成正比。

还有一种方案是使用非原生图数据库。这种方案中的数据库系统可以分为两类：

- 在现有不同数据结构之上对图 API 进行分层的系统，如键/值、关系、文档或基于列的存储。
- 多模型语义的系统，其中一个系统可以支持多个数据模型。

非原生图引擎针对替代存储模型进行了优化，如列、关系、文档或键/值数据，因此在处理图时，DBMS 与数据库主模型之间的转换成本昂贵。执行者可以尝试通过彻底的非归一化来优化这些转换，但在查询图时这种方法通常会导致高延迟。换句话说，非原生图数据库实际上永远不会像原生图数据库那样高效，原因很简单：它需要转换过程。

了解图的存储方式有助于为其创建更好的模型，图数据库的“原生”性质至关重要。这样的考虑也与本书的核心思想有关：

在一个成功的机器学习项目中，每一方面都与向最终用户提供高效和高性能的服务相关，其中高效和高性能不仅意味着准确，还意味着按时提供。

例如，如果网站上的精准推荐需要 30 秒才能提供，那么它将毫无用处，因为到那时，用户可能已切换到其他页面了。

有时，这些方面并非首要。常见的误解是认为非原生图技术已经足够好了。为了更好地理解机器学习项目数据库引擎中对图的原生支持的意义，可参考以下例子：

你必须执行一个供应链管理系统，其分析整个供应链，从而预测未来的库存问题或发现网络中的瓶颈。

供应链管理专业委员会将供应链管理定义为一个系统，该系统“涵盖采购、转换和所有物流管理活动中涉及的所有活动的规划和管理。”[1] 可以将供应链自然地建模为一个图，如图 2.28 所示。

1 http://mng.bz/8WXg。

现在假设你想利用关系数据库或任何其他基于全局索引的 NoSQL 数据库来存储供应链网络模型。供应链中各要素之间的关系如图 2.29 所示。

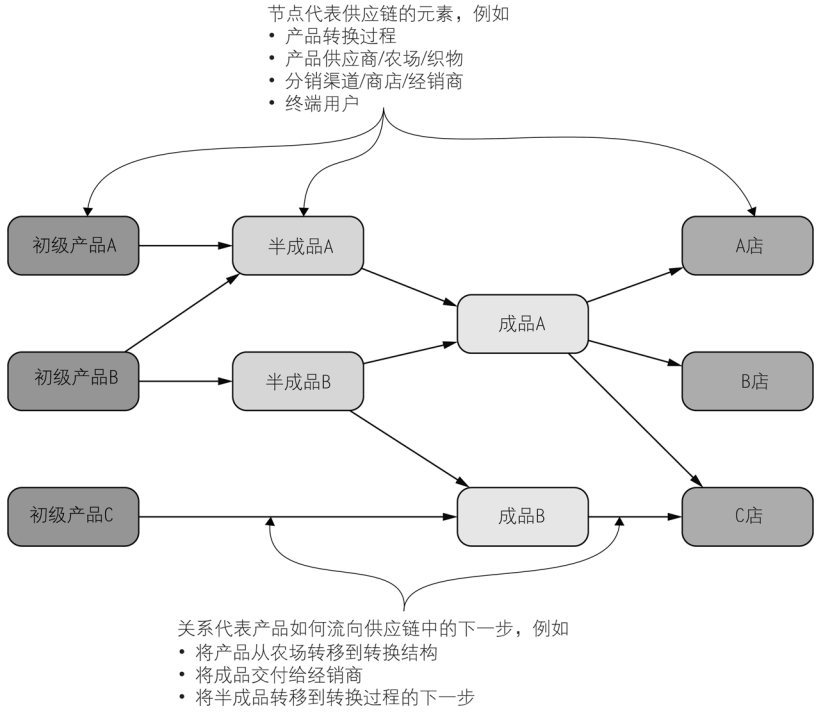

图 2.28　一个供应链网络

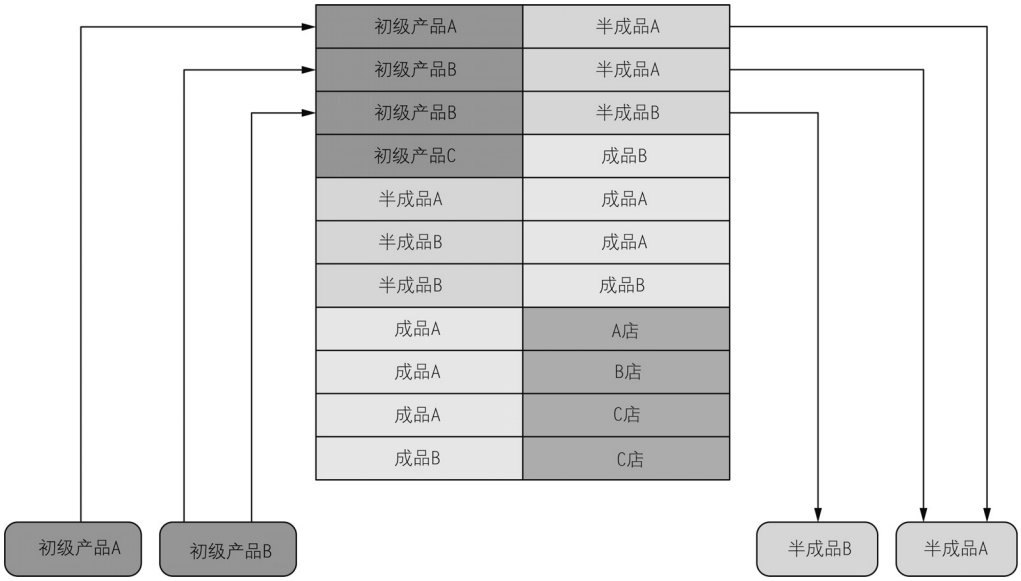

图 2.29　存储供应链网络的表格模型

如图所示，这些索引为每次遍历增加了一层间接性，从而增加了计算成本。要想查找产出成品 B 后交付的位置，我们首先必须进行索引查找，这需要耗时 O(log *n*)[1]，然后获取链中后续节点的列表。对于偶尔的或浅层查找，这种方法是可以接受的，但是当我们反转遍历方向(例如，为找到创建成品 C 所需的中间步骤)时，耗时则长得离谱。

假设现在初级产品 A 被污染或不再可用，我们需要在链中找到所有受此问题影响的产品或商店。因此我们不得不执行多个索引查找，对于初级产品和商店之间链中的每个节点执行一个索引查找，这使得耗时更长。找出成品 B 将交付到哪里需耗时 O(log *n*)，而要遍历 *m* 个步骤的网络，索引方法将耗时 O(*m* log *n*)。在具有无索引邻接的原生图数据库中，可有效地预先计算双向连接并将该连接作为关系存储在数据库中，如图 2.30 所示。

在这个表示中，当你有第一个节点时，遍历一个关系将耗时 O(1)，这意味着它直接指向下一个节点。现在，执行以前的相同遍历只需耗时 O(*m*)。图引擎不仅更快，而且成本仅与跳数(*m*)有关，与关系的总数(*n*)无关。

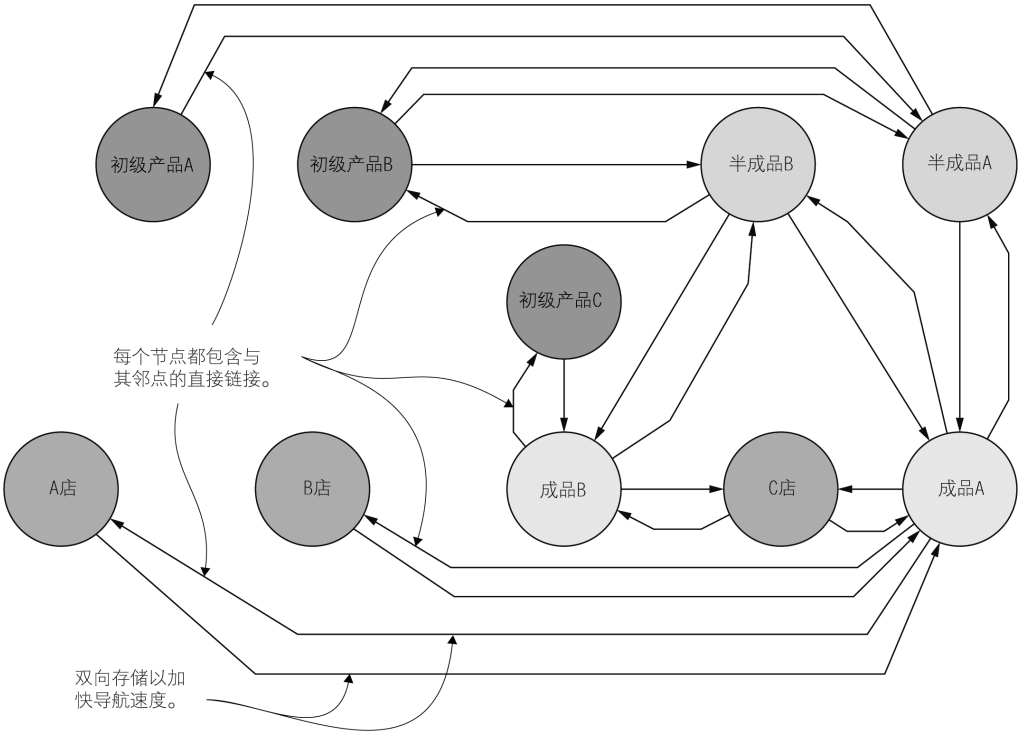

图 2.30 用于存储供应链的基于图的模型

现在假设你需要识别出供应链中的瓶颈。找出网络瓶颈的一种常用方法是利用介数中心性，其计算节点之间的最短路径，是对图的中心性(重要性)的一种度量。每个节点的介数中心性为通过该节点的最短路径的数量。在这种情况下，索引查找耗时——O(log*n*)——

1 大写 O 符号“用于描述计算机科学中算法的性能或复杂性。尤其用于描述最坏的情况，可用于描述算法所需的执行时间或使用的空间(例如在内存中或磁盘上)。”(来源和例子：https://mng.bz/8WXg)。

将极大地影响计算性能。

回顾一下，原生图架构具有许多优势，使其在管理图模型方面往往优于非原生方法。这些优势总结如下：

- *“分钟到毫秒”性能*——原生图数据库进行连接数据查询的速度远远快于非原生图数据库。即使在一般的硬件上，原生图数据库也可以轻松在单台机器上处理图中节点之间每秒数百万次的遍历和每秒数千次的事务性写入[Webber, 2017]。
- *读取效率*——原生图数据库可以使用无索引邻接进行恒定时间遍历，不需要复杂的架构设计和查询优化。直观的属性图模型不需要创建任何额外的、复杂的应用程序逻辑来处理连接[Webber, 2017]。
- *磁盘空间优化*——为了提高非原生图中的性能，可以将索引非归一化或创建新索引，或将二者结合使用，但这会影响存储相同数量信息所需的空间量。
- *写入效率*——索引非归一化也会对写入性能产生影响，因为所有这些额外的索引结构也需要更新。

2.3.5　标签属性图

相比简单的节点和关系列表，用于表示复杂网络的图需要存储的信息更多。幸运的是，可以很容易地将这种简单结构扩展为更丰富的模型，这种模型包含属性形式的附加信息。此外，还需要对类中的节点进行分组，并分配不同类型的关系。图数据库管理系统供应商引入了标签属性图模型，将一组属性与图结构(节点和关系)联系起来，并对节点和关系进行分类。该数据模型允许使用任何 DBMS 典型的更复杂的查询特征集，如投影、过滤、分组和计数。

根据 openCypher 项目[1]，标签属性图被定义为“具有自边[2]的有向、顶点标签、边标签的多重图，其中边有自己的身份。”在属性图中，我们使用节点来表示顶点，使用关系来表示边。

属性图具有以下属性(此处以与平台无关的方式定义)：

- 该图由一组实体组成。一个实体代表一个节点或一个关系。
- 每个实体都有一个标识符，在整个图中可以对其进行唯一标识。
- 每个关系都有一个方向、一个标识关系类型的名称、一个起始节点和一个结束节点。
- 实体可以具有一组属性，这些属性通常表示为键/值对。
- 可以用一个或多个标签对节点进行标记，这些标签将节点分组并表明它们在数据集中所起的作用。

属性图仍然是图，但沟通功能比以前更强大。在图 2.31 中，你很容易发现 `Person` 的属性 `Alessandro` 与 `Company` 的属性 `GraphAware` 存在 `WORKS_FOR` 关系，Michal 和 Christophe 也是如此。`name` 是节点 `Person` 的属性，而 `start_date` 和 `role` 是关系

1 http://mng.bz/N8wX。

2 自边，也称 sloop，是源节点和目标节点相同的边。

WORKS_FOR 的属性。通过使用关系 HAS_NATIONALITY 存储每个 Person 的国籍，就可以将 Person 连接到 Country 节点，该节点具有存储国家名称的属性 name。

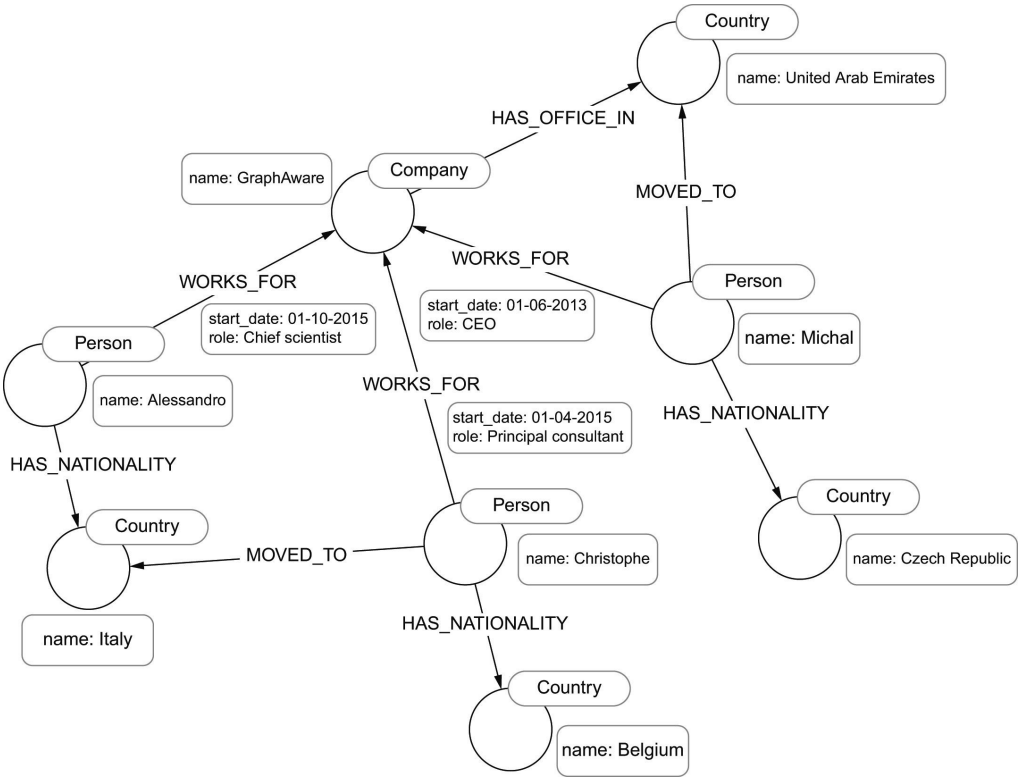

图 2.31 属性图

对于关系数据库，有一些定义图模型的最佳实践或样式规则。例如，节点的标签应该是单数，因为它们代表一个特定实体，而关系的名称则应该反映方向。

显然，可以用多种方式表示同一组概念。例如，在图 2.31 所示的模型中，可以将国籍存储为 Person 节点的一个属性。根据访问模式和潜在图 DBMS 的特定需求，模式可能会发生显著变化。在本书的第 II 部分，我们将看到许多用于表示数据的模型，每个模型都有特定范围并满足目标应用程序的特定要求。

2.4 本章小结

本章描述了机器学习应用程序中与数据管理相关的一些问题，并讨论了图模型如何帮助解决这些挑战。本章通过使用具体场景并描述相关的基于图的解决方案来说明某些特定方面。你学到了以下内容：

- 如何处理大数据的“四 V”：数量、速度、多样性和真实性。“四 V”模型描述了机器学习项目在数据规模、新数据生成速度、数据呈现的异构结构以及来源的不确定性方面面临的多个重要问题。
- 如何设计架构来处理大量训练数据。训练期间，预测分析和机器学习一般需要大量数据才能有效。拥有更多的数据比拥有更好的模型更为重要。
- 如何设计合适的 Lambda 架构，以使用图来存储数据视图。在基于图的 Lambda 架构中，图模型用于存储和访问批处理或实时视图。这些视图代表主数据集的预先计算和易于查询的视图，其中包含原始格式的原始数据。
- 如何规划你的 MDM 平台。MDM 指的是识别、清洗、存储和(最重要)管理数据。在这种情况下，图展示了数据模型中具有的更大灵活性和可扩展性以及搜索和索引功能。
- 如何确定适合应用程序需求的复制模式。复制使得你可以在图数据集群中的多个节点之间分配分析负载。
- 原生图数据库的优势是什么。原生图 DBMS 优于非原生图 DBMS，是因为它们将模型(我们表示数据的方式)与潜在数据引擎一对一映射。这样的匹配提高了性能。

(注：本章的参考文献，请扫描本书封底的二维码进行下载。)